IRONSCALES Overview – How it Works

Preface

This document outlines the IRONSCALES email security platform, detailing its integration with Microsoft 365 (M365) and Google Workspace (GWS), operational workflow of the AI and machine learning, and the importance of human intelligence in email security. It aims to provide clear information to ensure understanding of our advanced email protection solution.

Scope

This overview covers email security against phishing, BEC, and impersonation attacks. While our platform also protects against QR-code (quishing) and image-based attacks, these are not covered here. Additionally, integrated phishing simulation testing and security awareness training are not covered, you can learn more about those through our self-guided virtual tours.

1. Integration

Integrating IRONSCALES into your environment just takes a few minutes. Using native API integration, you can deploy our advanced email security within M365 and GWS without changing MX records, thus avoiding mail delivery disruptions common with traditional SEGs.

The same API enables automatic deployment of the IRONSCALES report phishing button to all employees, regardless of the endpoint.

2. Inbox-Level Protection

Our solution provides continuous security by operating directly within the inbox, unlike those that only inspect emails during transit.

3. Self-Learning

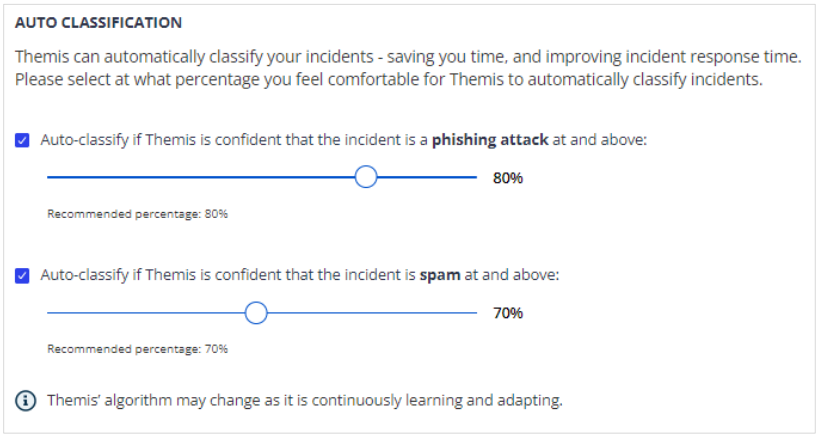
Once integrated, our AI begins a self-learning process, building a detailed social graph through natural language processing (NLP) and natural language understanding (NLU) to establish normal communication patterns for every employee inbox.

4. Continuous Threat Protection

Our Adaptive AI uses the social graph to perform real-time reputation, content, and behavioral analysis to detect the subtlest signs of malicious threats, including novel attacks. When a new email arrives, the AI assesses it based on predefined criteria: sender reputation, email content, embedded links, attachments, and other signals. Depending on its confidence [score], the AI automatically marks the email as safe or remediates it, moving it to a quarantine or spam folder.

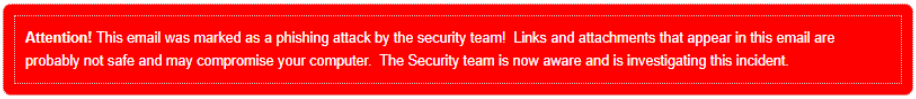
5. Automated Remediation

Our AI-powered MSOAR (Machine Security Orchestration Automation and Response) automatically classifies, clusters similar messages in your organization, and remediates threats, significantly reducing remediation time and enhancing security. Admins can customize the level of automation to their comfort levels.

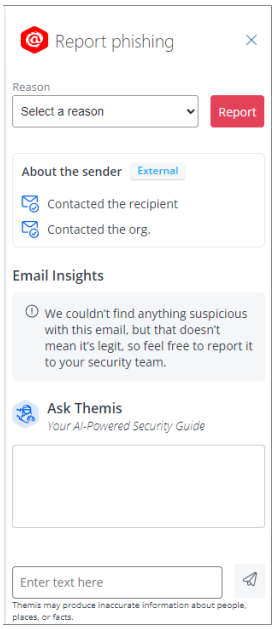


6. User Interaction

For emails where AI classification confidence score is below a customizable threshold, the email may get modified in the employees' inbox with (customizable) banners.

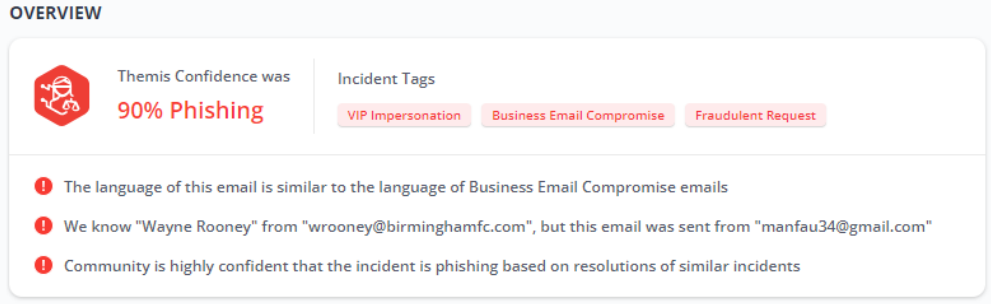


The IRONSCALES report phishing button allows employees to escalate suspicious emails to their IT Security team, empowering them to help identify a potential emerging attack. This is now critically important as adversaries are harnessing GenAI to develop more convincing, socially engineered attacks that can bypass many defenses.

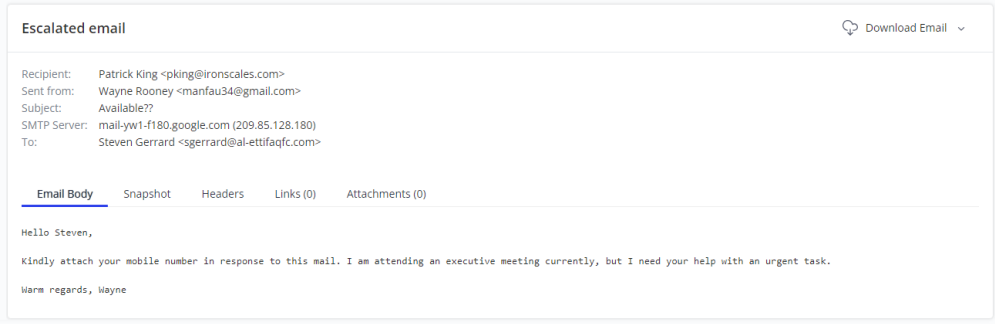


7. Escalated Review

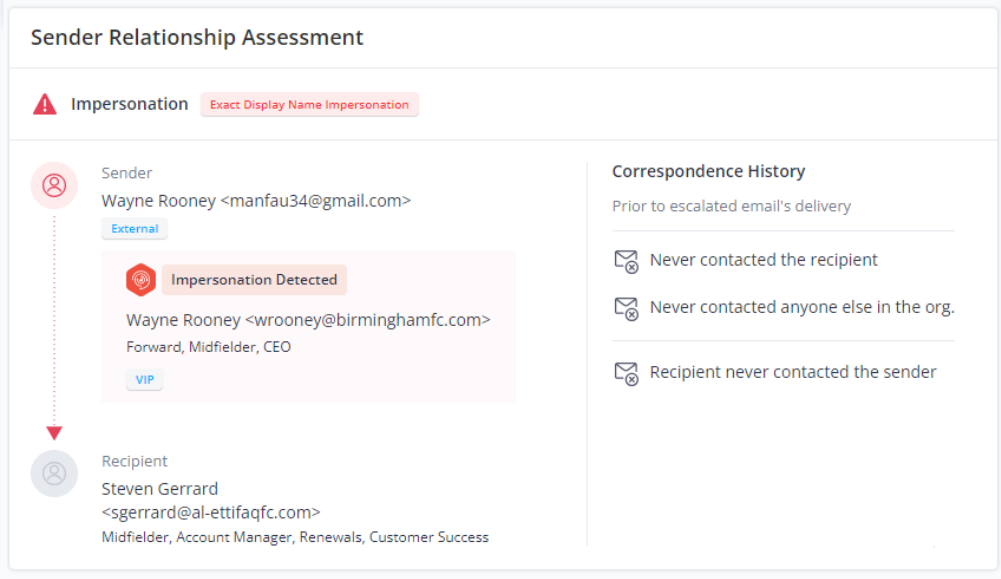
While over 99% of all email attacks are automatically detected, a few emails flagged by the AI or by employees, are escalated for a manual review by the IT Security administrators at each client company (take a step-by-step virtual tour here).



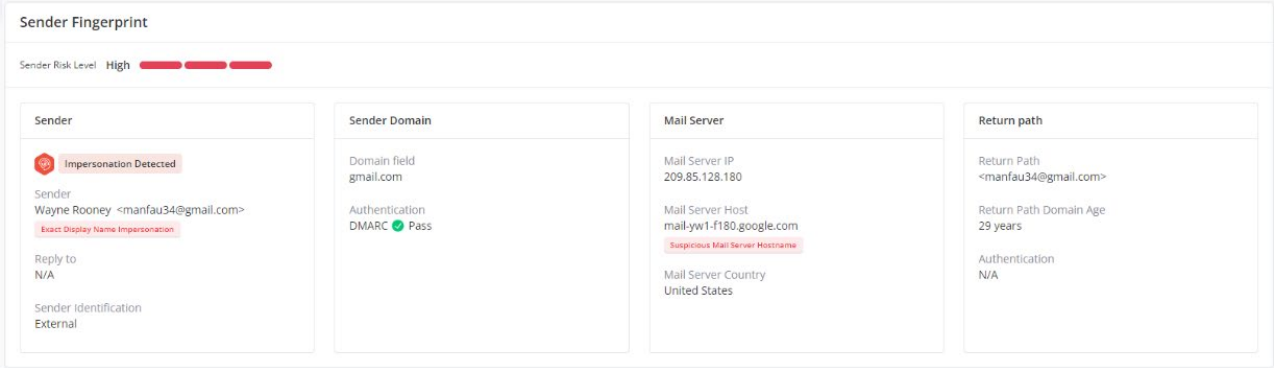
Security admins are provided everything they need to quickly assess escalated emails, diving into aspects like authentication, DMARC, reputation, headers, link, and attachment scanning results.



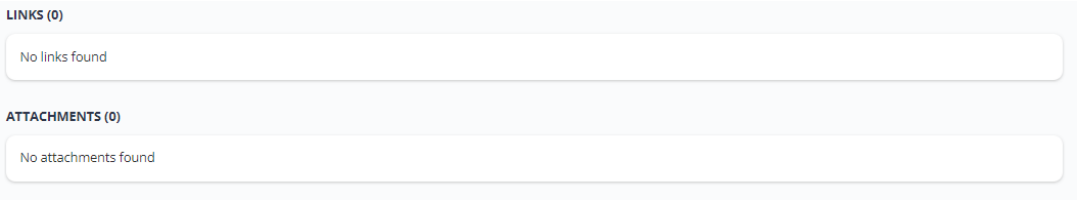
View the sender <> recipient relationship and correspondence history.



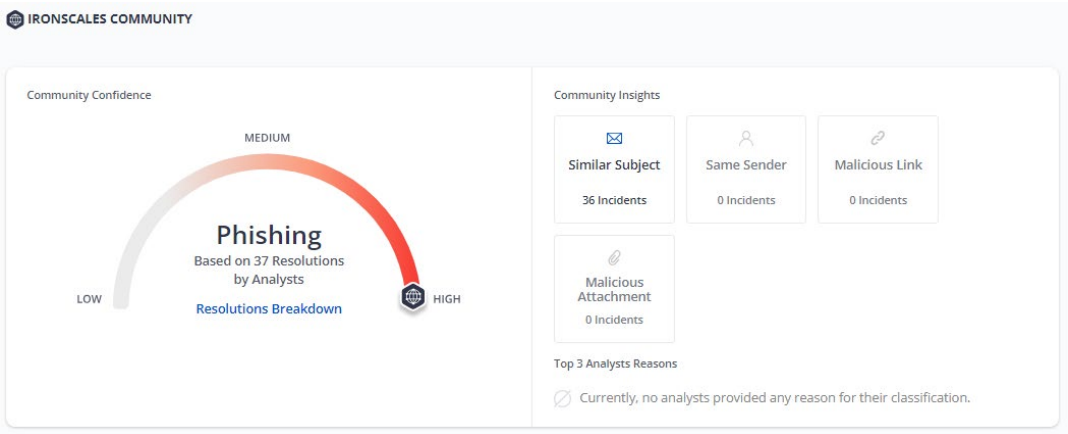
A “Sender Risk Level” is provided for every incident along with authentication results and mail server meta data.



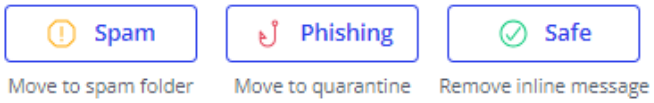
Although most email attacks with malicious links or attachments are instantly detected by our behavioral analysis, additional link and attachment scanning information is also provided (when applicable).



IT Security admins can also see how the same or similar message was assessed by a community of over 25,000 of their peers across the IRONSCALES network.



Based on the IT Security admin’s assessment, the email gets its final classification: phishing, safe, or spam.



8. IRONSCALES Analysts' Involvement

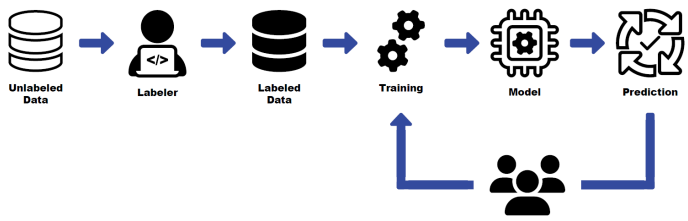
Alongside the community's line of site, data analysts at IRONSCALES investigate a subset of escalated emails, chosen through a blend of random and algorithmic processes. These analysts are highly trained experts that conduct deep investigations to classify emails, annotating each with tags and critical insights. This enriched data is vital, as it offers nuanced insights that help in refining the IRONSCALES machine learning models further.

9. Feedback Loop & Model Update

Classifications by the community and IRONSCALES analysts feed back into the machine learning model, central to our Human-in-the-Loop (HITL) process. This feedback loop, enriched by expert insights, continually refines our AI models.

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.



For a deeper understanding of our approach and the importance of integrating human insights with AI, read our full white paper: "Why AI Alone is Not Enough".

10. Communication Back to the User

After the comprehensive review process, the system or IT Security team may inform the end-user about the final classification of their escalated email. For instance, if an email they highlighted as suspicious is confirmed as phishing, they might receive a notification, fostering trust and reinforcing vigilance. Employees can optionally receive instant or scheduled notifications for emails removed from their inbox, and request that a specific email get removed from quarantine if they believe it is safe.

Summary of our Approach

Our email security platform integrates with Microsoft 365 and Google Workspace using API-based integration, providing continuous protection at the inbox level. Since 2018, we have been committed to maximizing the power of AI and have learned that AI alone is not enough. Our AI begins a self-learning process upon integration, establishing a detailed baseline to detect anomalous and malicious activity. The platform features intuitive warning banners and a report phishing button to enable employees to flag suspicious emails easily.

We automate the remediation of nearly all email threats, while offering a straightforward and informative experience for admins when reviewing occasional ambiguous emails. The system incorporates insights from employees, admins, and IRONSCALES analysts for continuous training of our Adaptive AI so we can detect never-seen-before threats.