SANS | Research Program

PRODUCT REVIEW

When Trusted Senders Become Threats: Stopping BEC and Supply Chain Attacks with Self-Learning AI

Written by **Matt Bromiley**

April 2026

DARKTRACE

Introduction

Cloud email security is easier said than done. Native email security controls were built for scale, not sophistication, and attackers know it. Time and time again, SANS sees the adversary's ability to thwart even the best laid plans and defenses. The challenge isn't a lack of tools; it's a fundamental gap in how some tools approach detection.

Native email-provider security catches commodity threats effectively. Attack-centric detection methods like signatures, reputation feeds, and known-bad indicators work well against known, high-volume attacks. Secure email gateways (SEGs) add another layer but operate on the same underlying methodology. Combining native security with an SEG often yields duplicated detection capabilities rather than true defense-in-depth. Sophisticated social engineering, business email compromise (BEC), and supply chain attacks exploit this gap. When a trusted vendor account is compromised or an adversary crafts a novel phishing approach, attack-centric tools tend to struggle.

This product review examines Darktrace / EMAIL an ICES solution that deploys in parallel to Microsoft 365 (M365) or Google Workspace as a last line of defense. Darktrace / EMAIL is part of the company's broader Active AI Security Platform, sharing its AI-driven core across network, cloud, and email products. While the review did not look at other components, current users of the Darktrace ecosystem should consider whether their existing tools cover the full spectrum of threats or simply duplicate detections already in place.

The integrated cloud email security (ICES) category was developed to address this issue. Unlike inline SEGs, ICES solutions deploy via API alongside existing controls, operating on fundamentally different detection methodologies that complement rather than duplicate.

Darktrace has a unique answer to this problem. Rather than training on historical attacks, Darktrace / EMAIL uses self-learning AI to establish what the company calls "patterns of life." These patterns establish behavioral baselines for each organization's internal users and external correspondents. Darktrace / EMAIL uses a "quick in, quick out" operational model, designed to minimize business impact in mature deployments where email security doesn't require dedicated full-time staffing.

This product review takes the perspective of an operational analyst examining how the platform operates, looks, and feels, and how readily it surfaces context on existing threats. Readers are encouraged to use the same lens, comparing the walkthrough against their current solution to gauge whether those capabilities are already within reach.

Architecture and Deployment

Darktrace / EMAIL deploys in parallel to M365 or Google Workspace rather than sitting in-line with mail flow. The solution receives email copies through journaling or API-based delivery, analyzing messages without introducing latency or creating mail flow dependencies. Because it operates as a security overlay rather than an email infrastructure component, organizations can deploy alongside existing controls without restructuring their email architecture. This eliminates the risk of Darktrace / EMAIL becoming a single point of failure in the mail path.

Platform Integration

Darktrace / EMAIL shares its AI core with the company's network and cloud security products, enabling what the vendor calls internal references. These are email decisions informed by activity observed elsewhere in the environment. For example, when the platform analyzed a suspicious link in an email, a rarity score was assigned to the link, reflecting not just email traffic but also internal network traffic. This cross-platform context surfaces behavioral signals that an email-only platform would miss. Again, although this review does not cover the full platform, it includes details on how this product integrates with existing Darktrace solutions to provide a more complete picture.

> For M365 environments in particular, journaling ensures the complete email corpus is available for behavioral analysis. This is essential for the self-learning AI to establish accurate baselines across all communication patterns in the organization.

Multichannel Coverage

The core email license covers inbound, outbound, and lateral email analysis. This is worth emphasizing because many email security tools focus almost exclusively on inbound threats. See Table 1 for a comparison of the three.

Inbound Email	Outbound Email	Latest Email
Email from external parties flowing into the organization. These are typical phishing emails but may also be abnormal given historical patterns.	Compromised accounts sending phishing to external contacts, data exfiltration attempts, and policy violations.	Internal email between employees, addressing scenarios where an adversary compromises an internal account and phishes colleagues—a technique that bypasses many email controls.

Table 1. Inbound, Outbound, and Lateral Email Analysis Comparison

Darktrace / EMAIL also goes beyond just emails. Microsoft Teams messaging support extends coverage into employee communications (see Figure 1).

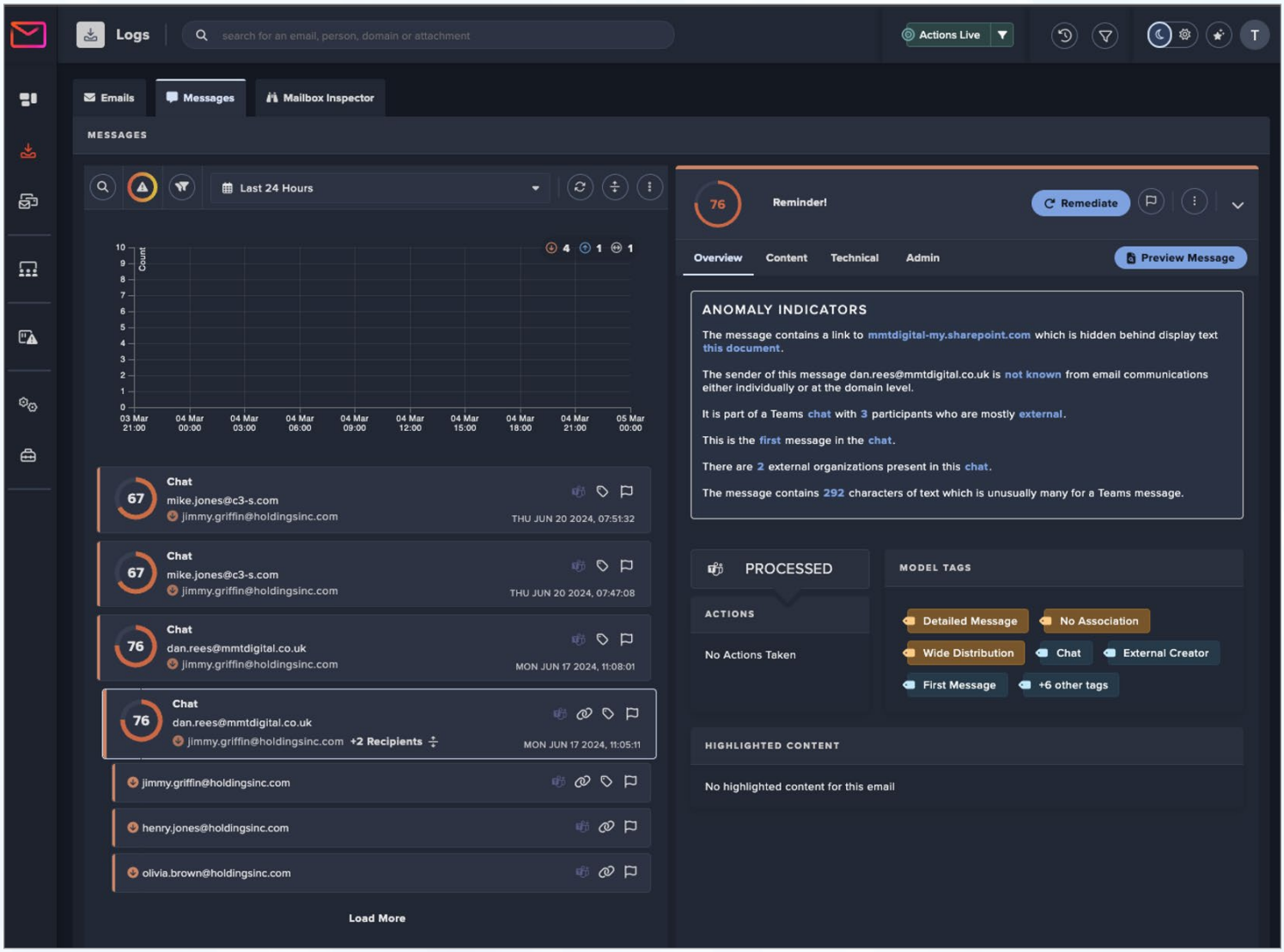

Figure 1. Messages Integration

Even in a brief snippet, the multichannel capabilities of Darktrace / EMAIL stand out immediately. Within the Logs interface are "Email" and "Messages" options, with Teams chat messages receiving the same depth of behavioral analysis as email. In examining the flagged Teams chat, the platform's AI narrative identified several channel-specific behavioral indicators. The message was the first in a new chat, contained a SharePoint link hidden behind display text, external parties were present, and the message length was unusual for the Teams medium. The platform adapts its behavioral baselines to channel-specific norms rather than simply applying heuristics.

Self-Learning AI and Detection Philosophy

The fundamental difference between Darktrace / EMAIL and traditional email security lies in detection methodology. Traditional tools train on historical attacks. They learn what a malicious email looks like and attempts to match incoming messages against that pattern. Darktrace takes the opposite approach; learning what normal business behavior looks like and flagging deviations from that baseline.

The platform builds behavioral models for every user and correspondent in the organization. For internal users, this includes:

- Communication patterns
- Typical content type
- Tone and sentiment
- The URLs and file types they normally share

For external correspondence, which includes vendors, partners, and customers, the platform builds equivalent baselines, learning how each third party typically communicates with the organization. Threat intelligence and sandboxing capabilities provide additional context, but they supplement the behavioral core rather than drive detection. The strength of this approach is detecting threats with no prior signatures. Novel social engineering, zero-day phishing, and compromised trusted accounts all produce behavioral deviations the platform can identify without requiring any historical knowledge.

Cross-Platform Intelligence

Email analysis doesn't happen in isolation. As Darktrace / EMAIL shares its AI core with the broader platform, behavioral analysis incorporates signals from network and cloud activity. When the platform scores a link's rarity, for example, it evaluates that URL against all traffic observed across the organization. Similarly, when a user's behavior shifts, the platform can correlate email changes with anomalous network or cloud activity to provide richer context before an alert fires.

Detecting Trusted Sender Compromise

One of the platform's standout capabilities is detection of trusted sender compromise. When a vendor, partner, or other established contact has their account compromised, the resulting phishing emails come from a legitimate sender with *real* communication history. This is exactly the type of threat signature-based tools struggle with because metadata about the sender, the headers, email content, frequency, and other factors, checks out.

Trusted sender compromise poses a significant threat to every organization because it takes advantage of pre-existing relationships and communication channels. Adversaries don't need to reinvent the wheel; they just need to appear to be someone else.

Darktrace / EMAIL detects these scenarios through “out of character” behavioral deviation. The platform flags messages when a trusted sender’s communication deviates from known behavior. Changes in tone, sign-off style, unexpected requests for bank detail charges, and unusual urgency patterns all register as behavioral anomalies against the sender’s established baseline. This lays the groundwork for detecting attacks like supply chain compromise, BEC, and account takeover (ATO).

Detection Architecture: Models and Recipes

Under the hood, the platform’s detection logic is organized into a two-tier architecture accessible through the Detection menu: Models and Recipes.

Recipes are modular detection logic components that draw upon AI models within the platform—reusable building blocks that evaluate specific behavioral or technical conditions. Each Recipe is constructed as a visual logic flow combining AND, OR, and NOT gates across detection objects (see Figure 2).

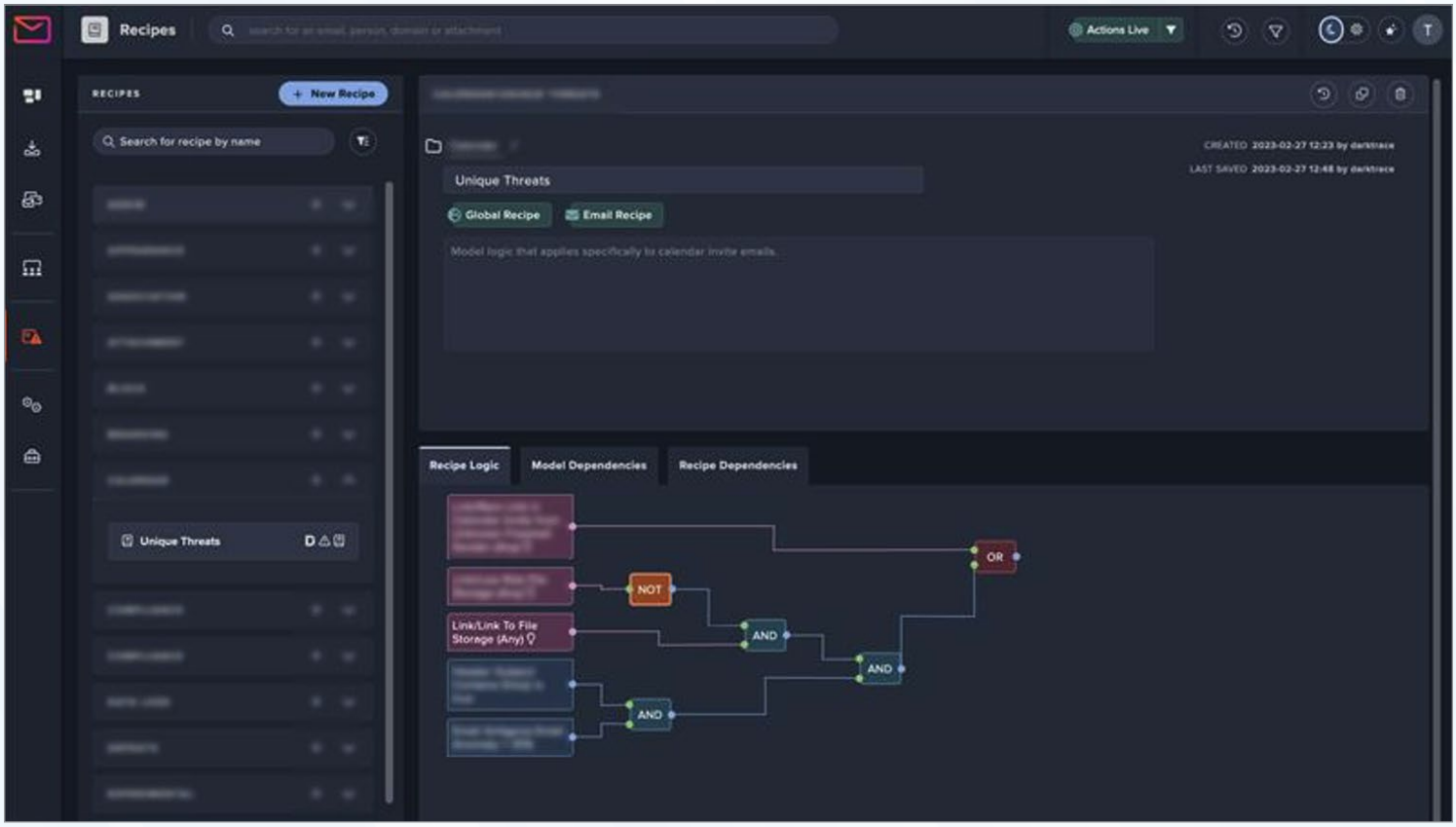

Figure 2. Recipe Construction

Examination of a Calendar Recipe for “Unique Threats” evaluates a combination of sender context, content characteristics, and behavioral thresholds, all orchestrated through a visual logic flow. Recipes are organized into categories including Appearance, Association, Attachment, Block, Branding, Calendar, Compliance, Data Loss, and more, with tabs showing Recipe Logic, Model Dependencies, and Recipe Dependencies that map how each component feeds into the broader detection system.

Models sit above Recipes and define both detection criteria and response actions. Each Model specifies what conditions trigger a detection, drawing on Recipes and direct detection objects, and what the platform does when those conditions are met. Figure 3 illustrates a Model named “Critical Attachment from New Address,” found within the platform’s Attachment category. The Model targets emails from contacts with no or limited history that contain an HTML file with minimal surrounding content.

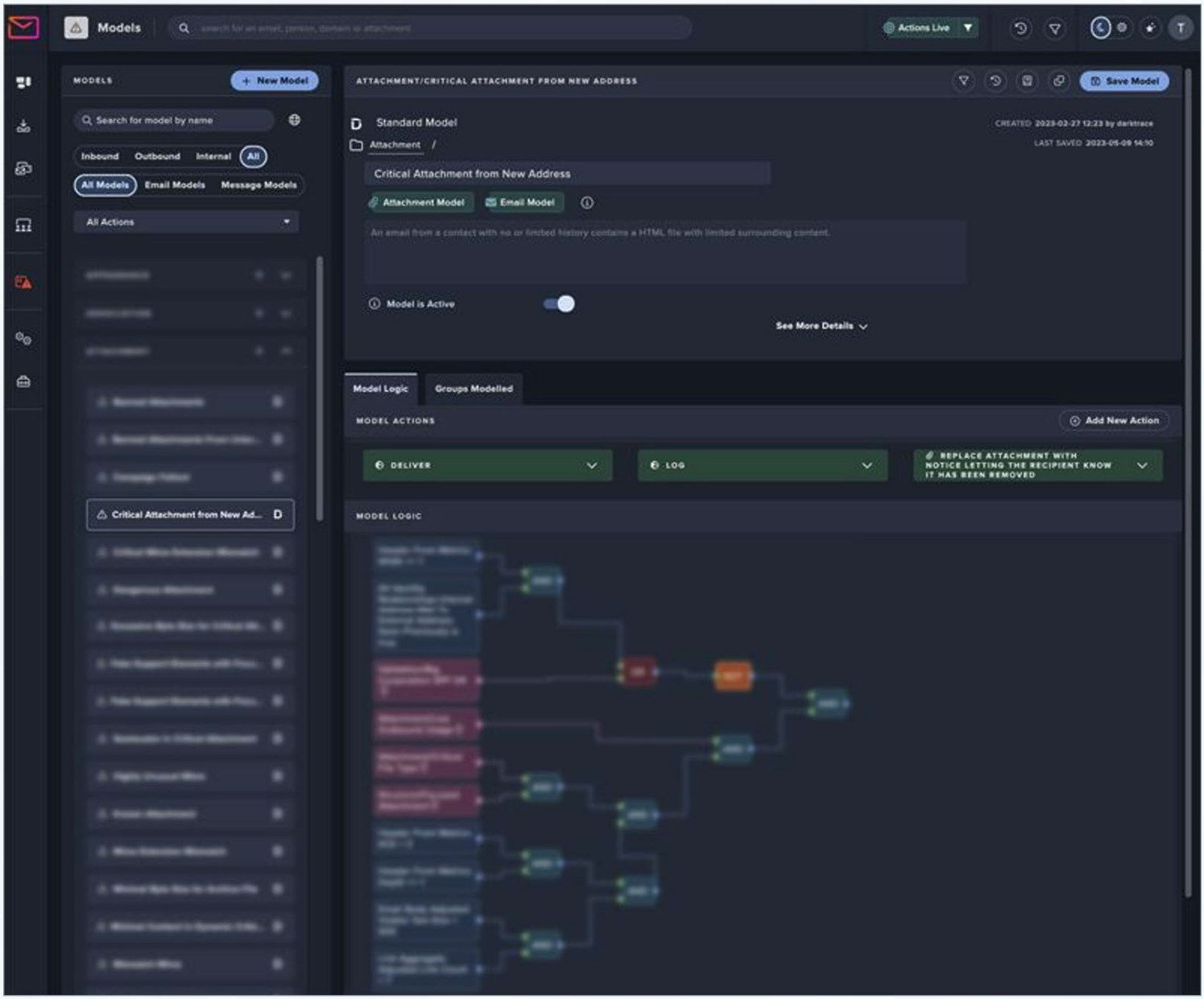

Figure 3. Model Construction

The visual logic flow chains multiple conditions:

- Sender metrics evaluating header width and conversation depth
- The sender is *not* from a known large corporation
- Attachment type classification
- Body characteristics evaluating visible text properties

What makes this transparent is the response configuration alongside the detection logic. For this Model, three actions fire simultaneously:

- DELIVER (the email reaches the recipient)
- LOG (the detection is recorded for analyst review)
- REPLACE ATTACHMENT WITH NOTICE LETTING THE RECIPIENT KNOW IT HAS BEEN REMOVED

The dangerous attachment is stripped, the recipient is informed, and business communication continues with minimal disruption (see Figure 4).

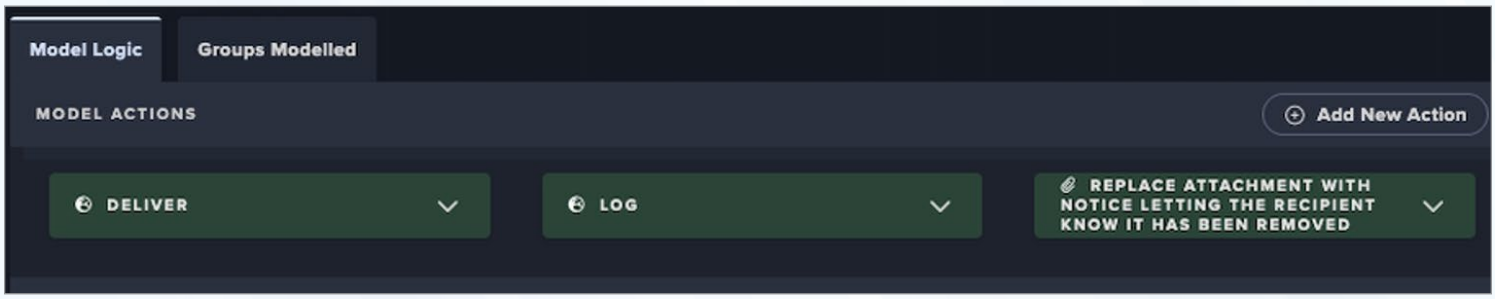

Figure 4. Model Response Actions

Models can be filtered by direction (Inbound, Outbound, Internal) and type (Email Models, Message Models), and the platform provides a “New Model” button for organizations ready to build custom detection logic. Darktrace recommends working collaboratively for significant Model modifications, but the full logic is visible for any detection the platform makes.

Threat Analysis and Investigation Workflow

The Darktrace / EMAIL dashboard serves as the analyst's primary entry point and reflects the platform's "quick in, quick out" operational philosophy (see Figure 5).

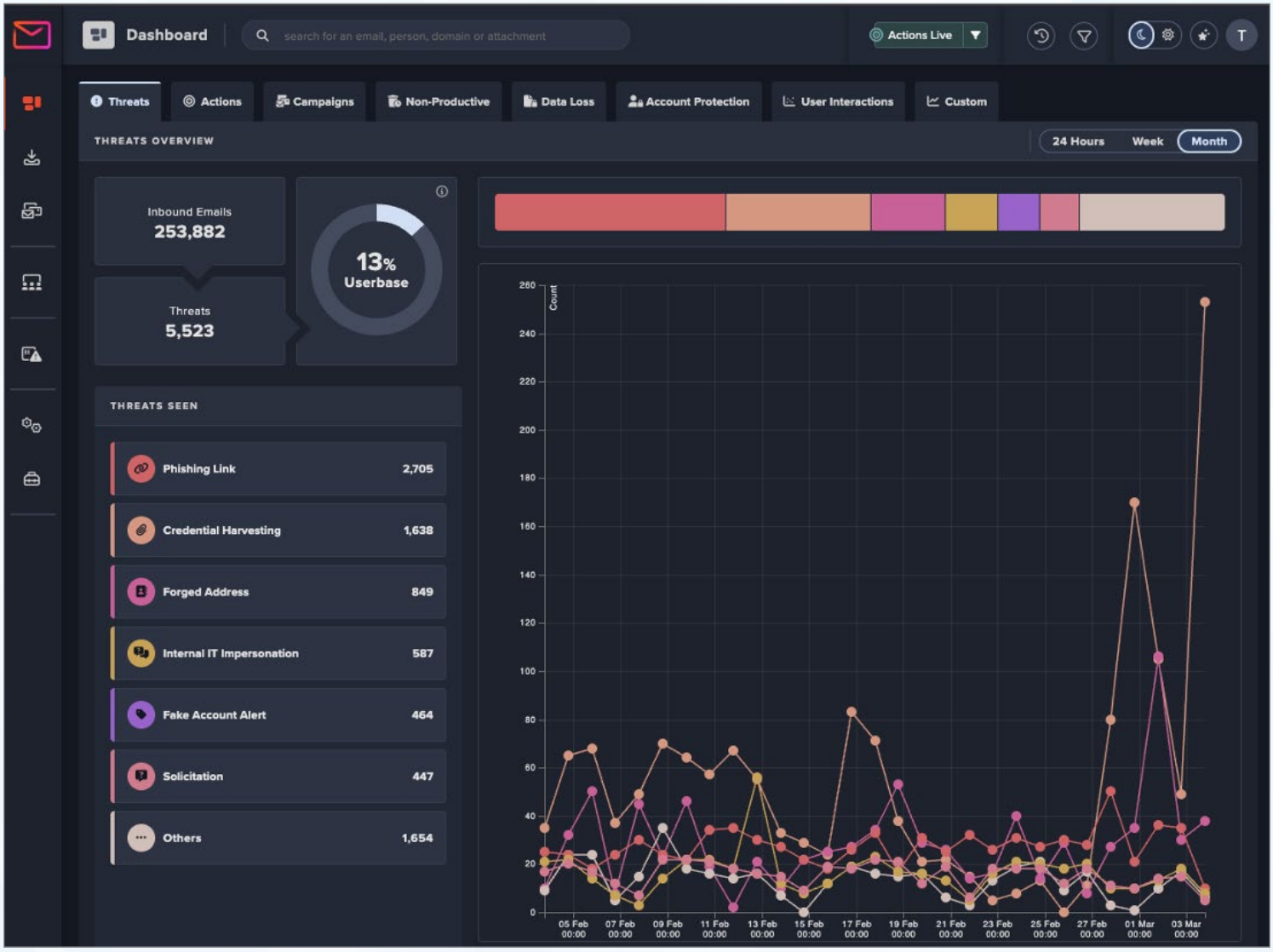

Figure 5. Dashboard upon Initial Login

The Threats tab, one of the eight top-level categories within the main dashboard, opens an overview summarizing email volume, targeted user percentage, and total threat count over a configurable time window (selectable by the user).

The provided stats were plentiful. Darktrace reported 253,882 inbound emails analyzed over the previous month, with 5,523 threats identified, affecting 13% of the monitored userbase. The Threats Seen breakdown immediately categorizes detections by type:

- Phishing Link (2,705)
- Credential Harvesting (1,638)
- Forged Address (849)
- Internal IT Impersonation (587)
- Fake Account Alert (464)
- Solicitation (447)

The operational stats are front and center with little to no fluff. This categorization gives analysts and security stakeholders an instant operational picture, not just how many threats were detected, but what kinds of threats are targeting the organization. The trend graph plots detection volume over time by category, making it easy to spot spikes or sustained campaigns. Below the graph, a Threat Trends panel highlights the most significant changes over the reporting period. During the review, Document Anomalies had increased 80% and Phishing Attachment was up 46%, while Extortion had decreased 51%. These trend indicators surface shifts in the threat landscape without requiring analysts to manually compare time periods. This is exactly the kind of operational intelligence that supports the "quick in, quick out" workflow.

Beyond Inbound: Data Loss and Account Protection

The dashboard's tab structure extends well beyond inbound threat detection. The Data Loss tab monitors outbound email (see Figure 6).

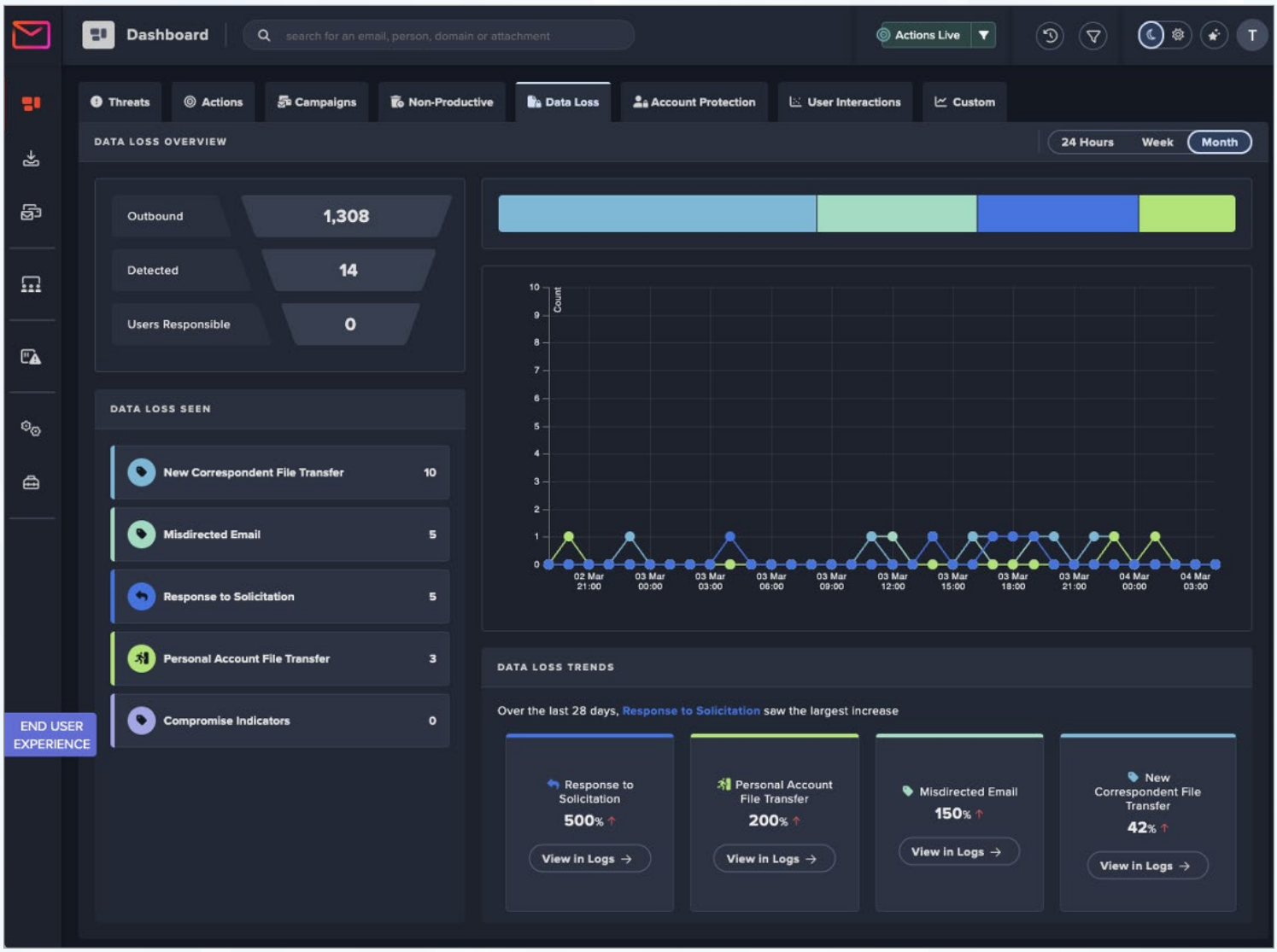

Figure 6. Data Loss Threats

The review period showed 1,308 outbound emails analyzed with 14 detections across categories including New Correspondent File Transfer, Misdirected Email, Response to Solicitation, and Personal Account File Transfer. The accompanying trend panel showed Response to Solicitation had increased 500% over the reporting period, the kind of outbound behavioral shift that could indicate a compromised account or policy violation.

Although this review won't cover every threat tab, the Account Protection tab is worth highlighting. It takes visibility even further, monitoring login behavior and surfacing anomalous account activity (see Figure 7).

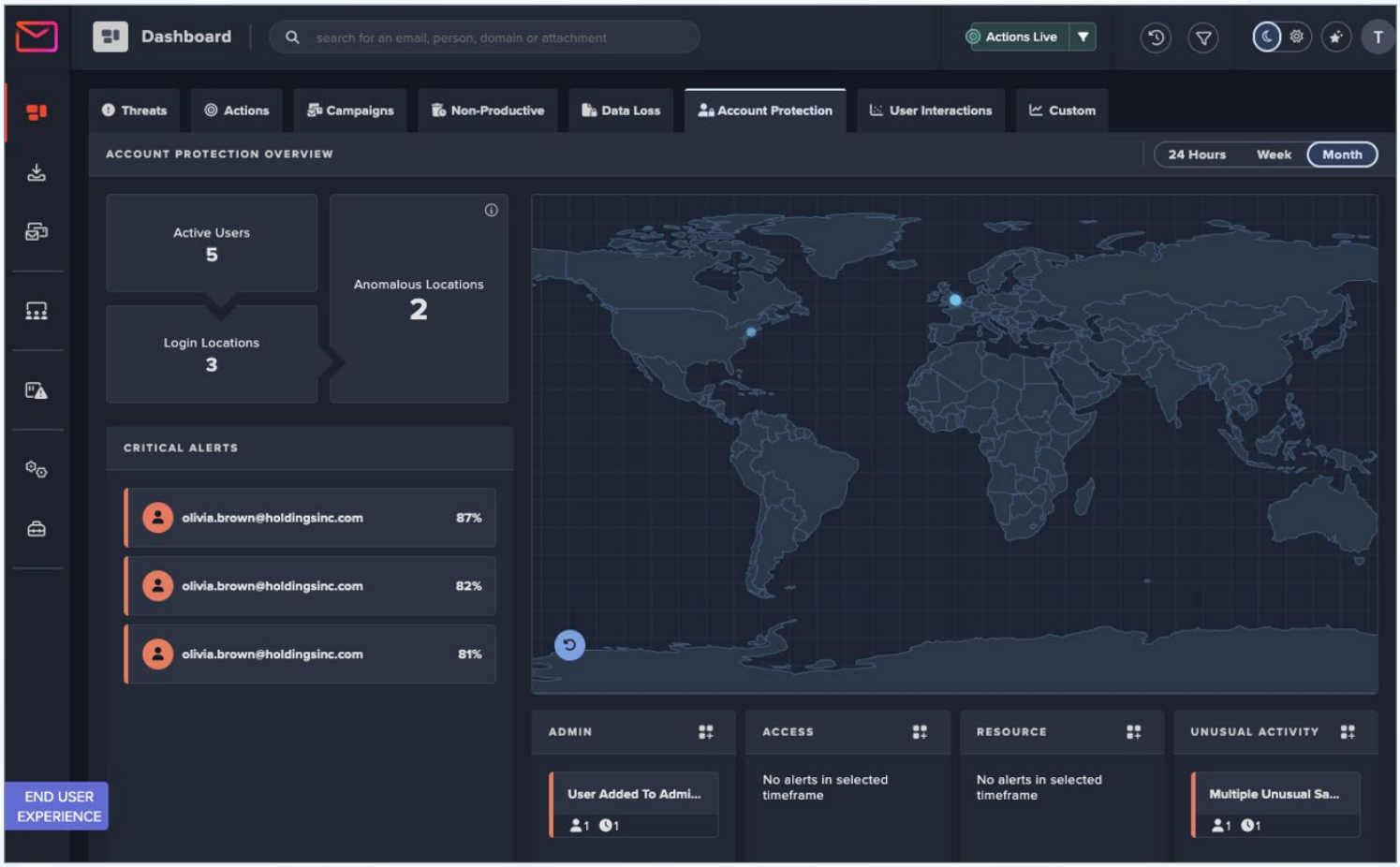

Figure 7. Account Protection Details

During the review, critical alerts identified a specific user scored 87%, 82%, and 81% anomaly scores across multiple events. Events are mapped out geographically where applicable, with key detections such as "User Added to Admin Group" and "Multiple Unusual SaaS Activity" extending well beyond email content to incorporate identity and access patterns.

The same user mentioned in the dashboard above also appeared in at least one email investigation with a held message flagged for compromise indicators. This correlation, which was apparent throughout the review, demonstrates the platform's power to connect email behavioral anomalies with account-level suspicious activity. This not only builds a comprehensive picture of potential compromise, it also instills confidence that analysts are investigating true positive activity, not benign activity.

Investigating a Flagged Email

From the Logs view, analysts work through flagged emails sorted by an anomaly score. The anomaly score, as seen in Figure 8, is a numerical indicator representing increasing deviations from learned baselines.

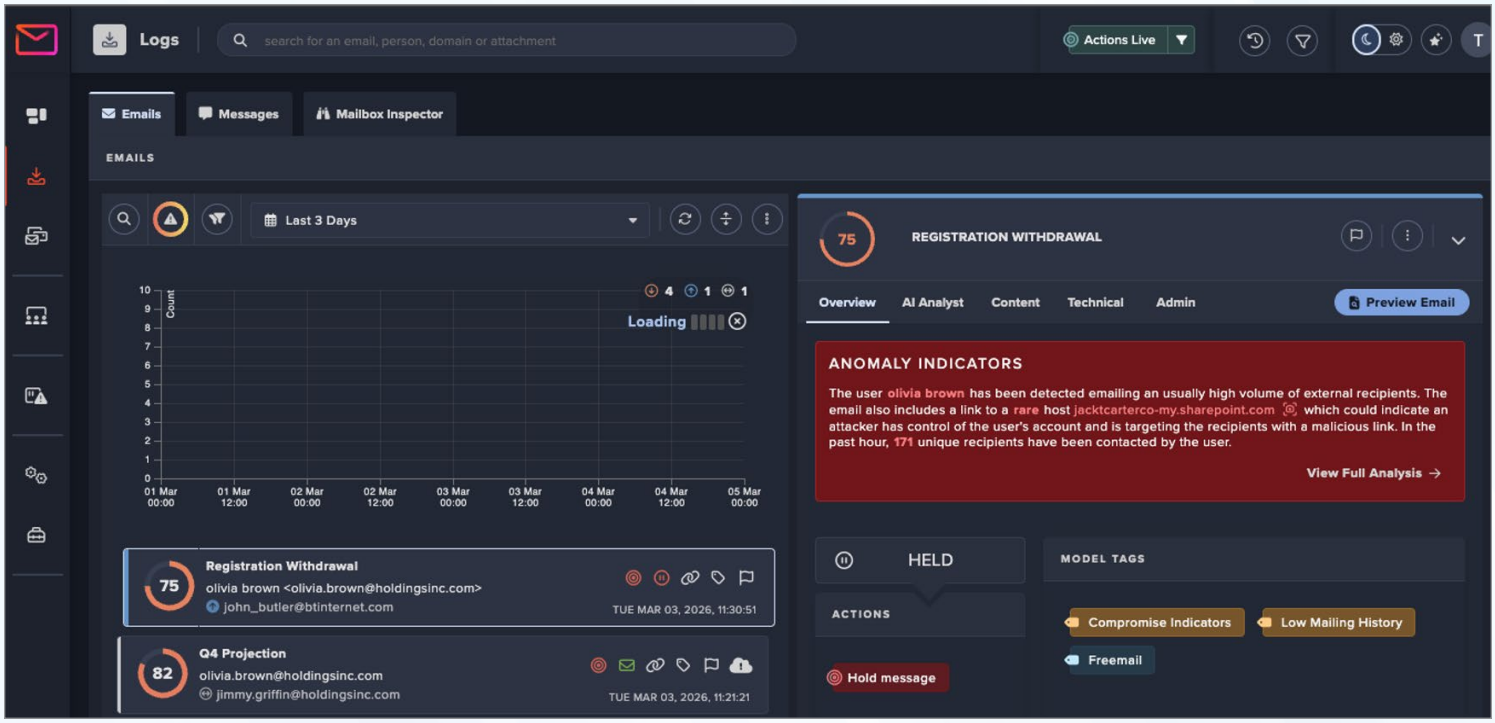

Figure 8. Flagged Email(s)

During the review, the email list displayed scores ranging from 75 to 100, with color-coded severity indicators providing immediate visual triage. This makes alerts clear and crisp, giving analysts an immediate sense of what to look for.

The Overview Tab immediately presents the Anomaly Indicators narrative; a plain-text AI-generated explanation contextualizing the detection against the organization's specific baseline. For this email, the narrative identified that the user emailed an unusually high volume of external recipients, email(s) included a link to a rare SharePoint host (which could indicate adversary control of the user's account), and that 171 recipients had been contacted in the past hour.

Below the narrative, the platform displays the action taken (HELD, in this case), and the Model tags that contributed to the detection:

- Compromise Indicators
- Low Mailing History
- Freemail

Additional tabs for AI Analyst, Content, Technical, and Admin provide progressively deeper investigation capabilities, ranging from AI-assisted analysis through raw email content and header inspection.

Visually, Darktrace / EMAIL serves as a one-stop destination for email analysis (formerly known as “single-pane-of-glass”), providing “Read Email” capabilities in addition to other actions (see Figure 9).

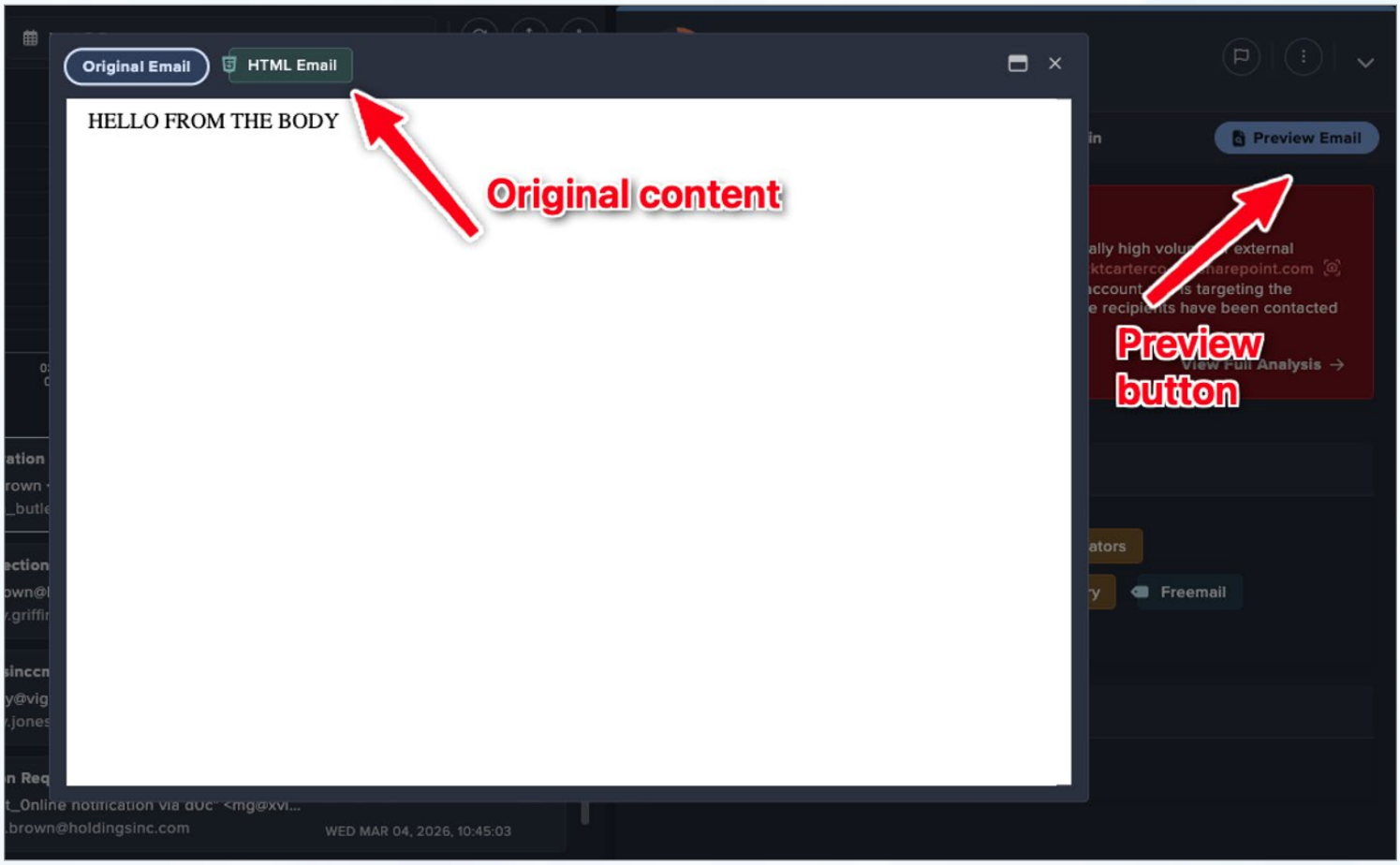

Figure 9. Read Email Capabilities Directly from an Alert

The practical value of this layered interface is speed-to-understanding. The anomaly score provides instant triage priority, the AI narrative explains why the email was flagged in organization-focused terms, the Model tags show which detection layers contributed, and the Action shows what the platform did about it. All this is visible in a single view without navigating away.

Speed-to-understanding is the difference in how quickly an analyst can go from “alert fired” to “I know what’s going on.” The platform surfaces contextual information and key alert details immediately, putting everything needed to understand an alert within easy reach.

Campaigns

When the platform identifies multiple emails as part of coordinated activity, they aggregate into the Campaigns view (see Figure 10).

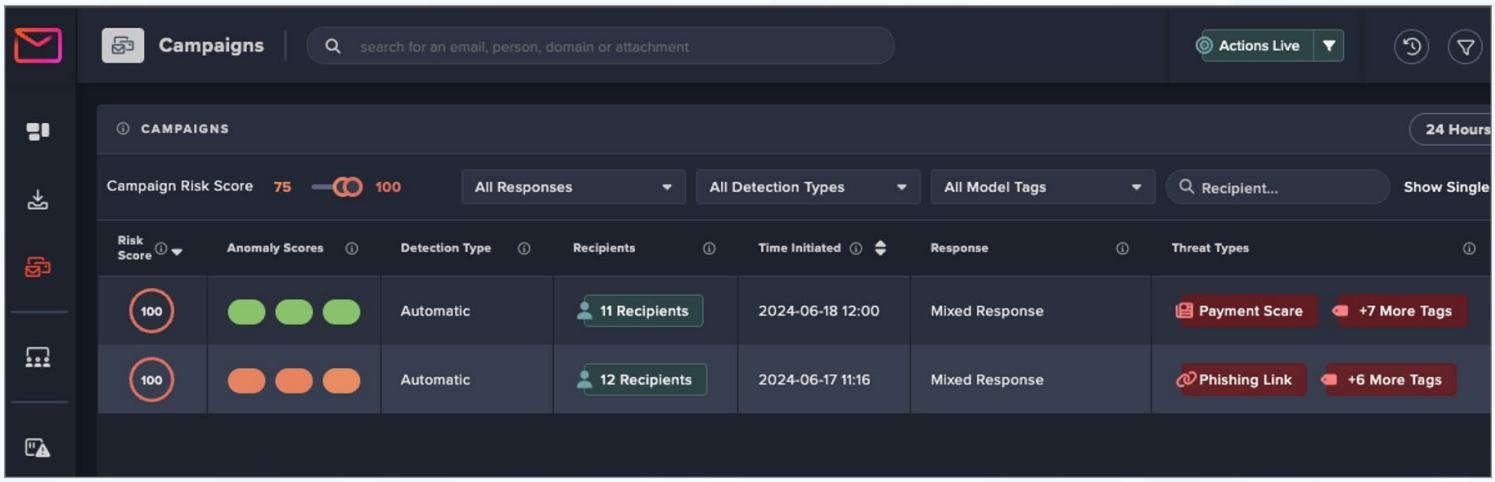

Figure 10. Read Email Capabilities Directly from an Alert

The Campaigns dashboard identified campaigns with risk scores, filterable by a sliding scale (75–100), response type, detection type, and model tags. Two campaigns were visible, both scored at 100, with automatic detection (one tagged Payment Scare with seven additional tags and the other Phishing Link with six additional tags). Each campaign entry showed the number of targeted recipients, response type (Mixed Response), and the time the campaign was initiated.

Darktrace / EMAIL earns high marks for information being easily surfaced, straightforward to parse, and immediately actionable.

Response Actions and Remediation

Darktrace / EMAIL's response philosophy emphasizes reducing risk while maintaining business operations. The platform follows graduated escalation. It starts with the minimum necessary action—like locking a suspicious link within an otherwise-delivered email or converting a potentially dangerous attachment to a safe format—and graduates to holding or quarantining only when the risk warrants it.

This graduated approach was evident throughout the review. The "Registration Withdraw" email (see Figure 11) from a potentially compromised account was held entirely.

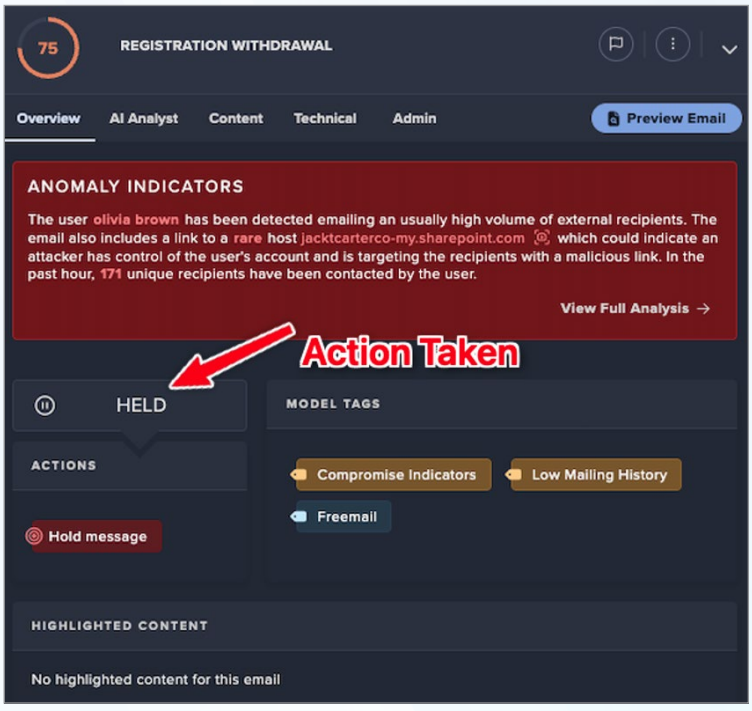

Figure 11. Zoomed in Snippet on a Malicious Email

Holding an email is an aggressive, disruptive action warranted by the combination of compromise indicators and mass external recipient targeting. In contrast, other observed alerts demonstrated a more nuanced response, such as the email being delivered and logged, with the attachment removed and informing the user of these actions. The recipient still receives the communication; however, the malicious component is neutralized.

The platform's response toolkit includes link locking and neutralization, attachment conversion or removal, message hold and quarantine, folder routing for nonproductive email categorization, and junk classification. Analysts can view the specific action taken on any email directly in the threat analysis view, along with the model tags that drove the decision.

When investigating user actions, analysts also can take drastic steps to fix an account, up to and including disabling a user in Office365, blocking an IP, forcing a logout, or (in certain cases) disabling an inbox rule (see Figure 12).

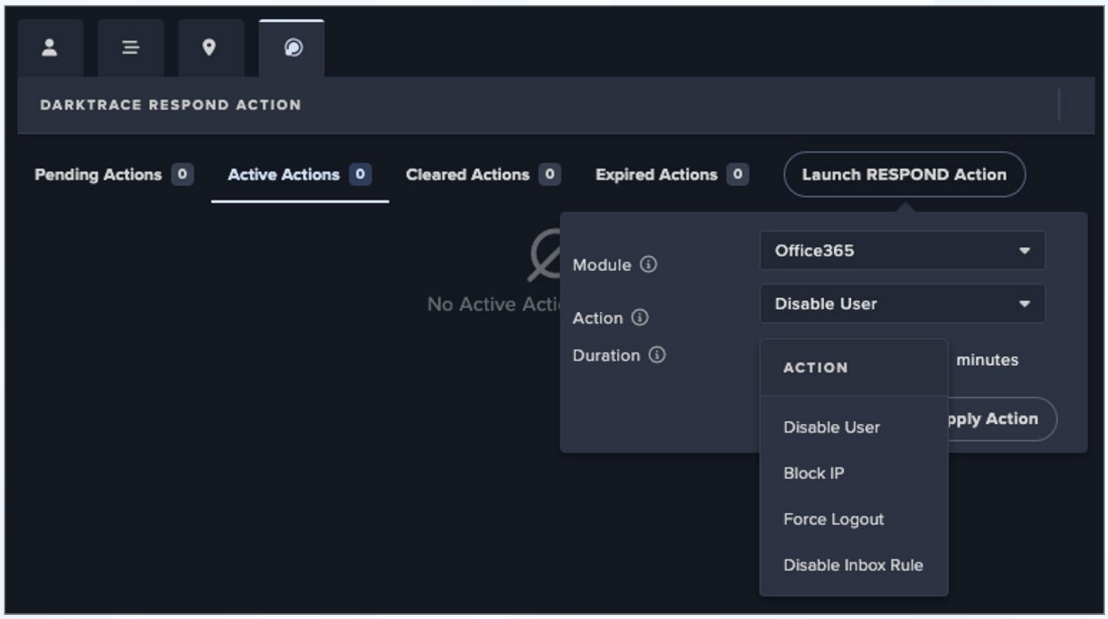

Figure 12. Darktrace Respond Action Menus

Darktrace / EMAIL does an excellent job of splitting the difference between automated actions that can be scripted and based on observed anomalies versus an analyst being able to enact quick, rapid-fire remediations as they investigate enterprise alerts.

Learning Exceptions

When an analyst determines a detection was a false positive, the platform's tuning workflow diverges significantly from traditional email security management. Clicking the "Remediate" button, which is visible throughout the platform, presents an AI-proposed exception tailored to the specific characteristics of that detection, combining the sender identity, the anomaly threshold that triggered, and the specific threat type into a targeted adjustment.

> Unlike static rule libraries that grow unwieldy over time, Darktrace / EMAIL's AI-driven tuning is self-correcting by design—exceptions expire as the platform learns, so the tool adapts to the organization rather than the other way around.

The key mechanism is that these exceptions expire. The platform assumes the false positive occurred because the AI's behavioral baseline had not yet fully learned this communication pattern, not because a permanent gap exists. As the AI continues to learn and incorporates the legitimate pattern into its baseline, the manual exception becomes unnecessary and naturally phases out. Analysts are not building an ever-growing library of static rules that require periodic review and create potential blind spots. The tuning is self-correcting, as it should be. The tool grows with your security team, not the other way around.

Final Thoughts

Darktrace / EMAIL addresses a genuine methodology gap in how organizations approach email security. Most environments already have native email controls and, in many cases, a SEG. But those tools share a common detection philosophy rooted in signatures, reputation, and known-bad indicators. Darktrace / EMAIL operates on a fundamentally different axis, learning what normal looks like for each organization and flagging deviations from that baseline. During the product review, this played out in practice: Trusted sender compromise, novel phishing with no prior signatures, and behavioral anomalies that would pass cleanly through attacker-centric tools were all surfaced and actioned by the platform.

From an analyst's perspective, the investigation workflow stands out, namely the progression from the dashboard's threat categorization and trend analysis through the Logs view, where anomaly scores, AI-generated narratives, and model tags converge in a single panel. This creates a triage experience that prioritizes speed-to-understanding. An analyst can assess a flagged email's risk, understand the reasoning behind the flag, and verify the platform's response in a matter of minutes. The AI-generated narratives don't just describe what was detected, they contextualize it against the organization's specific baseline, referencing sender history, communication patterns, and rarity scores in plain language.

The most impressive features were the detection architecture's transparency and the platform's scope beyond inbound email. The visual Model and Recipe editor exposes the full logic behind every detection, showing exactly which conditions triggered, how logic gates connect them, and what response actions follow. Beyond detection, the Data Loss and Account Protection dashboards extend monitoring into outbound email behavior and identity anomalies, connecting signals across these domains into a unified view.

Darktrace / EMAIL is best positioned for organizations running M365 or Google Workspace that have existing native email security controls in place and want to close the gap against threats those controls weren't designed to catch. These include, in no order, BEC, supply chain compromise, and novel social engineering. For security teams seeking advanced email threat detection that integrates into existing operations without creating new management overhead, look no further.

Sponsor

SANS would like to thank this paper's sponsor:

DARKTRACE