CISO STRATEGY GUIDE

Social Engineering in the Age of AI

How to Augment Native Email Security Defenses with AI

The Rise (and Rise) of Social Engineering

Whether created by a person, generative AI (often shortened to GenAI), or a blend of both, a phishing email must meet two criteria to be successful. First, it must go undetected by existing email security defenses. Then, once it's in the recipient's inbox, it must convince them to act.

For years, organizations have relied solely on the signature-based and reputation-based detection capabilities offered by legacy email defense vendors. While effective at detecting previously identified payloads that have been added to their definitions libraries and identifying on certain red flags in domain analysis, this technology can be readily exploited by cybercriminals.

Once an attack gets through the legacy technical layer, it then must socially engineer its target.

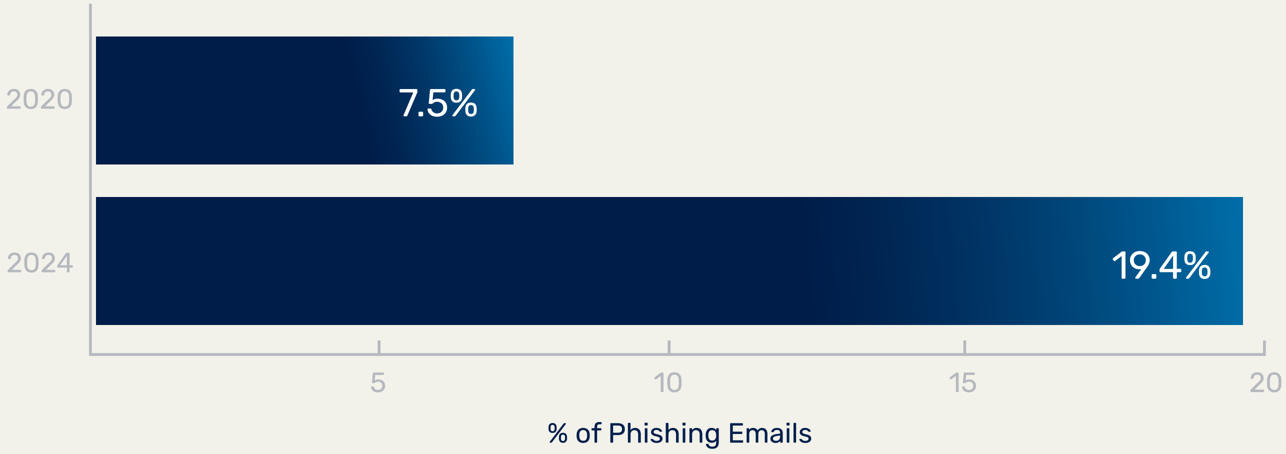
Even in this age of growing cybersecurity awareness, Verizon's 2024 Data Breach Investigations Report uncovered that:[1]

1. People are organizations' greatest risk, with 68% of security incidents involving the human element
2. Pretexting is the leading cause of cybersecurity incidents, with threat actors targeting people via existing email chains and context

Just as they've refined their phishing emails to avoid detection by native defenses and SEGs, cybercriminals also invest resource to make their attacks more convincing — something that's getting easier in the age of AI.

In fact, a growing number of phishing emails rely solely on text-based socially engineered messages and non-malicious attachments (such as fraudulent invoices).

Increase in Attacks Relying Soley on Social Engineering 2020-2024



In this CISO strategy guide, we'll look at why people are susceptible to social engineering; the tactics cybercriminals use to evade detection by SEGs and manipulate their targets; and how AI has changed the threat landscape.

Social Engineering in the Age of AI

AI is an incredibly powerful tool for increasing the sophistication and scale of all phishing attacks, especially when it comes to social engineering. It has irrevocably changed the threat landscape, and organizations must ensure their defenses can prevent this wave of advanced threats.

Reconnaissance: Setting the scene

By processing information at incredible speeds, AI helps cybercriminals aggregate and correlate data across numerous websites, online platforms and previous data breaches. This can be used to identify targets and channels for attack (especially for multi-channel campaigns) but most importantly, it can be used to produce believable pretexts.

GenAI: Creating convincing attacks at scale

The large language models (LLMs) and natural language generation (NLG) powering GenAI chatbots can turn simple prompts into well-written attacks, which can be enhanced using the information collected during reconnaissance. This includes creating realistic pretexts and leveraging other social engineering tactics, such as urgent language. Chatbots can also refine and translate the text in existing social engineering attacks to improve their efficacy.

Additionally, AI can process the visual and audio outputs of videos, podcasts and other media available online to generate deepfakes. A previously resource-heavy activity that delivered questionable outputs, we're now starting to see the impact of improved AI tools in accelerating and enhancing deepfake generation: Zoom and mobile phone calls have increased as the second step in multi-channel attacks, while SMS has decreased.[2]

Automation: Continuing the conversation

Cybercriminals don't even need to be at their computers to generate advanced phishing attacks. AI can automate the research, creation and sending of highly targeted attacks at scale. GenAI chatbots can be used to respond to replies from phishing emails in real time, engaging with victims and quickly directing them to the cybercriminals' intended outcome.

People Are Hardwired to Be Deceived

People fall victim to social engineering because we are not as rational as we'd often like to believe. In "Use Behavioral Economics to Influence Security Behavior and Individual Decisions," Gartner states:

"Behavioral economics suggests that human decisions are strongly influenced by context, including the way in which choices are presented to us. Behavior varies across time and location, and it is subject to cognitive biases, emotions and social influences. Decisions are the result of less deliberate, linear, rational and controlled processes than we would like to believe."[3]

These factors that influence behavior – time, location, cognitive biases, emotions, and social influences – can all be exploited by cybercriminals' social engineering tactics.

Heuristics and cognitive bias

Heuristics are "mental shortcuts" or "rules of thumb" that people use every day to process new information to simplify and speed up decision making. However, they're not always grounded in logic and accuracy, and can lead to cognitive bias and error.

There are several heuristics that cybercriminals exploit when socially engineering their targets:

Representativeness
Judgment of how likely an object belongs to a general category or class based on how similar it is to members of that category.

Availability
Assessment of how often an event occurs or how likely it will occur based on how easily that event can be brought to mind.

Halo effect
Snap judgments, which are particularly influenced by first impressions, prominent characteristics or recent experiences. The halo effect leads people to only consider one aspect of something to form a general opinion.

Authority bias
Unreasonably high confidence that information verified by a person or organization with formal authority is correct.

Hyperbolic discounting
Inclination to choose immediate rewards or satisfaction over future gratification.

Social proof
Copying the actions of others, especially in uncertain or ambiguous situations, or to fit in with the group.

On average it only takes a recipient 21 seconds to click on a malicious hyperlink and a further 28 seconds to enter credentials into a malicious website.

So, when a junior employee receives a CEO fraud attack impersonating a senior executive within their organization, authority bias can make them more susceptible to quickly carrying out instructions. They may assess whether the email fits with other communications from that person (representativeness and halo effect), meaning it must seem realistic, for example using appropriate language or making a passably reasonable request.[4]

On average it only takes a recipient 21 seconds to click on a malicious hyperlink and a further 28 seconds to enter credentials into a malicious website — so most phishing attacks only have to be good enough to evoke fast judgments.

Type 1 and Type 2 Thinking

Nobel Prize-winning psychologist Daniel Kahneman described two systems for thinking:[5]

Type 1

Type 1 thinking is referred to as "thinking fast." It is characterized as automatic decision making, which is done at velocity and, often, at volume. As it is based in instinctive and quick reactions, it is more frequently prone to error than Type 2 thinking. People spend most of their time in this state.

Type 2

Type 2 thinking is slow and deliberate, characterized as being logical and rational, and more accurate.

It's in cybercriminals' interest to generate fast, instinctive responses that are the hallmark of Type 1 thinking. Frequently phishing attacks will contain social engineering tactics that create pressure, urgency and even fear to trigger this response. For instance, in our CEO fraud attack example, the cybercriminal could offer a "reward" (such as social favor for carrying out the requested task) or a "punishment" (e.g. being the employee responsible for losing a customer because a task wasn't completed) — or both!

Cybercriminals, therefore, must engineer fast and instinctive reactions, before a target's critical thinking kicks in.

Not So Social: How Cybercriminals Manipulate Their Targets

As we've seen, on average it only takes 21 seconds for victims to interact with a phishing email. Social engineering tactics are designed to manipulate people's heuristics, cognitive biases and thinking patterns to elicit these quick responses.

Here are the top tactics used to manipulate targets:

Pretexting
The leading cause of cybersecurity incidents,[6] pretexting involves creating a false scenario to manipulate targets into taking specific actions, such as sharing sensitive data or paying a fraudulent invoice. Depending on the pretext created, many heuristics can be manipulated.

Impersonation
In almost every phishing email, cybercriminals must disguise their identities. This can be achieved both technically — i.e. domain spoofing to make the email appear to be from a legitimate sender — and within the body copy, whether it's the language used to impersonate a supplier or colleague, or a branded HTML template of a household name. The use of compromised accounts, particularly within the supply chain, is an effective impersonation tactic.

Baiting
Cybercriminals play with many of our basic instincts, such as greed. With baiting, a victim's heuristic for hyperbolic discounting is triggered and they are lured in with an attractive offer, such as unlocking a limited-time deal by logging into their account. Baiting can also include piquing a target's interest with exclusive information or offers.

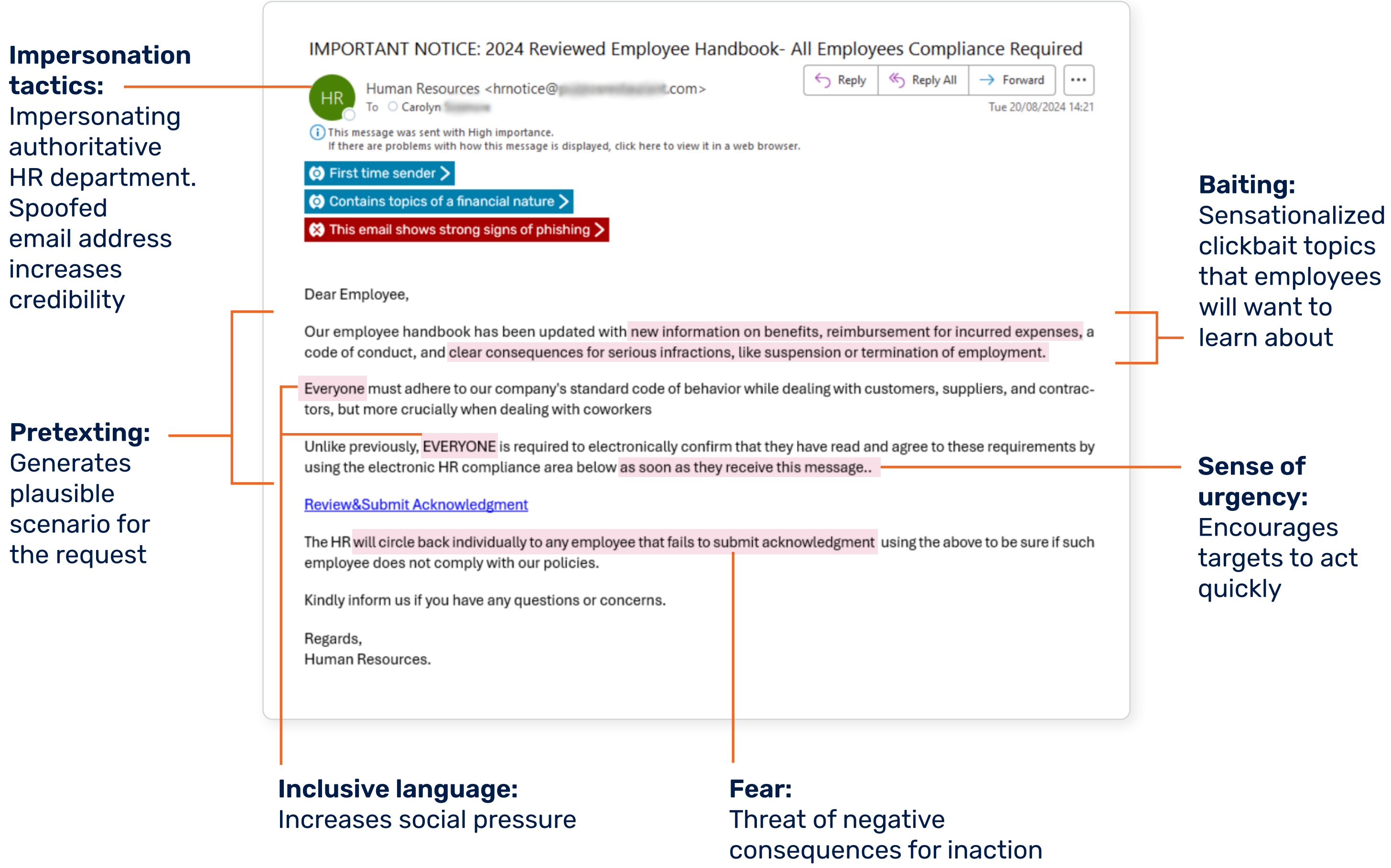
Fear
One of the most powerful emotions, fear is used to stimulate Type 1 thinking and galvanize a target into action. Cybercriminals make overt or implicit threats that doing nothing will lead to negative consequences, ranging from smaller outcomes, like the closure of an online account, to the more severe, such as being responsible for the loss of a client or having sensitive videos shared with relatives as part of sextortion attacks.

Urgency
Again, designed to provoke Type 1 thinking and make people act quickly, creating a sense of urgency taps into negative consequences for inaction, such as losing a discount within 24 hours or needing to resolve an invoice before the end of the work day.

Authoritative language
The most persuasive technique in phishing,[7] the use of authoritative language exploits people's obedience to hierarchy, triggering the authority bias heuristic and tricking them into carrying out the specified actions.

Social Engineering in the Real World

This phishing attack was missed by native defenses and a SEG, but was detected by KnowBe4 Defend. Which among other detection capabilities, utilized AI-enabled linguistic analysis to identify the social engineering tactics present, in addition to analyzing the email header to detect the company impersonation



IMPORTANT NOTICE: 2024 Reviewed Employee Handbook- All Employees Compliance Required

Human Resources <hrnotice@[redacted].com>
To ○ Carolyn [redacted]
Tue 20/08/2024 14:21

This message was sent with High importance.
If there are problems with how this message is displayed, click here to view it in a web browser.

First time sender >
Contains topics of a financial nature >
This email shows strong signs of phishing >

Dear Employee,

Our employee handbook has been updated with new information on benefits, reimbursement for incurred expenses, a code of conduct, and clear consequences for serious infractions, like suspension or termination of employment.

Everyone must adhere to our company's standard code of behavior while dealing with customers, suppliers, and contractors, but more crucially when dealing with coworkers

Unlike previously, EVERYONE is required to electronically confirm that they have read and agree to these requirements by using the electronic HR compliance area below as soon as they receive this message..

Review&Submit Acknowledgment

The HR will circle back individually to any employee that fails to submit acknowledgment using the above to be sure if such employee does not comply with our policies.

Kindly inform us if you have any questions or concerns.

Regards,
Human Resources.

Impersonation tactics: Impersonating authoritative HR department. Spoofed email address increases credibility

Pretexting: Generates plausible scenario for the request

Baiting: Sensationalized clickbait topics that employees will want to learn about

Sense of urgency: Encourages targets to act quickly

Inclusive language: Increases social pressure

Fear: Threat of negative consequences for inaction

Generating This Attack With an AI Chatbot

We asked AI to write short email on the same subject matter — and its output uses many of the same tactics as the original, including urgent language and fear.

USER

Please write a short email from HR about updates to the employee handbook on benefits, reimbursement for incurred expenses, a code of conduct, and clear consequences for serious infractions, like suspension or termination of employment. Use urgent language to request recipients log into the HR compliance area to accept the new policies asap.

ASSISTANT

Here's a draft email for HR with the requested content:

Subject: URGENT: Critical Employee Handbook Updates - Immediate Action Required

Dear [Company Name] Employees,

We have made important updates to the Employee Handbook that require your immediate attention and acknowledgment.

These updates cover several critical areas:

1. Benefits package revisions
2. New expense reimbursement procedures
3. Updated code of conduct
4. Clarified consequences for serious infractions, including suspension and termination policies

It is crucial that all employees review and accept these new policies as soon as possible. Please log into the HR Compliance Portal immediately to review and acknowledge the updated handbook.

Failure to accept these new policies within 48 hours may result in a temporary suspension of certain employee privileges. We cannot stress enough the importance of prompt action on your part.

If you have any questions or concerns, please contact the HR department directly.

Thank you for your immediate attention to this matter.

Best regards,

[Your Name] Human Resources Manager

The Limitations of Traditional Email Defenses Against Social Engineering Attacks

Almost every organization uses signature-based and reputation-based email security. It's what comes as standard in Microsoft Exchange Online Protection and is the detection capability also present in legacy SEGs.

So, getting through these defenses are simply the cost of doing business when it comes to phishing. Analyzing KnowBe4 Cloud Email Security platform data shows that 2024 saw a 51% increase in phishing attacks landing in the inbox.[8]

Cybercriminals can avoid detection by using tactics such as using compromised accounts, legitimate email services and domain ageing.

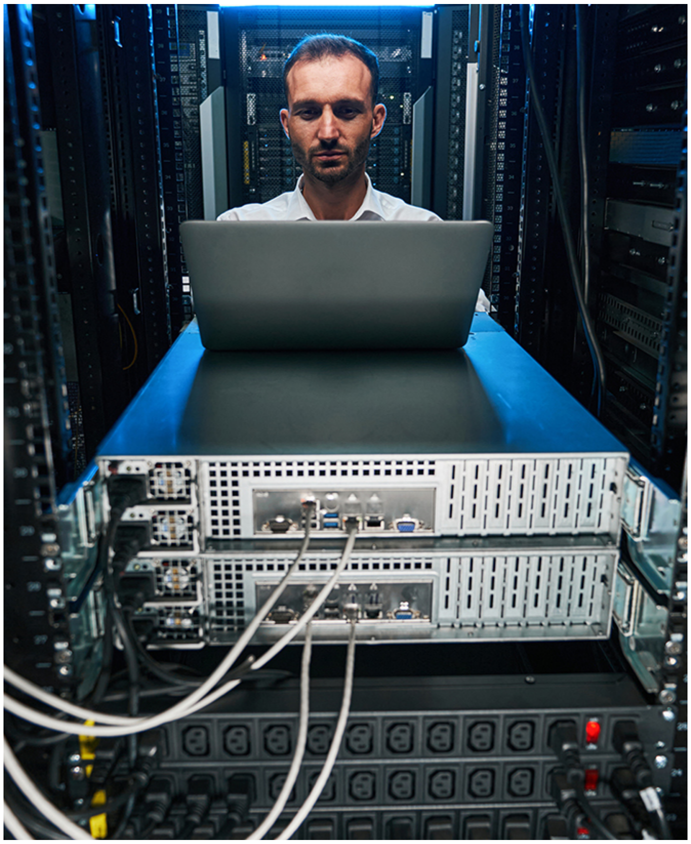
One surefire way to get through signature-based detection is to rely solely on social engineering and not include a traditional payload (malware or phishing hyperlink). Text-based attacks and those with "harmless" attachments (i.e. anything not classified as malware, such as fraudulent invoices) won't raise red flags in legacy or SEGs provided they can also bypass reputation-based checks. While many legacy systems use weighted combinations of email content (e.g. display name and subject line and message body key words) to detect impersonation-based phishing emails, these methods are only useful in the most rudimentary of attacks. As we've seen, simple prompts into AI engines can result in well-written attacks.

Where a traditional payload is used, definitions libraries take time to update which gives cybercriminals opportunity to bypass detection. While historically it has been resource-intensive to create new malware payloads, GenAI is reducing this burden and also making it easier to create text for new phishing websites.

Authentication protocols (DMARC, DKIM, and SPF) can help to detect spoofed emails in impersonation-based social engineering attacks by authenticating the sending domain and address. Similarly, Microsoft and SEGs can block emails sent from newly created or poorly rated domains. However, cybercriminals can avoid detection by using tactics such as using compromised accounts, legitimate email services and domain ageing.

Finally, multi-channel phishing campaigns aim to move their targets away from email and into another platform that has fewer controls and visibility, such as Zoom, Microsoft Teams, and SMS, and in combination, these attacks can be harder to detect.

With more phishing attacks getting through these SEG defenses, organizations urgently require more sophisticated phishing detection technology that leverages other analysis techniques to identify a broader range of threats.



Strategic CISO: Detecting Social Engineering With AI-Enabled Defenses

With more phishing attacks bypassing the signature-based and reputation-based detection used by native and/or legacy email defenses and SEGs, organizations need to implement an AI-powered layer to their defenses. Specifically, these controls must be able to detect the advanced social engineering attacks that people are particularly susceptible to.

Integrated cloud email security (ICES) platforms offer intelligent, behavioral-based threat detection that can identify the full spectrum of inbound phishing attacks, including socially engineered phishing emails – regardless of whether they're written by a person or a bot, or a bit of both!

Natural language processing (NLP) and natural language understanding (NLU) enable an ICES platform to analyze email content and attachments to identify linguistic identifiers of social engineering (such as urgent language) and the organizational context behind the words being used (for example, to detect CEO fraud). Additionally, the analysis can also highlight unusual requests or language that frequently appears in pretexting, such as financial phrasing. The platform should also detect and neutralize any obfuscation techniques implemented to prevent NLP and NLU working correctly.

Analysis of the sender's display name and the sending email address (including within the email header) enables the technology to detect email impersonation tactics that are used in social engineering. Analyzing a company's accepted domains, internal systems, and brand can also identify third parties impersonating their brand.

Finally, ICES platforms deeply understand how each individual employee uses email, such as which vendors and contacts within the supply chain a person communicates with. By analyzing both these patterns and the sending domain (e.g. through authentication protocols and domain age), the technology can detect phishing attacks sent from compromised email accounts, including those owned by trusted vendors.

By combining all these capabilities, an ICES platform offers powerful detection for even the most sophisticated socially engineered phishing attacks.

1 Data Breach Investigations Report, Verizon, May 2024.

2 Phishing Threat Trends Report, Vol. 3, Egress, April 2024.

3 'Use Behavioral Economics to Influence Security Behavior and Individual Decisions', Gartner.

4 Data Breach Investigations Report, Verizon, May 2024.

5 Thinking Fast and Slow, Kahneman, Daniel, 2012.

6 Data Breach Investigations Report, Verizon, May 2024.

7 https://pmc.ncbi.nlm.nih.gov/articles/PMC7252086/.

8 Phishing Threat Trends Report, Vol. 3, Egress, April 2024.

Detect and Neutralize Social Engineering Phishing Attacks With KnowBe4 Defend

Learn how KnowBe4 Defend uses AI-powered technology to detect social engineering tactics

Request Demo

See How

KnowBe4 Cloud Email Security

KnowBe4
Defend

Detect and defend against targeted phishing attacks

Learn More

KnowBe4
Prevent

Stop data breaches before they happen

Learn More

KnowBe4
Protect

Send and receive secure, encrypted email

Learn More

Inbound threat protection

Outbound threat protection

About KnowBe4

KnowBe4 empowers workforces to make smarter security decisions every day. Trusted by over 70,000 organizations worldwide, KnowBe4 helps to strengthen security culture and manage human risk. KnowBe4 offers a comprehensive AI-driven 'best-of-suite' platform for Human Risk Management, creating an adaptive defense layer that fortifies user behavior against the latest cybersecurity threats. The HRM+ platform includes modules for awareness & compliance training, cloud email security, real-time coaching, crowdsourced anti-phishing, AI Defense Agents, and more. As the only global security platform of its kind, KnowBe4 utilizes personalized and relevant cybersecurity protection content, tools and techniques to mobilize workforces to transform from the largest attack surface to an organization's biggest asset.

KnowBe4

KnowBe4, Inc. | 33 N Garden Ave, Suite 1200, Clearwater, FL 33755
855-KNOWBE4 (566-9234) | www.KnowBe4.com | Sales@KnowBe4.com