PHISHME COFENSE

The Future of Phishing Defense:

AI AND HUMAN COLLABORATION

Executive Summary

In the modern-day cybersecurity environment, AI-driven anti-phishing solutions operating on autopilot may seem highly appealing. With the growing cybersecurity skills gap, security operations center (SOC) teams are increasingly turning to automation.

While AI systems provide advantages in speed and efficiency, they often struggle with detecting nuanced phishing threats and maintaining transparency.

A more balanced approach—combining controlled automation with human oversight—bridges this gap. By integrating human-vetted threat intelligence, this approach enhances accuracy, ensures data control, and effectively identifies and mitigates malicious emails that may slip past traditional AI-powered defenses.

The Push for More Automation: Balancing Efficiency and Security

With SOC teams currently facing increasing pressure to manage heavier workloads amid a widening cybersecurity skills gap, the demand for automation in phishing defense solutions has surged. At the same time, the rise of AI has ushered in an entirely new level of automation capability.

While AI-driven solutions provide much-needed efficiency, they also introduce challenges in maintaining the accuracy and transparency required to address sophisticated phishing threats effectively.

Balancing efficiency and security in phishing defense requires a thoughtful integration of AI and human expertise.

While AI excels at processing vast amounts of data quickly, identifying patterns, and automating repetitive tasks, it can struggle with detecting context-specific phishing attempts.

To close this gap, organizations should adopt a hybrid approach that combines the speed and scalability of AI with the critical thinking and contextual awareness of human analysts. By leveraging AI to handle high-volume, low-complexity tasks and reserving human expertise for more intricate threat analysis, organizations can achieve both operational efficiency and robust security.

This allows SOC teams to focus on high-priority threats that require deeper investigation and strategic decision-making, rather than being overwhelmed by routine tasks.

Limitations of AI in Phishing Defense

While automation is beneficial and often necessary, complete automation when it comes to phishing protection is not always the best approach.

Human oversight is necessary to detect subtle nuances, context, and intent that AI might overlook. AI models are only as good as the data they are trained on.

Therefore, if threat actors innovate a new method of phishing, an AI-powered solution might fail to detect it simply because of the lack of representative training data and the reliance on pattern recognition. AI solution vendors often address this by highlighting “48-hour” intervals for their model updates, but in the fast-paced world of phishing, 48 hours is more than enough time for attackers to exploit vulnerabilities.

Many users of these AI-based phishing defense solutions select them because they believe “it catches everything.” Since these solutions catch more malicious emails than traditional secure email gateways (SEGs), users assume they’re catching everything.

However, in our own research we have found that these systems are, in fact, missing threats. The issue lies in the “False Negative” problem: these solutions are unable to identify or report threats that go undetected.

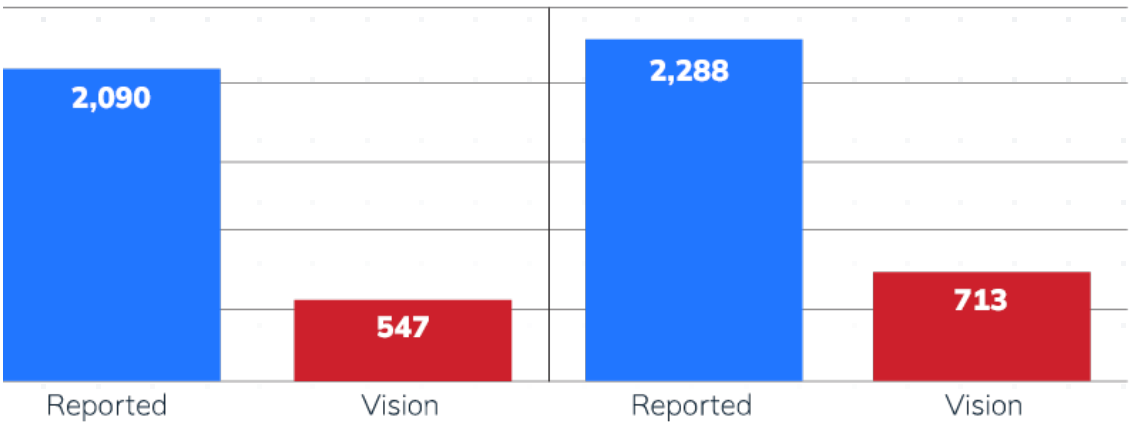
These identified malicious emails were neither reported or detected by the AI-based ICES system. The increase in undetected threats, compared to pre-ICES implementation, highlights the unknown and undetected threats organizations continue to face.

This example highlights an important point: while modern-day SEG and ICES solutions are highly effective, they are not infallible and can still let dangerous threats slip through. Bridging the gap in phishing defense requires a collaborative and innovative approach.

Human analysis plays a crucial role, as trained users can detect nuanced, context-specific signs of phishing that AI might overlook, providing a stronger and more adaptive defense.

BEFORE ICES IMPLEMENTATION

2 MONTHS AFTER ICES IMPLEMENTATION



Before implementing an ICES, employees reported an average of 2,090 emails per month. ***These customers also implemented Cofense Vision, which quarantined an additional 547 malicious emails that were undetected and unreported.***

After implementing an ICES, employees reported an increased average of 2,288 emails per month and ***Cofense Vision quarantined an additional 713 malicious emails that were previously undetected and unreported.***

Striking the Balance: Human Oversight Meets AI Efficiency

One of the most effective strategies for enhancing cybersecurity within your organization is to foster a strong reporting culture.

Frequent and consistent reporting allows for early threat detection and enables faster incident response. However, reporting malicious emails is just the first step; you also need a reliable way to analyze these reported emails. Many vendors are turning to AI to automate this process.

Unfortunately, this approach shares the same limitations as in-line gateway filtering. User-reported emails, initially missed by the in-line technology, are sent to an abuse mailbox where AI technology attempts to identify malicious content.

While this adds another layer of detection, it relies on the same flawed foundational technology that previously failed to catch these threats, and some malicious threats will evade detection by the AI/ML models being used.

Additionally, these solutions often produce a high rate of false positives, overwhelming security teams with unnecessary alerts. This constant barrage forces teams to spend countless hours sifting through benign reports, diverting attention from real, high-priority threats.

Meanwhile, the longer threats linger in employee inboxes, the greater the organization's risk of a successful attack.

Unpacking the Black Box: Transparency in AI-Powered Cybersecurity

Another area that customers who are considering an AI-powered anti-phishing solution should review is understanding how these systems make decisions and the data required to do so.

Many vendors downplay the critical issue of transparency, leaving customers questioning their data protection. For security teams, many of these solutions operate as “black boxes,” making decisions that are difficult to explain or audit.

This lack of transparency prevents teams from assessing the system’s accuracy or fairness, leaving them unable to fully trust or validate the technology they rely on to protect their organization.

This lack of clarity can be a serious issue in regulated industries or when customers demand accountability and transparency.

Many AI systems rely on complex algorithms that make it challenging to trace the reasoning behind their decisions. One commonly used method is called “Social Graph Analysis,” where AI collects and analyzes data about an organization’s users, such as their communication habits, key contacts, and typical patterns of interaction.

Then, if an email arrives that does not meet the usual criteria or usage patterns for that user, it can be flagged, tagged, or even removed as potentially malicious.

However, this strategy requires the collection and processing of large amounts of data about each user, raising serious concerns about data security, sovereignty, and user consent.

Questions arise about whether users should be subjected to such extensive data collection by a third party, particularly in ways they may not fully understand or control.

The use of such large amounts of personal user data to define what constitutes "normal behavior" is likely to be an issue for many enterprise customers.

Furthermore, poorly managed AI could violate privacy laws, such as the General Data Protection Regulation, which require transparency and accountability in the collection and processing of data, as well as the implementation of security measures to protect personal information.

Who Controls Your Data?

In addition to the lack of transparency, organizations often relinquish control over their data once it is processed by the vendor.

For example, when a user reports a suspicious email and it is analyzed by AI, the vendor typically assumes control over how that data is used, stored, and potentially shared with other customers. This shift in control can leave organizations unable to dictate or monitor how their sensitive information is handled, creating significant risks.

Organizations should have the flexibility to choose the solution that best meets their needs. Whether they opt for a fully on-premises, customer-hosted and managed solution for post-inbox analysis and remediation or a hosted, fully managed service, they should be able to retain full visibility.

A well-designed managed service should allow clients to log in anytime, review how their data is being managed, and monitor how the service is being run.

Conclusion

Effective phishing defense requires more than just speed and accuracy—it demands a strong focus on data transparency and control.

The most effective strategies combine human expertise with advanced technology, while maintaining control over sensitive data. This approach not only enhances threat detection and remediation but also ensures that sensitive information is managed responsibly and securely.

Cofense stands at the forefront of phishing defense, offering a powerful combination of human-supervised AI and advanced technology to help organizations combat evolving threats. With a global network of over 35 million trained end users, our solutions provide real-time detection and analysis of malicious emails, keeping organizations ahead of attackers.

Beyond technology, Cofense prioritizes employee education, equipping teams with the skills to identify and report phishing attempts effectively. This integrated approach—blending cutting-edge innovation with human expertise—not only enhances detection capabilities but also empowers employees as a critical line of defense.

Together, we build a resilient security culture that protects organizations from even the most sophisticated phishing attacks with speed, precision, and adaptability.

Want to learn more about how Cofense can enhance your phishing defense strategy with controlled AI automation? **Contact us today.**

Cofense is the only cybersecurity company leveraging expert-supervised AI for phishing detection and response—delivering human-vetted intelligence and real-world training to help enterprises stay ahead of modern threats. Built to augment existing email defenses, Cofense identifies attacks that bypass perimeter filters, remediates them in minutes, and continuously strengthens the human layer through simulations modeled on active phishing campaigns. Informed by insights from over 35 million trained users, Cofense enables faster containment of threats and measurable reductions in risk. Organizations like Mastercard and Blue Cross Blue Shield rely on Cofense to reduce exposure, meet regulatory demands, and build lasting resilience against the most persistent cyber threat: phishing.

Smarter phishing defense. Stronger human security.