IRONSCALES
SAFER TOGETHER

Why AI Alone is Not Enough

WHITE PAPER

Abstract

In the rapidly evolving field of cybersecurity, the sophistication of attacks, especially those leveraging generative AI (GenAI), presents an unprecedented challenge.

This whitepaper explores the intrinsic limitations of relying solely on Artificial Intelligence (AI) and Machine Learning (ML) for email security and underscores the indispensable value of human insights in augmenting these technologies.

Through a detailed examination of the IRONSCALES platform, we demonstrate how the integration of AI with collective human intelligence forms an adaptive defense mechanism against the advanced email attacks organizations see every day.

We explore the complexities of AI and ML in email security, highlighting their roles in detecting threats and the challenges they face, including the adaptability of attackers who now use GenAI to craft more convincing and personalized attacks. This whitepaper introduces IRONSCALES pioneering approach, which combines the scalability and precision of AI with the nuanced understanding of human insights.

Furthermore, we illustrate the power of crowdsourcing human insights across a network of over 13,000 organizations, enriching the AI models with a diverse range of experiences and expertise. This collaborative model not only accelerates the adaptation of security measures to new email threats but also significantly reduces the occurrence of false positives and negatives, optimizing the user experience.

“Why AI Alone is Not Enough” argues for a nuanced approach where AI-driven automation provides efficiency and scalability, while human expertise offers depth and adaptability.

This framework ensures an adaptive, dynamic, and proactive system capable of addressing the advanced attacks we see today and tomorrow.

TABLE OF CONTENTS

INTRODUCTION

Our reliance on technology and data has never been more profound, nor the threats against it more cunning and complex.

In the face of these challenges, nearly all organizations (90.5%) now augment their defenses with one or more AI-enabled email security solutions, recognizing that the basic protections offered by cloud email providers like Microsoft 365 or Google Workspace are insufficient. Email security stands as a critical front in the battle against cyber threats, a domain where Artificial Intelligence (AI) and Machine Learning (ML) have become indispensable. These technologies analyze patterns, detect anomalies, and shield users from the onslaught of deceptive tactics, including the pervasive and ever-evolving phishing attacks.

Yet, despite their sophistication, AI and ML are not the cure-all for cyber security. The intricate fabric of human communication and deception presents challenges that pure algorithmic approaches can struggle to navigate. AI, with its pattern-based logic and predictive prowess, occasionally falters in the face of human ingenuity and duplicity. The adversaries in cyber space are not static; they learn, adapt, and constantly devise new strategies to penetrate defenses, creating a moving target that AI alone can struggle to keep pace with.

Within this evolving environment, the indispensable nature of human insight is unmistakable. The discernment, instinct, and flexibility of human analysts are pivotal in catching and addressing threats that AI might miss, creating a powerful partnership between human insight and AI capabilities that strengthens our defense against complex dangers in our emails.

This white paper will explore the power of combining AI/ML and human insights to make email secure. We will introduce the foundational concepts of how AI and ML operate in this context, the role of variables and features in the models used, and crucially, how human insights are integrated to enhance accuracy and responsiveness. Through the lens of IRONSCALES, we will exemplify how human-augmented AI is not just a theoretical ideal, but a practical solution implemented across thousands of organizations for superior protection against increasingly sophisticated email threats.

Embark on this exploration with us as we articulate why AI, while powerful, is not a standalone solution, and how the human element remains an essential component of the most effective cyber security strategies.

SECTION 1

Understanding AI and ML in Email Security

Email, one of the most ubiquitous forms of digital communication, has been a prime target for exploitation for decades. Protecting this vector demands the most advanced defenses—enter Artificial Intelligence (AI) and Machine Learning (ML). The prevalence and financial impact of Business Email Compromise (BEC) attacks alone exceeded $50 billion globally in 2022 , underscoring the critical need for these advanced technological defenses in safeguarding digital communication channels.

The Role of AI and ML in Detecting Threats

AI has rapidly become an indispensable component in modern day email security. It leverages algorithms to analyze vast volumes of data, discern patterns, and predict potential threats with speed and accuracy that far exceed human capability. Machine Learning, a subset of AI, goes a step further by using statistical techniques to enable computers to 'learn' from past data. This learning process empowers ML models to improve their threat detection capabilities over time autonomously.

How ML Models Function

ML models operate by processing input data—emails, in this case—through algorithms that weigh various factors known as 'features.' A feature is a distinct attribute or property of the data that the model uses to make decisions. In email security, features could include the sender's email address, the time the email was sent, specific keywords, URLs contained within the message, and even the email's structure.

Variables and Features, The Building Blocks of ML

Variables are the individual data points that the model examines. They can be anything from the characters in an email's subject line to the frequency of communication between two email addresses. Features, on the other hand, are engineered from variables and are specifically selected for their relevance to the task at hand—in this case, identifying phishing attempts, spam, and other malicious content.

The relationship between variables and features is one of transformation and refinement. Through a process known as 'feature engineering,' raw data (variables) are transformed into formats (features) that the ML model can interpret and learn from. This could involve extracting the domain of an email sender, analyzing the sentiment of the email content, or detecting anomalies in email-sending patterns.

Machine Learning Challenges in Email Security

Despite the capabilities of AI and ML, there are inherent challenges that they face in the realm of email security:

- **Adaptation of Threats:** Cyber threats are constantly evolving, with attackers finding new ways to mimic legitimate communication patterns (using popular GenAI tools), thereby evading detection.
- **Data Overload:** The sheer volume of data to process can be overwhelming, leading to potential oversight and misclassification—criminals leverage automation tools to scale their attacks extensively, amplifying this challenge.
- **False Positives and False Negatives:** Distinguishing between legitimate emails and threats is not always clear-cut, leading to false alarms (false positives) or missed detections (false negatives).

The Limitations of AI and ML

AI and ML models lack the innate human ability to perceive context, intent, and subtle cues that might indicate a threat. As a result, they can sometimes be deceived by cleverly disguised malicious emails that exploit the gaps in their programmed logic.

This section has established how AI and ML are utilized for email security, the importance of variables and features in ML models, and the challenges these technologies face. The next section will explore the indispensable value of human insights via frameworks such as Human-in-the-Loop (HITL) and Reinforcement Learning from Human Feedback (RLHF) that bridge the gap between algorithmic efficiency and human discernment.

No matter how advanced, AI and ML models are confined by the data they have been trained on and the features they have been programmed to recognize.

SECTION 2

The Value of Human Insights in AI/ML Frameworks

As AI and ML models become increasingly adept at filtering email-based threats, attackers are not standing idle. They are harnessing AI's power to craft more compelling and convincing attacks, particularly Business Email Compromise (BEC) and socially engineered schemes. This use of AI by bad actors creates messages that mimic legitimate communication more closely than ever, challenging security systems to keep pace.

The use of AI by attackers has turned it into a double-edged sword. On one hand, it provides security vendors with powerful tools to analyze and predict threats. On the other, it equips attackers with the means to automate and refine their attacks, creating a dynamic and escalating threat environment. For instance, AI-powered programs can generate contextually relevant content that can bypass traditional filters and engage victims with personalized messages that appear entirely credible.

The Human Element - Discerning What AI Cannot

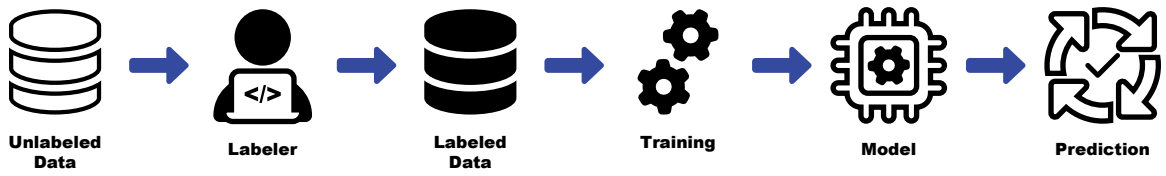
In the intricate "game" of cybersecurity, human insights emerge as a pivotal force—capable of detecting subtleties that that AI alone cannot predict or counteract. Humans bring context-sensitive analysis and emotional intelligence to the table—factors that are currently beyond the reach of AI. For example, AI might struggle to interpret the nuanced intent behind a seemingly benign email that's actually a prelude to a BEC attack, while a human could spot the red flags raised by the message's context and subtle cues.

Integrating Human Insights With AI/ML

The Human-in-the-Loop (HITL) approach is the embodiment of integrating human expertise with automated processes. Unlike standard supervised learning, where the AI model is trained on pre-labeled data and then applied to make predictions, the Adaptive AI Active Learning framework with HITL incorporates ongoing human feedback into the AI learning cycle. This allows human analysts to review, override, or validate AI decisions, ensuring that the final judgment on an email's legitimacy combines the best of AI efficiency with human discernment. Let's compare the two frameworks.

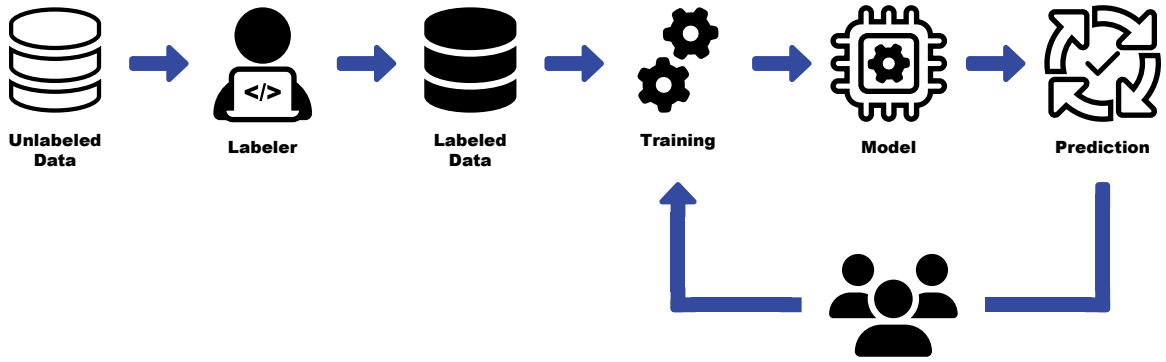
Standard AI Supervised Learning

In traditional supervised learning, AI models are trained on datasets that have been previously labeled, relying on historical data to predict future outcomes. While this approach can effectively identify known threats, it may not be as quick to adapt to new, sophisticated attacks that have not been seen before. This limitation can result in slower responses to emerging threats as the AI depends on the occurrence and subsequent labeling of new data types to learn and apply that knowledge to future predictions.



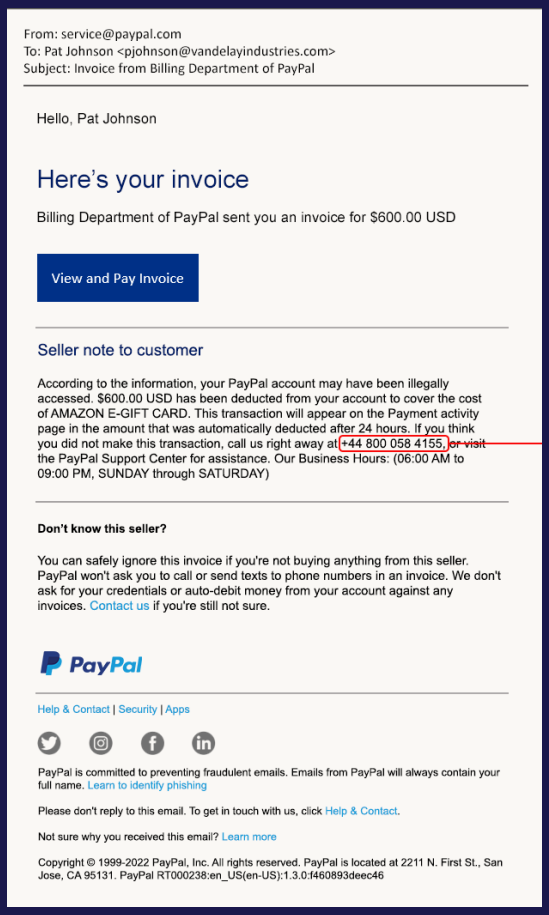
Adaptive AI Learning with Humans-in-the-Loop

In the HITL framework, when employees flag emails that seem suspicious, these escalations become critical data points. Human security experts review these escalations, providing nuanced insights that actively retrain and refine the AI, as shown in the ‘Adaptive AI Active Learning with Humans-in-the-Loop’ diagram.



Crowdsourced Intelligence

Compared to a standard AI learning processes that can be limited by the scope of their training data, the HITL approach at IRONSCALES is enhanced by crowdsourcing insights across a vast network of over 13,000 organizations. As illustrated in the ‘Adaptive AI Active Learning’ image, this collective intelligence rapidly informs the model’s learning, incorporating a broad range of human perspectives and experiences, and quickly adapting to new attacker tactics.



From: service@paypal.com
To: Pat Johnson <pjohnson@vandelayindustries.com>
Subject: Invoice from Billing Department of PayPal

Hello, Pat Johnson

Here's your invoice

Billing Department of PayPal sent you an invoice for $600.00 USD

View and Pay Invoice

Seller note to customer

According to the information, your PayPal account may have been illegally accessed. $600.00 USD has been deducted from your account to cover the cost of AMAZON E-GIFT CARD. This transaction will appear on the Payment activity page in the amount that was automatically deducted after 24 hours. If you think you did not make this transaction, call us right away at +44 800 058 4155, or visit the PayPal Support Center for assistance. Our Business Hours: (06:00 AM to 09:00 PM, SUNDAY through SATURDAY)

Don't know this seller?

You can safely ignore this invoice if you're not buying anything from this seller. PayPal won't ask you to call or send texts to phone numbers in an invoice. We don't ask for your credentials or auto-debit money from your account against any invoices. Contact us if you're still not sure.

PayPal

Help & Contact | Security | Apps

PayPal is committed to preventing fraudulent emails. Emails from PayPal will always contain your full name. Learn to identify phishing

Please don't reply to this email. To get in touch with us, click Help & Contact.

Not sure why you received this email? Learn more

Copyright © 1999-2022 PayPal, Inc. All rights reserved. PayPal is located at 2211 N. First St., San Jose, CA 95131. PayPal RT000238:en_US(en-US):1.3.0:f460893deec46

CASE STUDY
The PayPal Phishing Email

Consider a sophisticated phishing scheme that masqueraded as a legitimate PayPal invoice communication. The deceptive emails originated from PayPal's own domain, with links directing to the actual PayPal website. However, a subtle but significant anomaly was present, a request for recipients to call a number to dispute a charge. This anomaly, which could potentially be missed or dismissed by AI, was a red flag that prompted a recipient's suspicion, and subsequent escalation and administrative review.

Everything else in the email is genuine, the meta data, authentication and reputation checks, the consistency of the template, logos, colors, and hyperlinks—because it is sent from within the PayPal merchant portal. All these features weigh heavily in the ML analysis.

Integrating Advanced ML Techniques With Human Insights

In response to such intricate attacks, our email security systems must leverage both advanced ML techniques and human insights:

- **Enhanced Feature Set:** We enrich our ML models with a broader set of features, such as the presence and contextual relevance of phone numbers, enabling them to identify potential threats even within legitimate-looking emails.
- **NLP Enhancements:** By employing sophisticated NLP, our systems can detect subtle language patterns and anomalies that may indicate a phishing attempt, despite the email passing standard authentication protocols.
- **Dynamic Feature Weighting:** With insights gathered from human reviews, our models can dynamically adjust the weighting of certain features. If a security admin flags a phone number as suspicious, the model can be tuned to scrutinize similar patterns more closely in future assessments.
- **Specialized Training for Anomalies:** Real-world examples like the PayPal phishing email are used to train the AI system, focusing on the specific anomalies that led to human detection. This specialized training enhances the model's ability to detect similar fraudulent attempts.

The HITL framework serves as a vital component in the ongoing development of our email security system. It allows human experts to augment ML predictions, ensuring that the nuanced understanding of cybersecurity professionals continually informs and improves the accuracy of our AI models.

SECTION 3

The Symbiosis of Advanced AI and Collective Human Vigilance

IRONSCALES stands at the forefront of email security, leveraging advanced AI and ML technologies to autonomously intercept nearly 100% of malicious emails—while our pioneering integration of human insights enhances our security framework. This combination fosters a superior defense mechanism, outperforming what either humans or AI could accomplish alone, by blending the precision of technology with the nuanced understanding of human expertise.

Advanced AI at the Core

At the core of the IRONSCALES platform is a sophisticated AI engine capable of analyzing and intercepting the vast majority of email threats with precision and speed. The high level of accuracy the IRONSCALES platform achieves in intercepting email threats is a direct result of its sophisticated AI, which evolves continuously through machine learning. Named 'Adaptive AI,' it thrives on a rich diet of data, constantly updating its algorithms and integrating human insights to refine its efficacy.

- **Proactive Threat Detection:** IRONSCALES employs proactive threat detection mechanisms—powered by Adaptive AI—that predict and prevent attacks before they can do harm, significantly reducing the dependency on human intervention.
- **Autonomous Response:** Utilizing behavioral analysis, anomaly detection, and other advanced techniques like Natural Language Processing (NLP) and Computer Vision—the system autonomously responds to threats, continually improving through machine learning.

The Human Layer, a Strategic Enhancement

While the IRONSCALES Adaptive AI provides robust protection, it is understood that, like all AI-driven solutions, it may occasionally encounter false negatives. In these instances, the human layer acts as a strategic enhancement, stepping in on the rare occasions when a never-before-seen attack slips through. This human layer is not a primary line of defense but a complementary force, adding depth and resilience to the system.

- **Employee Empowerment:** By empowering protected employees to report suspicious emails, IRONSCALES ensures that even the most subtle threats are captured, enhancing the overall security posture for their organization.
- **The Collective Expertise of Security Practitioners:** Beyond individual employees, over 30,000 security practitioners across 13,000+ organizations contribute their expertise, forming a vast network of human intelligence that enriches the AI models.

Crowdsourcing, a Force Multiplier

The collective power of human insights acts as a force multiplier for the IRONSCALES platform. Each reported anomaly and each confirmed threat serve to refine and update the Adaptive AI models, making them smarter and more resilient.

Rapid Threat Sharing

When a new threat is identified, details are shared in real-time across the network, turning individual insights into community protection.

Prevision Tuning

The feedback from this expansive user base allows for precision tuning of the AI, minimizing the occurrence of false positives and negatives, and ensuring a seamless user experience.

SECTION 4

The Synergy of AI, ML, and Human Expertise

The dynamic and often unpredictable nature of cyber threats necessitates a security solution that combines the speed and scalability of AI with the discernment and adaptability of human expertise. IRONSCALES exemplifies this synergy, operating on the cutting edge of cybersecurity to protect against email threats with unparalleled effectiveness.

Balancing AI Automation With Human Oversight

The IRONSCALES strategy hinges on creating a balanced ecosystem where AI-driven automation handles the bulk of threat detection and prevention, while human oversight acts as a nuanced layer of verification and improvement.

- **Automated Efficiency:** AI systems work tirelessly, analyzing millions of data points and detecting threats with a level of efficiency unattainable by human teams alone.
- **Human Augmentation:** Human expertise complements these systems by providing contextual understanding and making judgment calls on edge cases that AI may find ambiguous.

Continuous Learning and Adaptation

The landscape of cyber threats is in constant flux, and so the defenses against them must be equally adaptive. IRONSCALES achieves this through a continuous learning loop that integrates human feedback directly into the AI models.

Real-Time Updates

As threats are identified and categorized by security professionals, these insights are fed back into the system, allowing for real-time updates to AI models.

Evolving with the Threat

The AI system evolves with each interaction, becoming increasingly proficient at identifying and responding to new types of attacks.

Reducing False Positives and False Negatives

A critical challenge in cybersecurity is to minimize false positives, which can lead to alert fatigue, and false negatives, which represent security risks. The IRONSCALES platform leverages its AI and human insight to fine-tune detection algorithms continuously.

- **Contextual Calibration:** By analyzing the decisions made by human security experts, AI models learn to calibrate their detection mechanisms, improving accuracy over time.
- **User Experience Optimization:** Keeping false positives low ensures that users maintain trust in the system and engage with security protocols, reinforcing the overall security culture.

A Testament to Synergy

The PayPal phishing case study discussed in Section 3 is a testament to the power of this synergistic approach. While the AI system provides robust defense against most attacks, the additional layer of human insight was crucial in identifying a sophisticated scam that could have otherwise gone unnoticed.

As cyber attackers increasingly employ AI in their tactics, the collaboration between AI and human expertise will become ever more critical. IRONSCALES is at the forefront of this evolution, demonstrating how proactive defenses, underpinned by AI and enriched by human insights, are the cornerstone of effective cybersecurity.

CONCLUSION

The Future of Cybersecurity With Adaptive AI and Human Insight

As the digital landscape evolves, so too does the sophistication of cyber threats. The advent of generative AI (GenAI) marks a new frontier in the arsenal of cyber attackers, enabling them to automate and refine phishing campaigns and other malicious endeavors with unprecedented precision and personalization. The increased use of GenAI by attackers not only highlights the escalating complexity of threats but also underscores the need for equally sophisticated defense mechanisms. This trend is substantiated in a report by Osterman Research, “The Role of AI in Email Security,” which revealed 91.1% of IT Security leaders said that cybercriminals are already using AI in the email attacks they are experiencing at their organization . These attacks are more scalable, targeted, and difficult to detect, underscoring the imperative for adaptive AI and human insight in cybersecurity strategies.

The IRONSCALES Approach - Adaptive AI and Gen AI for Defense (And Offense)

IRONSCALES is at the forefront of cybersecurity innovation, leveraging Adaptive AI to neutralize existing threats and employing Generative Adversarial Networks (GAN) machine learning models for modeling and preparing for future attack methods. This advanced strategy anticipates how attackers might harness GenAI, enabling IRONSCALES to predict and prepare for future attacks before they can do harm. The platform’s use of GenAI extends the capabilities of Adaptive AI, creating a more resilient defense that evolves in lockstep with emerging threats.

The Human Element - An Irreplaceable Ally in the AI Arms Race

In this high-stakes environment, the human element becomes more crucial than ever. Human intuition and expertise provide a critical layer of insight, capable of identifying subtleties that AI might overlook. The collaborative ecosystem within IRONSCALES harnesses the collective intelligence of over 30,000 security practitioners, enriching the Adaptive AI with a depth of human understanding that ensures a nuanced and comprehensive defense strategy.

Facing Forward, the Synergy of Human Insight and AI Innovation

The future of cybersecurity will be characterized by the dual escalation of threat sophistication and defense innovation. As attackers utilize GenAI to craft more insidious and convincing attacks, defense mechanisms must not only keep pace but stay ahead. The integrated approach of Adaptive AI, empowered by GenAI and augmented by human collaboration, offers a blueprint for this future, ensuring that organizations can navigate the cyber threat landscape with confidence and agility.

For organizations worldwide, the message is clear: The integration of cutting-edge AI technologies with the strategic application of human expertise is no longer a luxury but a necessity. In the era of GenAI-powered cyber threats, adopting a security posture that anticipates future challenges is imperative. IRONSCALES represents this proactive and adaptive approach, providing not just a shield against today's threats but a dynamic, evolving defense for the cybersecurity challenges we'll see tomorrow.

As we venture beyond the current horizon of AI capabilities, the collaborative synergy between technology and human insight will define the effectiveness of our cybersecurity defenses. IRONSCALES exemplifies this future-ready approach, embracing both the power of Adaptive AI and the irreplaceable value of human expertise to secure the way we communicate.

i Osterman Research, The Role of AI in Email Security, 2023 (link)

ii FBI, Internet Crime Complaint Center (IC3) Report, 2023 (link)

iii Osterman Research, The Role of AI in Email Security (page 7), 2023 (link)

Secure Your Inboxes. **Unburden** Your Team. **Empower** Your People.

The IRONSCALES™ platform stops the most elusive BEC, ATO, and VIP attacks that breach perimeter defenses including native cloud-hosted email security controls. By combing AI and human insights from every mailbox user and 20,000+ analysts across the IRONSCALES network of global admins, IRONSCALES protects your organization where it matters most—in your user's inbox.

IRONSCALES
SAFER TOGETHER