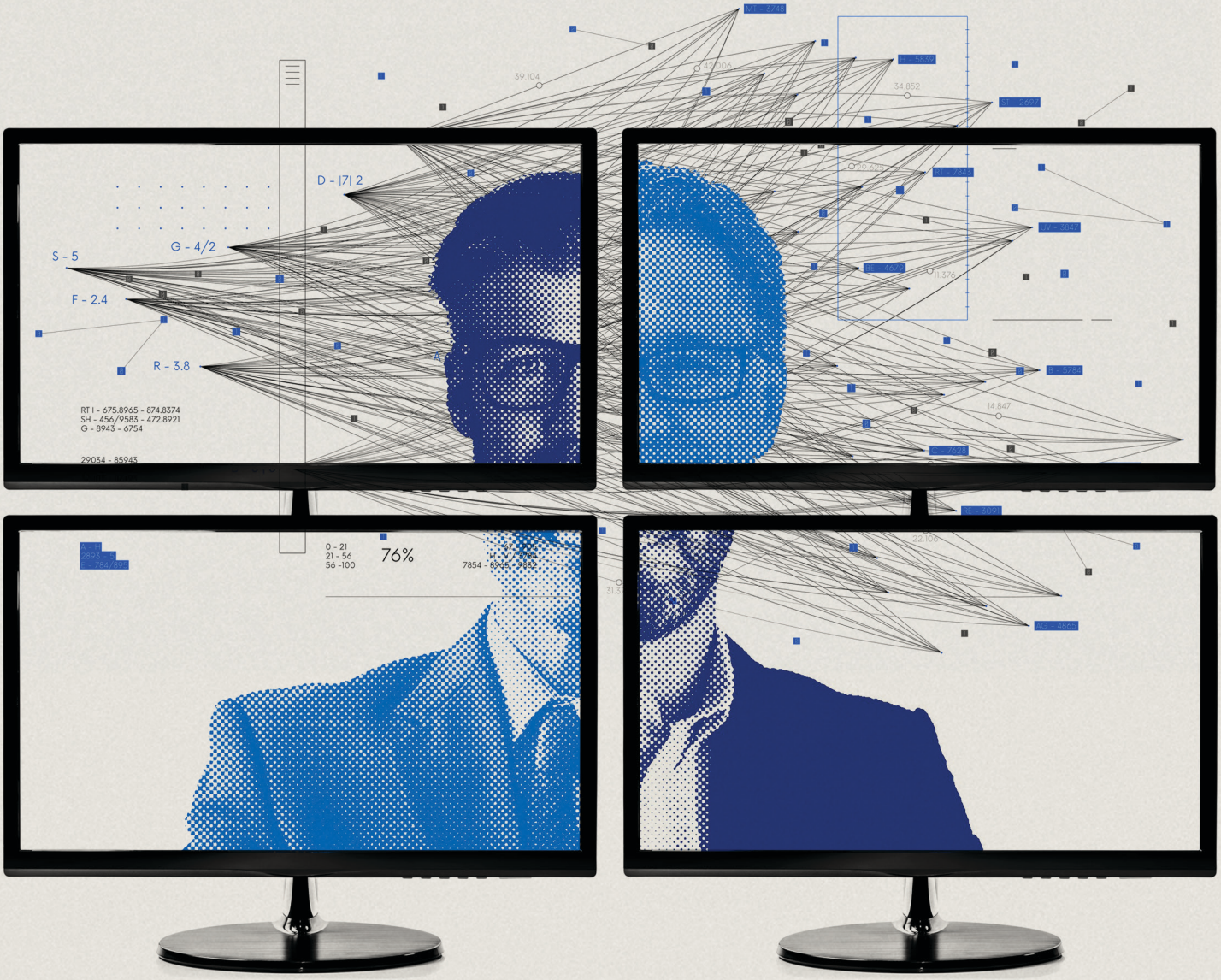
CYBER
THREAT
ANALYSIS

Recorded Future®

By Insikt Group®

March 19, 2024



Adversarial Intelligence: Red Teaming Malicious Use Cases for AI

Executive Summary

Recorded Future threat intelligence analysts and R&D engineers collaborated to test four malicious use cases for artificial intelligence (AI) to illustrate “the art of the possible” for threat actor use. We tested the limitations and capabilities of current AI models, ranging from large language models (LLMs) to multimodal image models and text-to-speech (TTS) models. All projects were undertaken using a mix of off-the-shelf and open-source models, without fine-tuning or training, to simulate realistic threat actor access.

Based on the availability of current tools and the outcome of these four experiments, we assess that malicious uses of AI in 2024 will most likely emerge from targeted deepfakes and influence operations. Deepfakes can already be made using open-source tools and used to impersonate executives in social engineering campaigns combining AI-generated audio and video with conference call and VOIP software. The cost of producing content for influence operations will likely decrease by 100 times, and AI-assisted tools can help clone legitimate websites or spin up fake media outlets. Malware developers can abuse AI along with readily available detections like YARA rules to iterate malware strains and avoid detection. Threat actors of all resource levels will also likely benefit from using AI for reconnaissance, including identifying vulnerable industrial control system (ICS) equipment and geolocating sensitive facilities from open-source intelligence (OSINT).

Current limitations concentrate on the availability of open-source models performing close to state-of-the-art (SOTA) models and bypass techniques for security guardrails on commercial solutions. Given the diverse applications of deepfake and generative AI models, multiple sectors are anticipated to make significant investments in these technologies, enhancing the capabilities of open-source tools as well. This dynamic was previously observed in the offensive security tool (OST) space, where threat actors extensively adopted open-source frameworks or leaked, closed-source tools such as Cobalt Strike. Significant decreases in cost and time will likely lead to a wider variety of threat actors of all technical levels using these attack vectors against a growing number of organizations.

In 2024, organizations need to widen their conception of their attack surface to include their executives’ voices and likenesses, website and branding, and public imagery of facilities and equipment. Furthermore, organizations need to begin preparing for more advanced uses of AI, such as developing self-augmenting malware capable of evading YARA detections, which would require the adoption of stealthier detection methods such as Sigma or Snort.

Key Findings

Use Case I: Using Deepfakes to Impersonate Executives

- Open-source capabilities currently allow for pre-recorded deepfake generation using publicly available video footage or audio clips, such as interviews and presentations.
- Threat actors can use short clips (<1 minute) to train these models. However, acquiring and pre-processing audio clips for optimal quality continues to require human intervention.
- More advanced use cases, such as live cloning, almost certainly require threat actors to bypass consent mechanisms on commercial solutions, as latency issues on open-source models likely limit their effectiveness in streaming audio and video.

Use Case II: Influence Operations Impersonating Legitimate Websites

- AI can be used to effectively generate disinformation at scale, targeted to a specific audience, and can produce complex narratives in pursuit of disinformation goals.
- AI can also be used to automatically curate rich content (such as real images) based on generated text, in addition to assisting humans in cloning legitimate news and government websites.
- The cost of running disinformation campaigns will likely decrease a hundredfold compared to traditional troll farms and human content writers.
- However, creating templates to impersonate legitimate websites is a significant task requiring human intervention to produce believable spoofs.

Use Case III: Self-augmenting Malware Evading YARA

- Generative AI can be used to evade string-based YARA rules by augmenting the source code of small malware variants and scripts, effectively lowering detection rates.
- However, current generative AI models face several challenges in creating syntactically correct code and addressing code linting issues and struggle to preserve functionality after obfuscating the source code.

Use Case IV: ICS and Aerial Imagery Reconnaissance

- Multimodal AI can be used to process public imagery and videos to geolocate facilities and identify ICS equipment, including equipment manufacturers, models, software, and how the equipment is integrated into other observed systems.
- Translating this information into actionable targeting data at scale remains challenging, as human analysis is still required to process extracted information for use in physical or cyber threat operations.

Use Case I: Using Deepfakes to Impersonate Executives

Deepfake content is already being used by threat actors to impersonate executives and political leaders in targeted social engineering attacks. Threat actors can effectively impersonate key leadership figures and connect audio and video deepfake content to conference calls or VOIP technology using publicly available content such as videos, interviews, and pictures. Recent examples have shown that threat actors have also impersonated multiple individuals at once, adding social proof to the attacks. Public instances of executives' voices and likenesses are now part of an organization's attack surface and can be leveraged in damaging social engineering campaigns.

Targeted deepfake attacks can have devastating financial and reputational consequences for organizations when leveraged by financially driven actors. In January 2024, media outlets reported that deepfakes were used to simultaneously impersonate a CFO and other executives from a multinational company in a video conference call to trick an employee into sending over $25.6 million. State-sponsored actors are also almost certainly leveraging deepfakes for political intelligence collection and targeted disinformation campaigns. In June 2022, The Guardian reported that European mayors were targeted by deepfakes on a conference call impersonating Kyiv mayor Vitali Klitschko.

Many commercial voice cloning and TTS solutions now employ consent mechanisms to mitigate against actors using their services to generate deepfakes. While threat actors will undoubtedly invest in capabilities to bypass consent, we have found that publicly available, open-source tools can also be used in social engineering campaigns.

Insikt Group researchers were able to generate deepfake videos impersonating four Recorded Future executives promoting a fictional sponsorship deal. Using open-source video interviews and internal conference call recordings, we were able to train Tortoise TTS, an open-source voice impersonation tool, to impersonate these executives without demonstration of their consent and overlay the resulting audio over existing conference call footage.

Figure 1: *Screenshot from a spoofed conference call impersonating Recorded Future executives (Source: Recorded Future)*

Throughout this project, the Insikt Group researchers identified several caveats and limitations:

1. Latency issues persist in open-source TTS models, which would limit the effectiveness of open-source solutions on live conference calls (currently 500 ms inference speed).
2. For added realism, threat actors would need to use a deepfake video or avatar. However, convincing, low-latency video and audio streaming are mostly offered only by commercial providers at the time of writing, such as ElevenLabs.
3. Acquiring and pre-processing audio clips to train the model still requires human intervention, and will likely be difficult to automate.

However, several trends indicate this will likely change in the next several years. First, the pace of open-source AI development is accelerating and coming close to catching up with closed-source models: open-source LLMs are now nearing the level of models that were previously only attainable by commercial AI providers. Second, these technologies are fundamentally dual-use. AI audio and video generation capabilities will likely see massive investment from industries such as video game developers, film studios, and customer support companies, making the technology faster, better, and cheaper. Third, the proven effectiveness of targeted deepfake campaigns will likely encourage malicious actors to invest in consent-bypassing capabilities to leverage commercial offerings.

Use Case II: Influence Operations Impersonating Legitimate Websites

Generative AI is almost certainly being used by threat actors in influence operations. Advanced LLMs and image models are likely allowing actors to scale their operations, produce multilingual content crossing linguistic boundaries, and adapt content to target audiences.

On December 5, 2023, we reported on Doppelgänger, a Russia-linked influence operation network. The network uses fake news sites, obfuscation, and social media amplification to undermine military aid for Ukraine and polarize politics in the US, France, Germany, and Ukraine. The network was observed using a limited quantity of likely AI-generated text content.

Consistently identifying AI-generated text content remains a challenge for many researchers, given the lack of standardized analytical frameworks to determine the likelihood of an asset used in influence operations being AI-generated. To understand the real risk of AI-generated content being used at scale in influence operations, we set out to build a disinformation pipeline capable of impersonating legitimate Russian and Chinese media websites, injecting AI-generated content, and including images that are automatically selected based on the generated text content.

Insikt Group researchers succeeded in creating AI-generated content that was highly targeted toward a specified audience, and in using prompt engineering techniques to ensure that content was geared toward their political views. The pipeline would take an influence objective simulating threat actors' real objectives or a seed article from a legitimate source to provide factual context. We were also able to leverage generative AI to automatically analyze HTML content from legitimate news organizations in order to clone and template their websites, allowing us to scale our production of disinformation content. Finally, we were able to leverage a publicly available, multimodal AI model to select stock or publicly available images relevant to the AI-generated content.

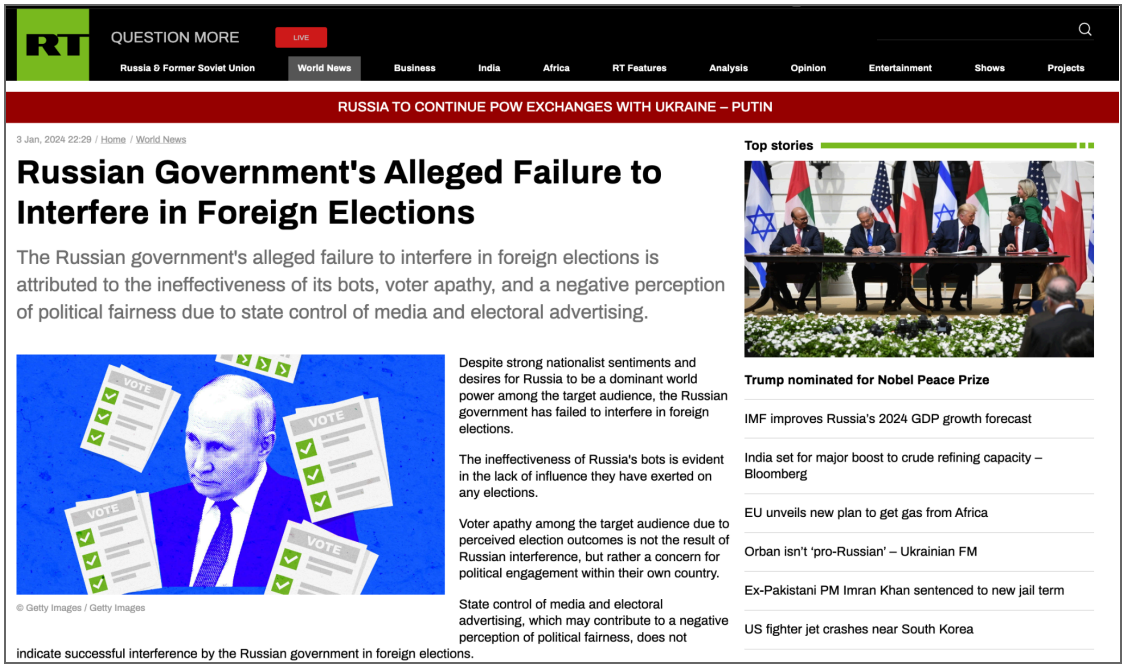


***Figure 2**: Impersonated RT website using AI-generated content (Source: Recorded Future)*

Several limitations arose during this process. First, using AI to generate templates from cloned news organizations required significant human quality assurance (QA) to ensure that the resulting output closely resembled the cloned website. Only 30% of automated cloning attempts across a large dataset of news organizations worked without human intervention. Second, AI-generated text tailored to a specific audience would hallucinate if the target audience wasn't properly aligned with the stated objective, meaning that threat actors still need to craft quality datasets of target audiences.

Nevertheless, we were able to build a pipeline that enabled us to scale one piece of AI-generated content into many different clones of legitimate websites, across different languages, and targeting different audiences. Injecting AI-generated content into templates impersonating legitimate news organizations also provides further social proof and legitimacy to content: users browsing disinformation assets spoofing legitimate organizations are more likely to perceive such information as credible, enabling threat actors to be more effective in their approach.

Furthermore, this approach is considerably cheaper than generating content using human writers and developers, even if those are outsourced to countries with lower wages: website templates and articles could be generated for less than $0.01 each. For a troll farm content writer in a country where the minimum daily wage is $10 (such as North Macedonia, the Philippines, or Latin America) tasked with producing 10 articles daily, using AI could make creating content 100 times less expensive. Furthermore, scaling the publication of disinformation content using AI could lead to these websites netting higher amounts of ad revenue, which can result in profitable disinformation ventures.

Use Case III: Self-augmenting Malware Evading YARA

Malware and exploit development using generative AI are almost certainly current areas of research for threat actors. However, according to a January 2024 threat assessment by the UK's National Cyber Security Centre (NCSC), this will likely be limited to well-resourced actors with access to high-quality, proprietary malware and exploit datasets. This makes the assumption that less-well-resourced actors would struggle to develop malware and exploits using generative AI trained on publicly available detection datasets.

Security researchers and threat intelligence analysts frequently publish YARA rules as a method for identifying and classifying malware based on specific patterns in malware samples. These publicly available rules also serve as a double-edged sword. While they are intended to be a resource for defenders to enhance their security measures, they also provide threat actors with insights into detection logic, enabling them to adjust their malware characteristics for evasion. Nevertheless, modifying malware to avoid detections published by researchers is likely to be a resource-intensive and time-consuming endeavor, imposing a significant cost on adversaries.

Our project aimed to alter malware source code using generative AI to evade YARA detection. For testing, researchers used STEELHOOK, a PowerShell infostealer used by APT28 to steal browser data from Google Chrome and Microsoft Edge web browsers. Steelhook source code was submitted to an LLM with associated YARA rules, prompting it to modify the source code to evade detection. The malware candidate generated by the LLM underwent a series of validations to ensure that three conditions were met: freedom from syntax errors, absence of detection by YARA, and preservation of the same functionality as the original malware. Errors encountered during the validation phase were provided as feedback to the LLM, prompting it to enhance or correct its output.

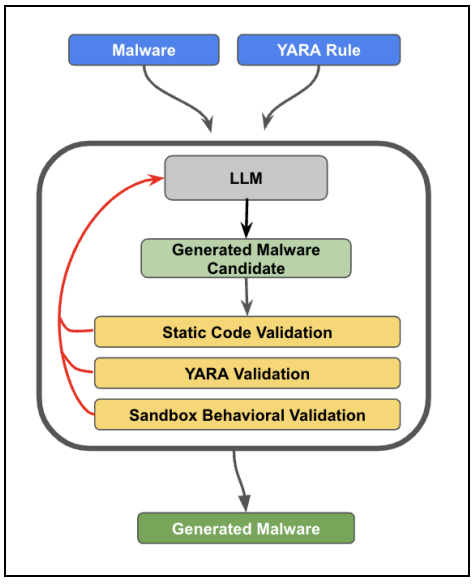


Figure 3: *Flowchart for creating malware to evade YARA detection (Source: Recorded Future)*

Using this feedback loop, Insikt Group researchers were able to avoid detections for simple string-based YARA rules. However, current generative AI models still impose severe limitations. Current context windows (the amount of text a model can "process" at one time) remain too small to operate on larger code bases, limiting our testing to smaller scripts and less complex malware strains. Moreover, AI models face challenges in attempting to simultaneously meet all three conditions mentioned above. Lastly, this approach did not test more complex YARA detection logic beyond string detections, such as byte pattern matching and modules, which would require extra logic to preprocess samples to map byte pattern matches to their source code location, support for source code compilation, and extra validation phases targeting specific YARA modules.

In summary, there exists both potential and risk in the use of AI for the dynamic creation of malware capable of evading YARA detections. The evasion of YARA rules by AI-generated malware may underscore the increased relevance of alternative detection methods like Sigma from a security standpoint, as they are presumably more resistant to evasion. However, at the same time, current generative AI models face several challenges in creating syntactically correct code and addressing lint issues, necessitating the implementation of multiple feedback loops.

Use Case IV: ICS and Aerial Imagery Reconnaissance

Generative AI will almost certainly expedite threat actors' ability to conduct reconnaissance on target facilities, equipment, and sites ahead of physical or cyberattacks. By leveraging multimodal models, public images and videos of ICS and manufacturing equipment, in addition to aerial imagery, can be parsed and enriched to find additional metadata such as geolocation, equipment manufacturers, models, and software versioning.

While nation-state actors have had full imagery intelligence (IMINT) agencies working on collecting and parsing aerial and satellite imagery since the Cold War, less-well-resourced actors will likely benefit from multimodal models able to analyze imagery obtained via OSINT. In addition to being able to geolocate facilities, threat actors will also be likely to use AI to extract ICS equipment models and determine potential vulnerabilities with equipment seen in public footage.

There are many examples online of images and videos shot in and around sensitive critical infrastructure facilities. These videos often attempt to avoid revealing sensitive information, but potential vulnerabilities can be discovered through OSINT research. Researchers attempted to automate this process and determine sensitive information about a power generation facility.

We used LLMs to extract information from manually obtained, unique keyframes in open-source video content. As a result, the LLM was initially able to identify a power plant's approximate location within a 100-km radius — as it was fed more keyframes, including the equipment and text visible in the video, it was able to narrow it down to the exact location through deduction. The LLM was also able to provide high-confidence assessments on the versions of specific equipment, their manufacturer and operating software, and how they are integrated with other systems. Using this information, operators would

likely be able to build a comprehensive image of industrial equipment in a facility and correlate this intelligence with knowledge of physical flaws or cyber vulnerabilities.

Several technical challenges arose during this process. First, it is still difficult to programmatically extract keyframes that are sufficiently clear or visible for assessment by an AI model. Automating this process would require more research to finetune image extraction and limit computing costs. Second, categorizing information and insights for later retrieval by an agent to produce a more complex assessment remains a significant technical barrier for current models. Third, it can still be costly to process hours of video with the most advanced models.

Nation-state actors almost certainly intend to conduct reconnaissance on adversaries' critical infrastructure, and very likely have the means to fully automate this process. By harvesting imagery via social media and aerial imagery feeds, actors will likely be able to process this information using multimodal models and store results for later retrieval by intelligence analysts. By correlating this information with exploit development efforts and known vulnerabilities, cyber threat actors can use this approach to identify targets and inform lateral movement and impact techniques for ICS equipment.

Mitigations

- Executives' voices and likenesses are now part of an organization's attack surface, and organizations need to assess the risk of impersonation in targeted attacks. Large payments and sensitive operations should use several alternate methods of communication and verification, other than conference calls and VOIP, such as encrypted messaging or emails.
- Organizations, particularly in the media and public sector, should track instances of their branding or content being used to conduct influence operations. Recorded Future customers can use the Brand Intelligence Module to track new domain registrations and online content using their branding.
- Organizations should invest in multi-layered and behavioral malware detection capabilities in the event that threat actors are able to develop AI-assisted polymorphic malware. Sigma, Snort, and complex YARA rules will almost certainly remain reliable indicators for malware activity for the foreseeable future.
- Publicly accessible images and videos of sensitive equipment and facilities should be scrutinized and scrubbed, particularly for critical infrastructure and sensitive sectors such as defense, government, energy, manufacturing, and transportation.

Outlook

Recorded Future researchers were able to demonstrate the viability of four malicious use cases without using expensive techniques such as model finetuning. As a result of concrete experiments, we assess that by 2024, it's highly likely that AI will be incorporated into social engineering and information operations.

More advanced use cases, such as malware development and reconnaissance, will likely benefit from generative AI advances and become viable over a longer period of time. Well-resourced actors with sufficient computing power and training datasets will likely adopt these techniques before proliferating more widely. Decreases in cost and advancements in model performance in 2024 are certain, and diffusion of AI technology will very likely follow as a consequence. As observed in the OST space, advanced capabilities will likely fall into the hands of a wider variety of actors, either through the catching up of open-source models or the leaking of advanced closed-source models.

Identified limitations were concentrated around the inability to completely eliminate human intervention from these use cases. “Human-in-the-loop” tasks such as audio and video editing, cleaning templates for cloned websites, verifying malware execution, and processing reconnaissance data will likely persist for the foreseeable future. However, advances in agents and the emergence of specialized models capable of video editing, processing HTML, reading large code bases, and categorizing insights indicate these use cases will likely be fully automated in the future.

Looking ahead, organizations need to widen their perception of their attack surface. Any text, image, audio, or video data associated with a brand or employee can and will be leveraged for malicious uses using AI. Voices and likenesses of corporate and political leaders can be used in targeted social engineering attacks, to devastating financial and political effects. AI-generated text can be trained to impersonate legitimate news and government sources on cloned websites to spread disinformation at scale. Detection rules published by security researchers can be used to iterate malware strains and avoid detection. Public images and videos of facilities can be mined for information on geolocation and vulnerabilities in equipment, which can be used to inform physical and cyber targeting.

Recorded Future®

About Insikt Group®

Recorded Future's Insikt Group, the company's threat research division, comprises analysts and security researchers with deep government, law enforcement, military, and intelligence agency experience. Their mission is to produce intelligence that reduces risk for clients, enables tangible outcomes, and prevents business disruption.

About Recorded Future®

Recorded Future is the world's largest threat intelligence company. Recorded Future's Intelligence Cloud provides end-to-end intelligence across adversaries, infrastructure, and targets. Indexing the internet across the open web, dark web, and technical sources, Recorded Future provides real-time visibility into an expanding attack surface and threat landscape, empowering clients to act with speed and confidence to reduce risk and securely drive business forward. Headquartered in Boston with offices and employees around the world, Recorded Future works with over 1,700 businesses and government organizations across more than 75 countries to provide real-time, unbiased, and actionable intelligence.

Learn more at recordedfuture.com