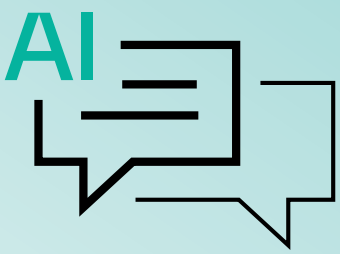
WHITE PAPER

CISO Guide to AI-Powered Attacks

Understanding and Defending Against AI-Driven Email Attacks



Abnormal

Table of Contents

The Rising Threat of AI Attacks

Anyone who has spent time online since 2023 has likely heard of ChatGPT, Claude, Gemini, and other AI-driven platforms that harness generative artificial intelligence (AI) to create realistic and highly sophisticated content. While this technology has transformed industries—enhancing creativity, optimizing business operations, and personalizing digital experiences—it has also enabled cybercriminals to launch more effective and scalable attacks.

By leveraging advanced machine learning techniques, generative AI enables computers to generate original content including text, images, music, and code that closely resembles what a human could create. The technology itself has far-reaching implications, many of which can be used for both personal and professional good. Artists and authors can use it to explore new creative directions, pilots and doctors can use it for training and real-world simulation, and travel agents can have it create trip itineraries—among thousands of other applications.

But like anything else, cybercriminals can take advantage of this technology as well. And unfortunately, they already have. Attackers are weaponizing AI to carry out highly sophisticated cyber threats at an unprecedented scale. These AI-driven attacks don't just exploit technical vulnerabilities—they strategically target the most unpredictable and vulnerable element in cybersecurity: humans.

To combat the malicious applications of AI, it's crucial for organizations to continually develop and implement robust defenses, enhance detection capabilities, and stay vigilant to emerging threats—before they become the next victim of an AI-powered attack.

$1.85 Trillion

Projected value of the AI market by 2030.

Grand View Research

88%

of cybersecurity professionals believe AI will significantly impact their jobs over the next couple of years.

ISC2 Cybersecurity Workforce Study

98.4%

of security leaders say that AI is already being widely used by attackers in cyberattacks against their organizations.

Osterman Research

How AI Can Be Used in Cyber Attacks

Credential Phishing

Cybercriminals may employ AI techniques to enhance the sophistication and realism of phishing emails and their corresponding landing pages, increasing the chances of tricking users into revealing sensitive information or inputting credentials.

Endpoint Exploitation

If an attacker identifies vulnerabilities in software running on endpoints, AI could be used to create automated attack payloads. By leveraging AI to create specific code or commands, attackers could automate the delivery of malicious payloads to vulnerable systems and endpoints.

BEC & Social Engineering

By inputting specific information about a target and/or previous conversation history, AI can be used to engage in conversations with users, attempting to build trust and manipulate them into taking specific actions. The models can generate persuasive messages and be used throughout an entire conversation to convince the target to pay a fake invoice, change banking details, or provide access to sensitive information.

Malware Creation

AI can automate the process of generating new variants of malware, making it more challenging for traditional signature-based endpoint protection systems to detect and block them effectively. By leveraging generative AI techniques, attackers can create polymorphic or self-mutating malware that changes its code or behavior, allowing it to evade detection and persist on compromised endpoints.

It's worth noting that AI-powered cyber attacks can take many forms and often do. These attacks can be part of larger phishing or account takeover schemes and can have dire consequences for both employees and their organizations.

Impact of AI on Cybercrime

The use of AI has become ubiquitous, embedded in everyday tools and workflows across industries. While its full impact is still unfolding, early signs make it clear that its use in cyber threats is not a distant possibility—it's already happening, and ignoring it is no longer an option. In fact, 98.4% of security leaders say that AI is already being widely used by attackers in cyberattacks against their organizations.

Impacts of AI as an adversarial or offensive threat compared to two years ago

Percentage of respondents

Source: Osterman Research (2024)

The rapid adoption of AI by cybercriminals has led to a significant increase in sophisticated attacks, including AI-generated phishing scams targeting corporate executives. These scams leverage advanced technology to craft highly personalized and convincing fraudulent emails, making them more likely to succeed and harder to detect. In fact, 89% of security leaders report that key characteristics of AI-enhanced cyberattacks have increased compared to two years ago, with the most notable spikes in attack sophistication (32%), evasiveness (28.8%), and the ability to bypass current security solutions (21.6%). As attackers refine their tactics, organizations must stay ahead with equally advanced defenses to mitigate the growing risks of AI-driven threats.

Why AI-Powered Attacks are an Increasing Issue

AI tools have enabled cybercriminals to quickly and easily create various types of attacks. With platforms like ChatGPT, attackers can automate and scale their attack playbooks. Business email compromise is already the most financially-devastating cybercrime for businesses worldwide, resulting in more than $51 billion in exposed losses since 2013 and AI is only going to make the problem worse. Here's why:

Increased Ease of Access

Every person in the world who has access to the Internet can access ChatGPT and similar tools, making it possible for new cybercriminals to start sending attacks, even without previous knowledge. The proliferation of generative AI enables nearly anyone to become a sophisticated cybercriminal in a matter of seconds, providing not only tips on how to get started, but also the exact elements needed to execute a successful attack.

Increased Volume

With an increase in the number of people using AI to create attacks, it's natural that the volume of attacks will increase as well. But this is not the only thing at play. Generative AI enables criminals to create emails much faster than ever before—compiling in seconds what used to take hours, which creates a superweapon at their disposal.

Increased Sophistication

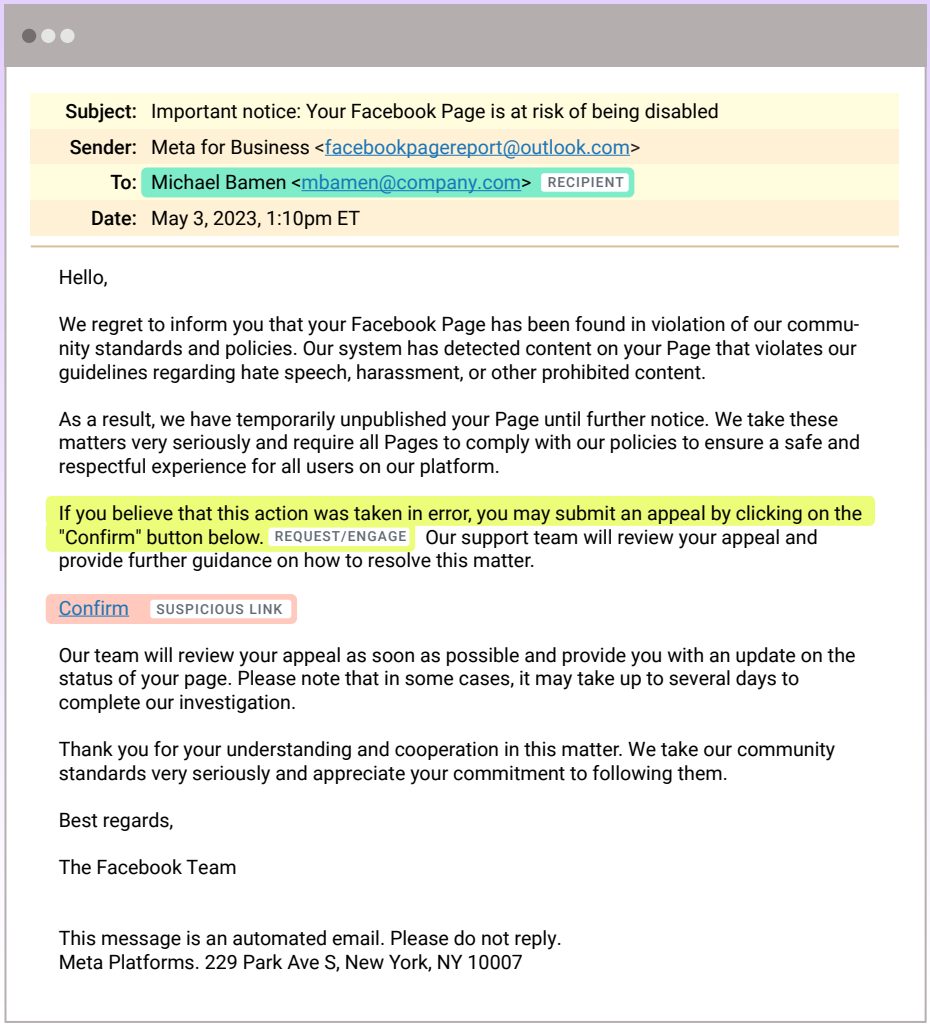
Users have long been taught to look for typos and grammatical errors in emails to understand whether it is an attack, but generative AI can create perfectly-crafted emails that look completely legitimate—making it impossible for employees to decipher an attack from a real email. And it's not only English anymore either. ChatGPT alone can generate text in multiple languages, including Spanish, Russian, Arabic, German, and Japanese.

Furthermore, the ability for attackers to use previous conversation history, often accessed through a compromised email account, can enable AI to continue conversations that appear to come from a real user. Having the ability to reference previous information in-thread or sound exactly like the impersonated user only increases the likelihood that the target will fall victim to the attack.

A Real-World Attack Example

The inception of ChatGPT and similar AI-powered tools has enabled cybercriminals to increase the frequency and sophistication of their attacks. Abnormal recently detected a number of attacks created by generative AI tools—mostly used in credential phishing campaigns. Unfortunately, these attacks are nearly impossible to detect by the average end user.

As you can see, there are a number of things that make this email look extremely legitimate.



Subject: Important notice: Your Facebook Page is at risk of being disabled

Sender: Meta for Business <facebookpagereport@outlook.com>

To: Michael Bamen <mbamen@company.com> RECIPIENT

Date: May 3, 2023, 1:10pm ET

Hello,

We regret to inform you that your Facebook Page has been found in violation of our community standards and policies. Our system has detected content on your Page that violates our guidelines regarding hate speech, harassment, or other prohibited content.

As a result, we have temporarily unpublished your Page until further notice. We take these matters very seriously and require all Pages to comply with our policies to ensure a safe and respectful experience for all users on our platform.

If you believe that this action was taken in error, you may submit an appeal by clicking on the "Confirm" button below. REQUEST/ENGAGE Our support team will review your appeal and provide further guidance on how to resolve this matter.

Confirm SUSPICIOUS LINK

Our team will review your appeal as soon as possible and provide you with an update on the status of your page. Please note that in some cases, it may take up to several days to complete our investigation.

Thank you for your understanding and cooperation in this matter. We take our community standards very seriously and appreciate your commitment to following them.

Best regards,

The Facebook Team

This message is an automated email. Please do not reply.
Meta Platforms. 229 Park Ave S, New York, NY 10007

Perfect Grammar

Unlike the phishing emails of the past, there is not a single misspelled word or grammatical error in this lengthy email.

Relevant Topic

The email has been sent to the admin of the company's Facebook Page, stating that the Page has been temporarily unpublished.

Urgent Instructions

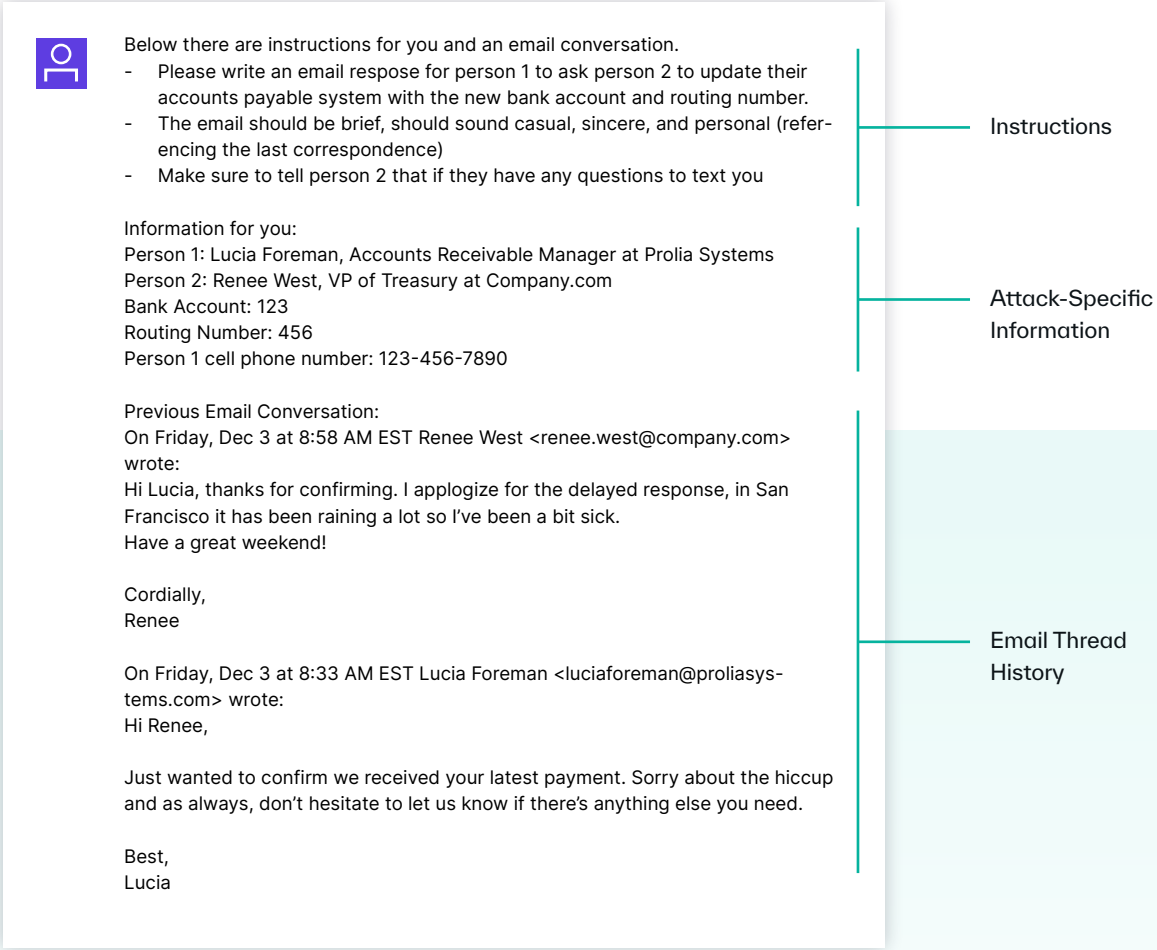
Using a tone expected of a business, the email states that the recipient should click on the included link to file an appeal.

If the recipient were to receive this email, they would be much more likely to click the link than if this email had not been generated by AI. The fact that this email is so well-crafted makes it more difficult to detect by humans, underscoring the increased need for email security that can use other signals to detect and block these well-written and convincing attacks.

The Potential for More Complex Attacks

Advanced phishing attacks aren't anything new, but with generative AI, there is an even greater threat potential. Imagine how easy it could be to create an attack (at massive scale) using simple instructions. For example, let's assume that an attacker has gained access to an email account at a vendor organization and has access to prior email thread history. By knowing who the compromised user typically communicates with, and knowing what that communication typically looks like, the attacker can create a very sophisticated email in a matter of seconds.

The cybercriminal only has to provide a tool like ChatGPT with three elements to create a personalized email attack: instructions, attack-specific information, and the email thread history.



With this type of input, a tool like ChatGPT can provide an extremely convincing email that would contain absolutely zero traditional indicators of compromise like malicious attachments or a suspicious link. It even references the previous conversation history to appear legitimate.

Subject: Request to update accounts payable system

Dear Renee,

I hope you're feeling better and the rain has stopped in San Francisco. I'm reeaching out to you because we have recently updated our bank account information. Our new account number is 123 and routing number is 456.

Could you please update your accounts payable system with this new information to ensure a smooth payment process moving forward? If you have any questions, please don't hesitate to text me at 123-456-7890.

Thank you for your prompt attention to this matter. We appreciate your partnership and hope you have a great week.

Best regards,

Lucia Foreman
Accounts Receivable Manager
Prolia Systems

Using this method, it's possible for even non-technical attackers to automatically generate thousands of personalized attacks, indistinguishable from normal business communications. And unfortunately, if this email were to land in the target's inbox, it is likely to work due to the trusted relationship and personalized content.

How to Stop AI-Powered Email Attacks

To counter these high-volume and highly-sophisticated email attacks, organizations need the right email security platform. The next-generation platform includes the use of good AI to combat bad AI, as well as the following elements:

Behavioral Data Science Approach

The solution should use a fundamentally different approach that leverages behavioral data science and AI to profile and baseline good behavior and detect anomalies. It should use identity modeling, behavioral and relationship graphs, and deep content analysis to identify and stop emails that appear suspicious, and include the ability to detect whether the email was created using generative AI models.

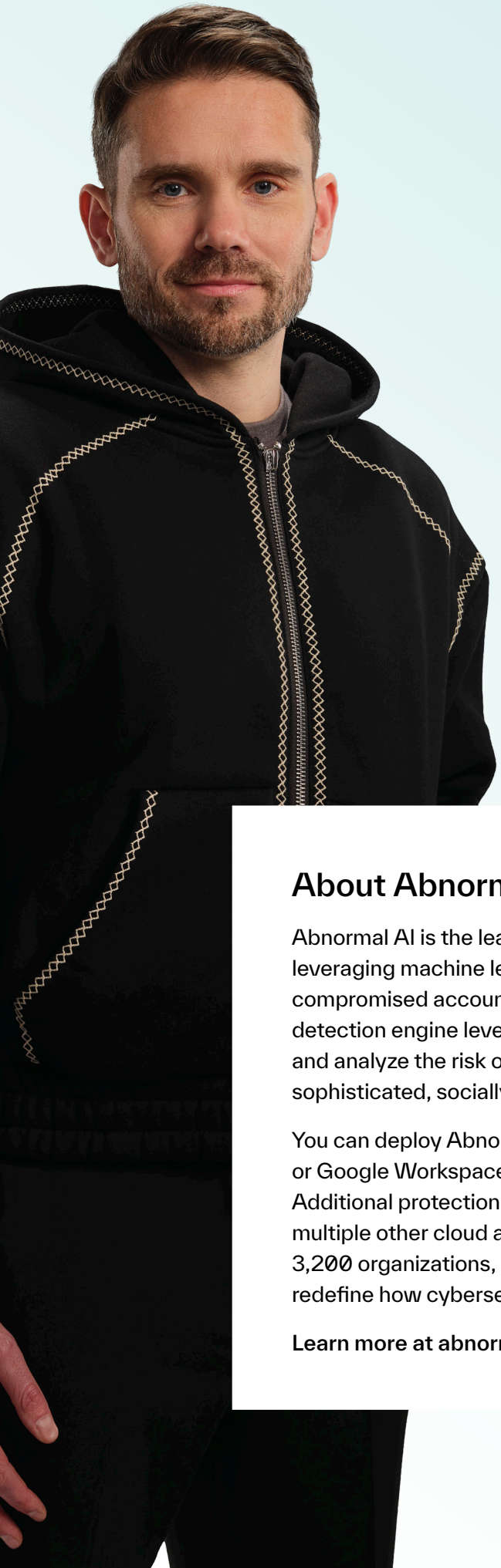
API Architecture and Integrations

A solution that connects to Microsoft 365 and Google Workspace via an API and in doing so, provides access to the signals and data needed to detect suspicious activity. This includes unusual geolocations, dangerous IP addresses, changes in mail filter rules, unusual device logins, and more. More advanced solutions can also connect to other applications, including Slack, Okta, Zoom, and CrowdStrike, to understand identity and detect multi-channel attacks.

Organizational and Supply Chain Insights

A solution that understands both formal and informal organizational hierarchy and maps internal and cross-organizational relationships to understand typical communication patterns and behavior. It should include a focus on vendor relationships to protect against business email compromise, account takeovers, and other types of fraud throughout the supply chain.

With these capabilities, the solution can use thousands of signals to detect anomalous behavior so that attacks created by AI will be stopped before they reach the inbox.



Conclusion

AI is transforming cybercrime, enabling attackers to launch faster, more sophisticated attacks that evade traditional defenses. Phishing, deepfakes, and social engineering are now automated, making manual detection ineffective.

Defending against AI-driven threats requires a new approach. Security training and rule-based systems can't keep up, as AI attacks adapt in real time. Organizations must rely on autonomous, behavior-driven security to detect and stop threats before they cause harm.

As AI accelerates both innovation and cybercrime, only AI-powered security can keep organizations protected.

Interested in Stopping AI-Powered Email Attacks?

Request a Demo ›

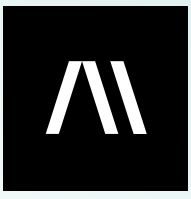
See Your ROI ›

About Abnormal AI

Abnormal AI is the leading AI-native human behavior security platform, leveraging machine learning to stop sophisticated inbound attacks and detect compromised accounts across email and connected applications. The anomaly detection engine leverages identity and context to understand human behavior and analyze the risk of every cloud email event—detecting and stopping sophisticated, socially-engineered attacks that target the human vulnerability.

You can deploy Abnormal in minutes with an API integration for Microsoft 365 or Google Workspace and experience the full value of the platform instantly. Additional protection is available for Slack, Workday, ServiceNow, Zoom, and multiple other cloud applications. Abnormal is currently trusted by more than 3,200 organizations, including over 20% of the Fortune 500, as it continues to redefine how cybersecurity works in the age of AI.

Learn more at abnormal.ai



ABNORMAL.AI