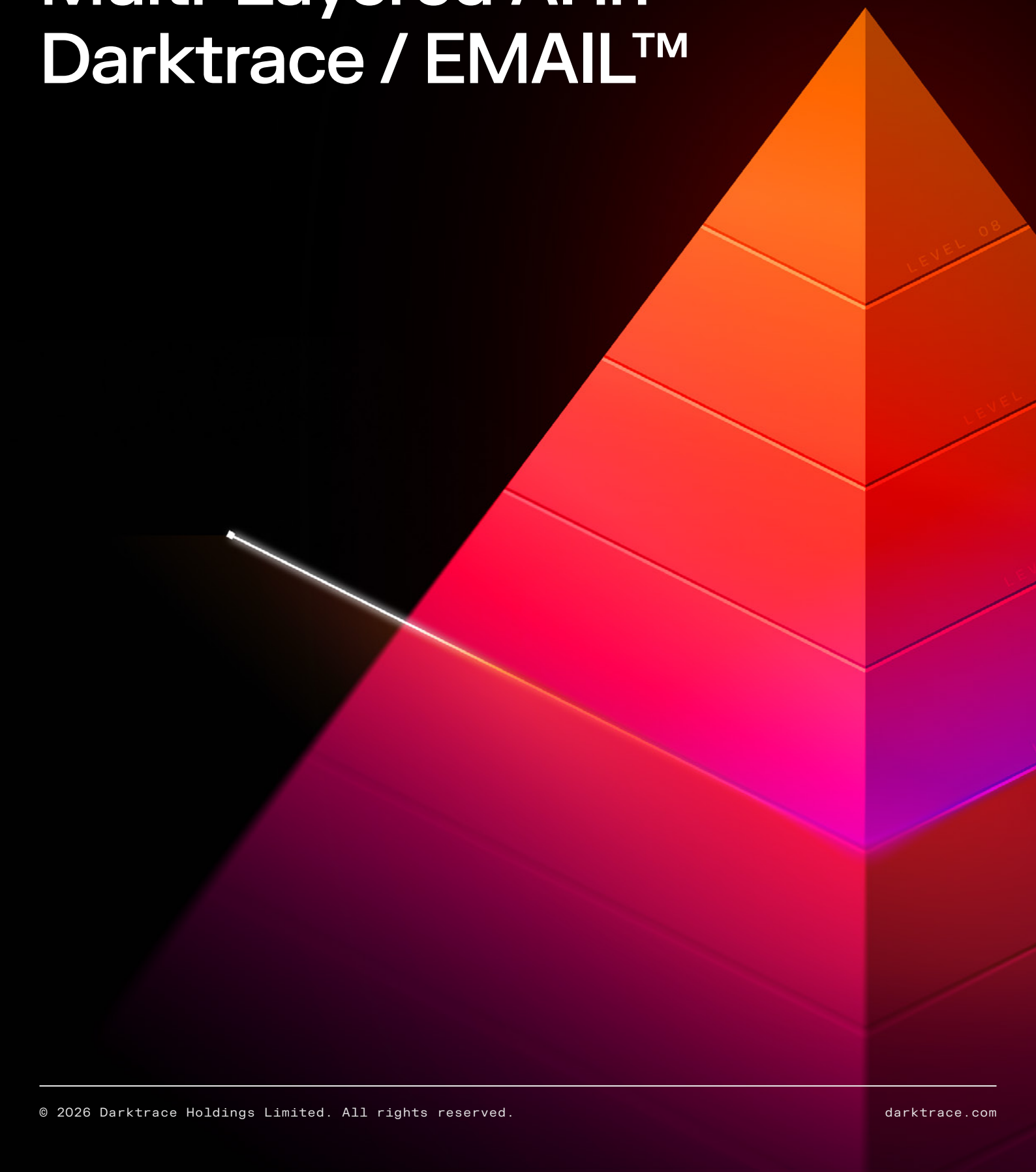
DARKTRACE

A Guide to the Multi-Layered AI in Darktrace / EMAIL™



© 2026 Darktrace Holdings Limited. All rights reserved.

Contents

Abstract

Darktrace / EMAIL is an example of the core Darktrace methodology – a multi-layered AI system built into a single product. Darktrace / EMAIL learns the expected behaviors of an organization without reliance on prior rules, signatures, and historical data.

Automatically building models that are completely unique to each and every customer to identify novel attacks, targeted threats and anomalous activity.

This paper examines the multiple layers which make up Darktrace / EMAIL, and the AI systems involved to provide a comprehensive explanation of how Darktrace learns what an anomalous email looks like.

Darktrace / EMAIL

Multi-Layered AI Defense

Today, the cybersecurity community is experiencing a surge of interest – and uncertainty – around AI. Vendors in the email security space which operate on the principle of identifying and blocking known threats using pre-defined threat intelligence are now bolting pattern-based or siloed AI onto existing products.

This results in a market full of AI claims with little clarity on what these claims actually mean in practice.

Why does it matter?

AI-driven security decisions have real operational consequences: what gets blocked, what's labeled risky, and what needs investigating. A black-box approach doesn't hold up – customers deserve to understand the decisions AI makes, just as security teams deserve autonomy over the tools they rely on.

This paper is designed to deliver that clarity. The goal is simple: to demystify our multi-layered AI, explain why the architecture matters, and show how Darktrace / EMAIL understands the entire organization to deliver contextual detection and real-world operational value.

How is Darktrace's AI different?

Many solutions claim to use AI but often this means learning patterns of known attacks or modelling how attackers behave based on a pre-defined series of rules and signatures. These rules and signatures are not always translatable to novel threats, placing gaps in customer cyber defense strategies.

Darktrace goes beyond this surface level behavioral learning, understanding **the normal behavior of your entire organization from day one:** not just classifying individual behaviors, but also identifying how your people communicate, the relationships they maintain, and the workflows unique to your organization. From that foundation, Darktrace intelligently identifies and analyzes subtle deviations from that "normal" to indicate risk **without relying on preconceived assumptions of what is "good" or "bad".**

The core principle of our approach is understanding what is normal for an environment without any pre-existing baselines. This understanding acts as a foundation for our multiple layers of complex analysis. This approach enables detection that is personal to the environment – where "normal" is discovered, not assumed – and continuously refined as behavior evolves over time. This paper breaks down every stage of that process.

AI built with integrity

One of our core concerns as a business is the responsible development and use of AI. The speed at which AI has advanced over the last five years is tremendous, but in the race to be the first and the best, there is room for integrity to be sacrificed.

At Darktrace, we believe that to realize the huge potential of defensive AI, vendors must build with the right philosophy, and be able to demonstrate to their customers that they have taken "do no harm" approach, focusing on six key pillars : privacy, interpretability, security and robustness, accuracy, and transparency. Darktrace can do just that, putting our development cycle and AI practices to the test by becoming one of the first cybersecurity vendors in the UK to obtain an ISO42001 certification.

Obtaining this standard officially confirms that we uphold a high standard for AI governance, ensuring ethical development (data sourcing) and transparency. All AI systems we develop follow the key principles of our approach towards responsible AI use.

Our AI Architecture

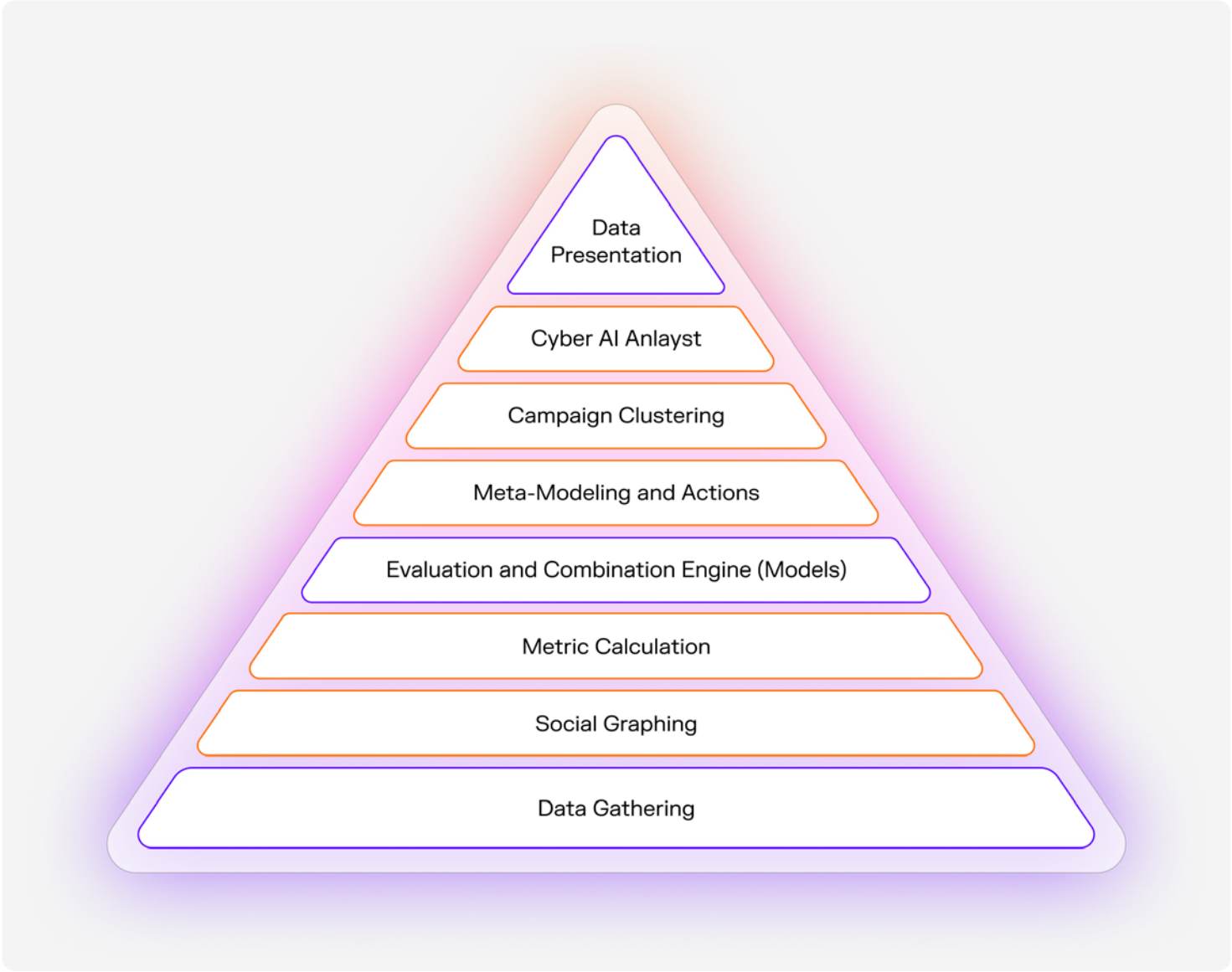
Long before "AI" became a buzzword, Darktrace was building, testing, and operationalizing it in live environments to solve real-world security challenges. We've been at the forefront of cybersecurity innovation since 2013, with AI embedded in our mission from the very beginning.

The pyramid below represents the architecture of Darktrace / EMAIL's multi-layered AI; a structured visualization of how intelligence is built, step by step, from the collection of raw data to the point of actionable insight.

Each layer plays a distinct role and feeds into the next: collecting data, understanding behavior, analyzing intent, making decisions, and presenting clear outcomes.

This is not a question of better or worse infrastructure, but of design philosophy. The layered approach shown here has been built by Darktrace because it enables the most accurate and contextual outcomes, with the software continuously adapting to each organization and delivering results that static or attack-led models cannot replicate.

What follows breaks down each layer of this pyramid and how they work together to deliver these capabilities.



LEVEL 01

Data Gathering

Darktrace will ingest and process data on point-of-transit for inbound, outbound, and lateral (internal) emails, plus data from Teams, Zoom, and Slack – essentially, platforms used for inter-person communication. For instances in API+Journaling Mode [1], this means inbound and lateral emails are processed in parallel with native defenses, enabling faster detection and response to threats. [2]

The first step in processing an email is feature extraction.

Darktrace / EMAIL extracts pertinent and relevant contextual information from the metadata of an email such as the delivery routing, content type, and the results of various authentication checks, as well as HTML and CSS information and content from the headers. [3]

Extracting as much contextual information about the email as possible is a critical foundation for use in the proceeding AI-driven metric calculations.

"Unlike other solutions, Darktrace's AI looks beyond known threat signatures, learning what's normal for our environment and flagging what's not. That was the missing piece – something that could **help us even when everything else failed."**

- CISO
  Global Telecoms Provider

LEVEL 02

Social Graphing

Darktrace / EMAIL utilizes the information extracted at Level 01 to perform social graphing. Social graphing involves the creation of clusters of peer groups and is key to Darktrace / EMAIL's Pattern of Life analysis.

Various unsupervised clustering algorithms are used to identify relationships between individuals within the organization as well as with external parties. Darktrace / EMAIL uses directly extracted information to perform this clustering, relying on historical data to identify behavioural patterns of users within an organization (such as identifying a group of users who frequently communicate with each other).

These clusters are used to algorithmically identify peer-groups.

Peer groups are hierarchical in nature, meaning that users can be clustered into high-level groups, such as working from the same office or who have similar sending schedules, or low-level groups, such as users who frequently contact the same related addresses or external domains.

Clustering of peer groups ensures a better representation of the organization and allows Darktrace to understand the responsibilities of individual users – for example, one user may perform actions or send emails which are normal for them but may look suspicious if performed by another user. This understanding is developed without any preconceived notion of what "normal" or "bad" looks like, and is not based on comparisons with other organizations.

Instead, it is learned entirely from within the organization itself, meaning the individual differences which are key to Darktrace's precise analysis are accurately captured.

1 https://www.darktrace.com/blog/flexible-deployment

2 https://www.darktrace.com/blog/why-api-journaling-delivers-faster-sla-backed-email-security-for-microsoft-365

3 This data is kept only on the clients' own, isolated Darktrace instance and is used for analysis and training of self-learning models. Information about what data is extracted, who can see it, and where it is used can be found in the Email Data Schedule in the Darktrace Trust Centre.

LEVEL 03

Metric Calculation

After information has been directly extracted from an email, Darktrace / EMAIL calculates a series of more complex metrics such as inducement classification, email topic classification, and counts of sensitive data appearing in an email message. These metrics represent more complex characteristics of an email and are required for Darktrace's Pattern of Life analysis.

Algorithmically Calculated (non-AI) Metrics

Certain metrics in Darktrace / EMAIL are calculated algorithmically, such as those which use historical data to indicate the volume and frequency of mails from a particular domain, and a phantom text score indicating the appearance of hidden, font size 0 text. These metrics contain important information by themselves and are used directly in the later decision algorithms, but they also act as a pre-cursor to Darktrace's AI-driven metrics which are required for further analysis.

AI-Driven Metrics

Darktrace / EMAIL features over 1000 metrics (both direct features and calculated values) which each have their own function, but the key metrics which make Darktrace stand out from the competition are the AI-driven metrics which make up our core system. The following section includes some examples of these AI-driven metrics, but this list is not exhaustive.

1000+ metrics

- **calculated by Darktrace**
  to determine an email's status

Email Likelihood

Each inbox within an environment will receive many emails from many different users, both internal and external. One key metric Darktrace calculates is a "mail expectancy score" which indicates a probability that a given mailbox would receive unsolicited external emails.

By analyzing a selection of metrics including the spread and variation in anomaly scores of received emails by each mailbox, Darktrace can determine whether an inbox is public facing or internal. Based on this analysis, the classifier estimates a likelihood that they will receive unsolicited emails. This metric feeds into other calculations, such as calculation of other metrics or for determining lists of VIP users.

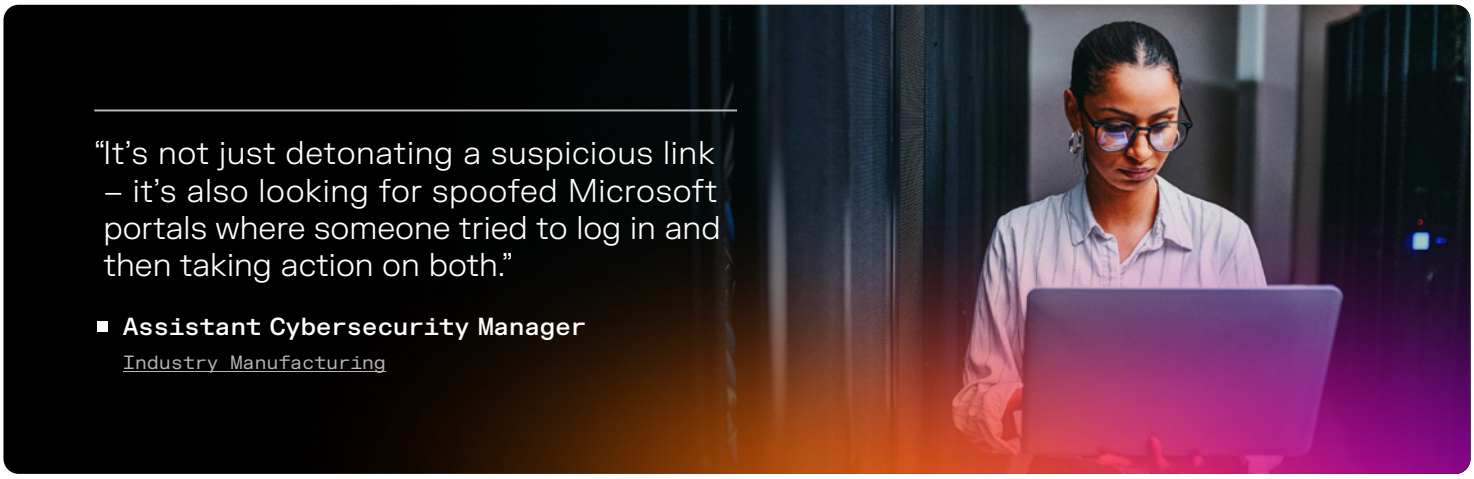
On deployment, this classifier performs a "look back" on several weeks' worth of historical email data from the organization to perform initial training and pattern of life identification, and will continually update itself as more emails are received by each mailbox in real time. This look back is performed for all pattern of life metrics to ensure they are as precisely as possible from day one.

Spoofing Detection

Some metrics, such as those used for identifying spoofing [4], use Bayesian probability analysis. When checking for spoofing, Bayesian Analysis is used to assess the likelihood that a user is impersonating an employee, versus the probability that an email comes from an external mailbox under the same name.

The spoofing detection algorithm is trained uniquely on each deployment. Utilizing an unsupervised learning approach on real-time live data allows the detector to adapt to the organization, as well as account for cultural differences.

As with all other metrics which are extracted, this score alone is not sufficient to confidently determine if an email is coming from a spoofed address. This score will be analyzed along with various other metrics to determine if the email is truly a malicious impersonation attempt.



4 https://www.darktrace.com/cyber-ai-glossary/spoofing

Shift Metrics

A key group of metrics essential for Darktrace's Pattern of Life analysis are shift metrics. Many of our metrics are based on calculations which are localized to the email being processed, using a pre-trained classifier. Darktrace / EMAIL uses an additional dynamic classifier to compare certain metrics to their occurrence in previous emails from the same user. The value produced by this dynamic classifier is known as a shift metric which is used to identify the variance of a metric for a given user.

Shift metrics are calculated for both internal and external parties, allowing Darktrace to provide a complete view of inbound, outbound, and lateral sending behaviours.

If an email has a particularly high shift metric, this suggests a sudden change in behaviour. This could be an indication of a larger issue such as Business Email Compromise or a Supply Chain Attack.

PRACTICAL USE CASE

Detecting Business Email Compromise (BEC) and Supply Chain Attacks

A trusted contact suddenly changes how they communicate. Their tone, timing, or request type is off. Shift metrics highlight this deviation instantly, even when the sender, domain, and authentication appear legitimate. This enables Darktrace to detect BEC and Supply Chain attacks without relying on known threat indicators.

Natural Language Processing

Darktrace / EMAIL features multiple systems which perform Natural Language Processing (NLP). The key NLP systems utilized by Darktrace / EMAIL are an Inducement Classifier (for analyzing language and structure to indicate attacks such as phishing and solicitation that use social engineering and pretexting) and Named Entity Recognition which is used for PII detection and Data Loss Prevention.

Inducement Classification

One important set of metrics which are calculated for every email are the Inducement Classification scores. These scores are used to determine if an email has been sent with a malicious intent. There are 4 different critical inducement categories which Darktrace will check for and present to users – extortion, phishing, solicitation, and spam.

Inducement classification is performed by passing a selection of email metrics through an AI-powered classification system, trained through a supervised learning approach. The classifier considers a selection of natural language metrics, as well as a collection of structural details, such as formatting of addressing fields, the location of any links in an email, and various other aspects of text construction.

By giving a prominent role to structural metrics, the system can be more language agnostic, increasing accuracy when performing inducement classification of non-English emails. By not only relying on regular expressions like many email tools, we aren't limited by the languages stored in our database. For a deeper exploration, see how Darktrace analyzes email structure to determine malicious intent.

This classifier returns a confidence score for each intent classification. The confidence scores are utilized by later decision algorithms to determine if an email is malicious.

Topic Classification

As well as inducement classification, NLP is used for Topic Classification. By classifying the topic of a given email, Darktrace can be more confident when reporting emails as phishing or solicitation (due to common correlations between topic and inducement) and can identify shifts in topic discussed by users.

Topic classification is performed using a probabilistic model, trained using an unsupervised learning approach, which can cluster messages into topic groups. The classifier produces a confidence value per-category, rather than just returning a single category, meaning that one email could be identified as featuring multiple topics.

The training happens live on deployment, using real emails from the organization. Training on deployment using real business data allows the system to adapt to and support the language(s) of the deployment.

Topic Shift

As well as being useful for boosting the accuracy of other calculations (such as correlations identified between topic and inducement classification), Topic Classification can be used to calculate another Shift Metric – namely Topic Shift.

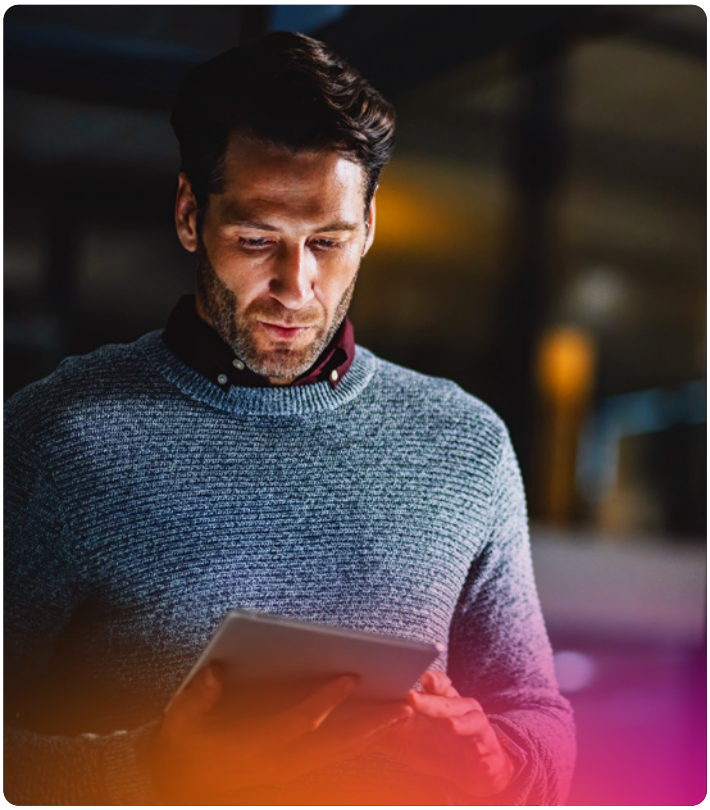
If a user who typically only sends emails containing a few topics suddenly starts sending many messages with different topics, they'll be identified as having a high topic shift, boosting the anomaly score of their emails and potentially flagging them for further investigation. This allows Darktrace to potentially identify Insider Threat or Compromised Mailboxes where a traditional email management system would not.

PRACTICAL USE CASE

Insider Threat & Compromised Mailboxes

By learning the typical topics a user discusses – and who they normally communicate with – Darktrace builds a baseline of expected behavior.

When a user suddenly changes both *who* they are emailing and, especially alongside potentially malicious content, Topic Shift highlights this combined deviation. This layered change can indicate a compromised mailbox or insider threat that would be difficult to detect using rule-based systems.



"Darktrace has stopped attempted data transfers immediately, and those events helped us **refine our internal security policies."**

- **Head of Systems**
  Manufacturing

Data Loss Prevention

Data Loss Prevention (DLP) is a task many email management systems struggle with. Traditional email filtering systems must rely on pre-labelling of data to determine when sensitive data is being exfiltrated. Darktrace / EMAIL features a stack of multiple models and metrics for detecting potential data loss, including behavioural and topic metrics, identification of personal freemail addresses, and PII identification.

Named Entity Recognition

One of the key models Darktrace / EMAIL uses for DLP is a Named Entity Recognition (NER) model capable of detecting PII in an email. This model is a multilingual large language model (LLM) which has been fine-tuned for the task of Named Entity Recognition. The model uses an encoder-only architecture meaning it can't produce generative text output or be used as a chat bot, and can only be used for the creation of text embeddings. These embeddings are all classified and tagged with an entity type where applicable.

The model can identify a wide range of classifications including names, meeting details, financial information, phrases which indicate urgency, and many more.

By identifying all named entities and potentially sensitive information within an email, and tagging them with the specific entity type, additional metrics can be produced to indicate the amount of sensitive data included. These metrics indicate the volume, density, and ratio of each different named entity type. This information can then be used by later decision algorithms to identify potential Data Loss incidents, or to help with the identification of account compromise and insider threat.

PRACTICAL USE CASE

Detecting Sensitive Data Exfiltration

Let's imagine an employee shares an email containing sensitive information – such as a spreadsheet of financial details – with an external recipient. Named Entity Recognition identifies and classifies this data in real time, while behavioral context highlights whether this type of sharing is expected. This combination allows Darktrace to flag potential data loss incidents even when no predefined labels or rules are in place.

LEVEL 04

Evaluation and Combination Engine (Models)

Darktrace / EMAIL performs threat classification and contextualization, triggering actions through an evaluation and combination engine. This engine is built of a series of decision algorithms called "models". Models are blocks of complex logic which build upon the statistics and metrics calculated previously, including AI model outputs. Models evaluate these metrics to determine the intent of an email, whether it contains any potential malicious content, and decide any actions to take.

Clients are also able to create their own workflows in the evaluation and combination engine using any of the metrics which have been calculated, and can assign autonomous actions to these. This includes AI outputs and Pattern of Life data. Clients will have access to a vast range of metrics produced by Darktrace's sophisticated AI models.

Furthermore, each of the models in this engine can be individually enabled or disabled, allowing fine control over what does or doesn't cause an alert. This can allow end-users to reduce false positives, by disabling models which alert erroneously for their environment.

Based on the outcome of these models, Darktrace will perform actions to remediate or mitigate any risk imposed by an email. This includes holding emails, disarming attachments (e.g. converting files to pdfs) or removing them entirely, locking and re-writing links, adding banners to an email, or appending labels to the subject to indicate a potential risk.

Having all models, actions, and AI metric outputs visible and accessible, and the ability to create custom workflows using the provided metrics, allows for complete transparency and customizability over the processing done by the / EMAIL system. **Transparency and interpretability are principles in our responsible AI approach.**

Models are composed of metrics (pink and blue boxes) specifying certain criteria or thresholds, joined together with logic gates created to determine whether or not an email contains a threat.

Patterns of Life

Models bring all the metrics of an email together. Models utilize key metrics, such as shift and anomaly metrics, to identify a Pattern of Life for the organization as well as for each mailbox. The pattern of life itself isn't a separate, manipulatable statistic, but rather a combination of metrics which rely on historical data.

By identifying abnormalities from a user's pattern of life, Darktrace / EMAIL can identify Business Email Compromise (BEC) or Vendor Compromise (VC), as well as Insider Threats.

Vendor and Business Email Compromise

To detect suspicious or malicious emails, traditional email security systems typically rely on SPF, DKIM, and DMARC failures, manually updated block-lists or static signatures to recognize known-bad indicators.

These methods are often incapable of identifying Business Email Compromise (where an email account of a trusted domain has been compromised) as the typical indicators would all return cleanly.

In the case of BEC or Vendor Compromise, Darktrace / EMAIL will detect abnormalities in a sender's tone and behaviors, using these indicators to identify and block attacks.

Darktrace / EMAIL isn't limited to just identifying BEC and VC on inbound emails, but is also capable of scanning outgoing and lateral emails for a potential insider threat. A sudden shift in the tone or technical content of an outbound email could indicate potential Data Loss (either through a malicious insider, or an accident from over-reliance on auto-filled addresses, for example).

LEVEL 05

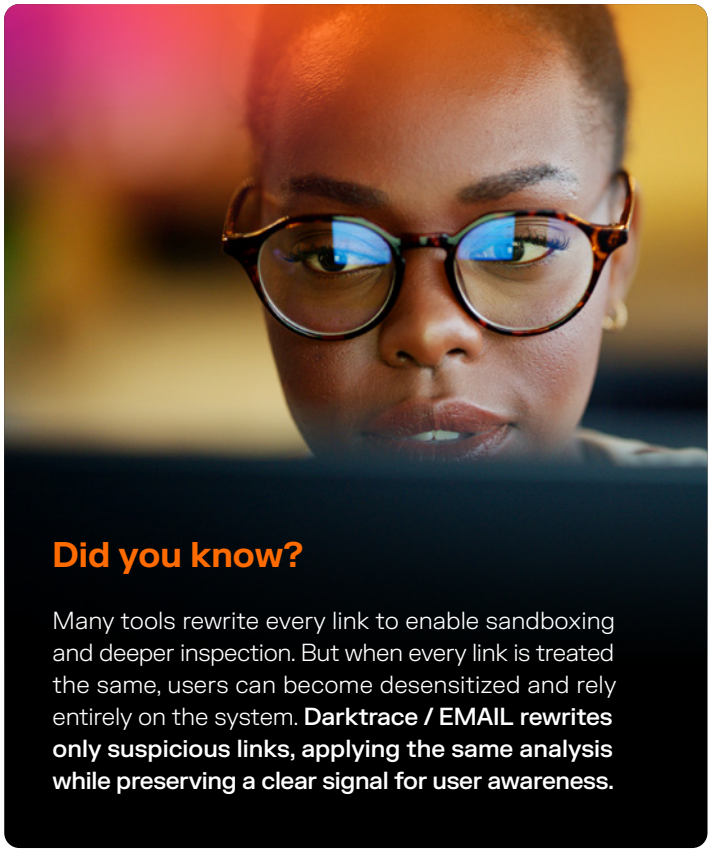
Meta-Modeling and Actions

Meta-modeling

The next step of Darktrace / EMAIL's analysis, after calculating metrics and applying models, is a meta-classifier which is used to calculate an overall email anomaly score. The score summarizes how abnormal a given email is, relative to the other emails received by a user or the organization.

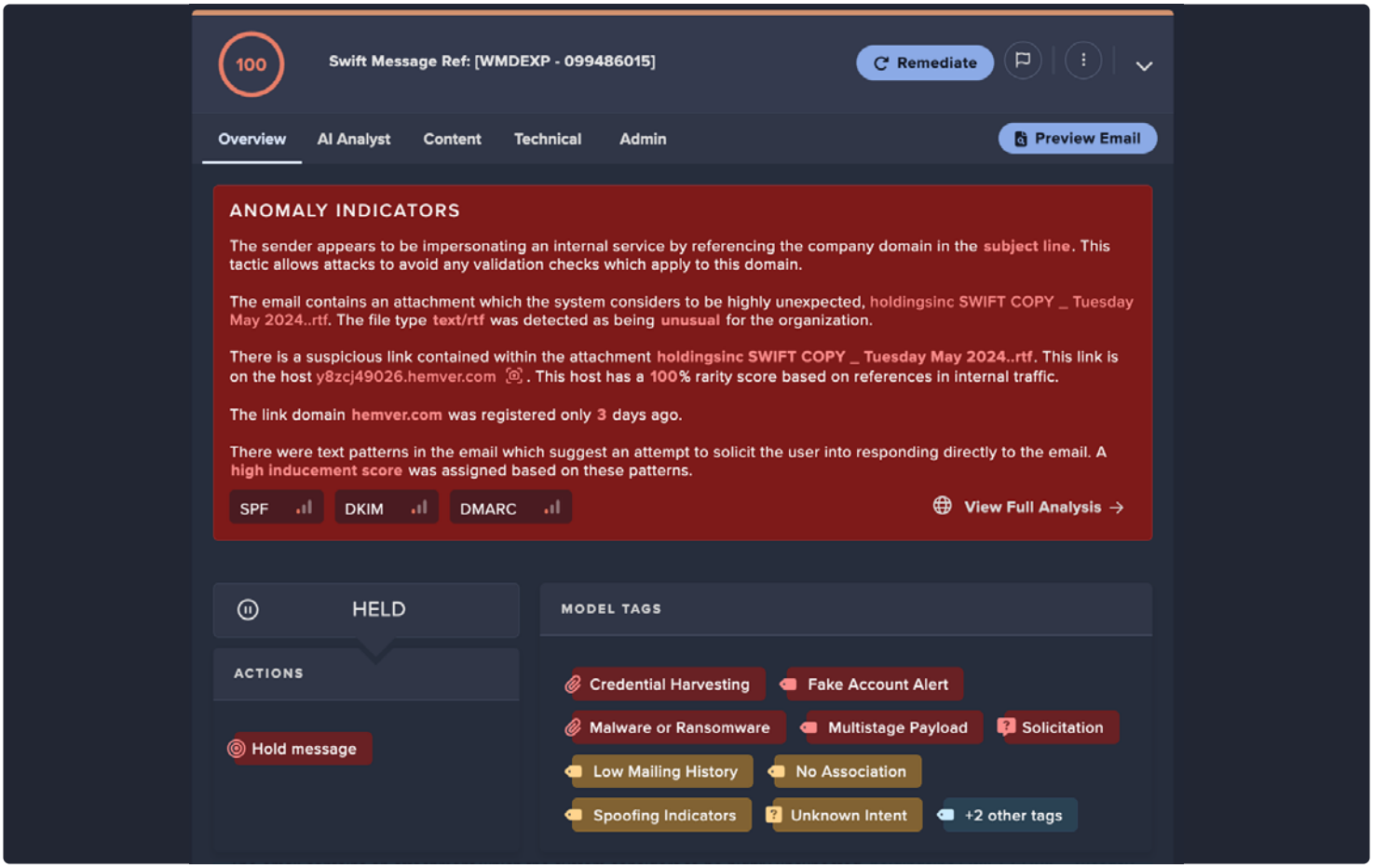
An email with a high anomaly score means that the content of the email differs from what is expected of the organization or user's pattern of life. It's important to note that the anomaly score does not always directly correlate with how "dangerous" an email is - a highly anomalous email isn't always a dangerous email.

This means that autonomous actions cannot be performed against emails based purely on the anomaly score they receive. Instead, autonomous actions are determined by the evaluation and combination engine described previously.



Did you know?

Many tools rewrite every link to enable sandboxing and deeper inspection. But when every link is treated the same, users can become desensitized and rely entirely on the system. **Darktrace / EMAIL rewrites only suspicious links, applying the same analysis while preserving a clear signal for user awareness.**



The overall anomaly score for each email is based on how much the email differs from the pattern of life for the user, boosted by any actions that have taken place.

Actions

Darktrace / EMAIL performs automated actions against emails as required. These actions could be to add banners or warning tags to the headers of an email, or in the case of more malicious emails could block an email outright. Integrations with other Darktrace products may include other actions, such as Just-In-Time training provided by Darktrace / Adaptive Human Defense.

Link Locking and Rewriting

Hyperlinks are a very common and essential component of email communication. Threat actors often use links in emails to instigate phishing attacks. Links which are placed in suspicious or enticing locations; links which have been masked, hidden, or disguised; or links which point to domains which are rare or unusual will be locked by / EMAIL. When a link is locked, it gets re-written to a Darktrace-specific URL.

Clicking on this re-written hyperlink will trigger a second stage triage, sending the user to the Darktrace page where a second, more in-depth analysis and behavioural sandboxing takes place. If the site is deemed safe, the user will be re-directed to the address, otherwise the link will be double-locked, completely denying the user access to the malicious site.

Webpage Analysis

When a URL gets sent for processing, the site is opened in a sandbox environment. Darktrace analyzes the webpage, identifying a series of metrics for processing. Content is taken from the SSL certificate (or lack thereof), and the HTML, CSS, and JavaScript content. Links to other pages, drive-by downloads, and masked text will also be noted in webpage metrics.

Computer Vision

As well as pulling metrics directly from the page content, Darktrace will take a screenshot of the webpage which will be analyzed using Computer Vision. The primary function of the screenshot analysis is to detect fake login pages which are used for credential harvesting campaigns.

Darktrace uses a classifier which is pre-trained using a supervised learning approach. The training data includes a selection of screenshots of login pages from various companies, taken from multiple different browsers and devices (mobile and desktop). The variety of training data helps the classifier when it comes to generalizing to unseen sites.

It also uses Optical Character Recognition (OCR) to identify phrases associated with login pages ("sign in", "remember me", etc.). The webpage classifier will analyze the extracted information and metrics and returns a confidence value indicating the likelihood that the site is a login page.

This can then be analyzed alongside other metrics to determine if the login page is real or an impersonation.

LEVEL 06

Campaign Clustering

Email-driven attacks are rarely isolated incidents. Threat actors will instead send a series of related emails (or emails with the same intent) to an individual or group of individuals within an organization.

Other vendors rely on previous attack data to identify campaigns. When a new campaign gets reported by end users, the vendor must reactively update their rules, retrain their AI models, and manually adjust their lists of abusive addresses to detect new campaigns.

To identify campaigns more effectively and to proactively against email campaigns, Darktrace / EMAIL uses a multi-layered model which creates clusters based on related features within emails, and identifies campaigns based on these clusters. By utilizing the huge number of direct and calculated metrics available, the Campaign Identification System can detect campaigns without having to be trained on pre-labeled data, boosting its ability to detect novel campaigns.

Email Campaign Identification

The Email Campaign Classifier first extracts a selection of information from an email, such as sender meta-data, subject line, URL content, and attachments. The campaign identifier also considers a wide range of metrics previously calculated, as well as email anomaly scores.

These metrics and features are passed into a pattern recognition engine. The engine performs correlational analysis to find clusters of related emails, returning either an existing cluster which the email fits into, or a new cluster with other emails which haven't yet been put into a campaign grouping.

Not all these clusters will get converted into a campaign. Additional analysis is performed on each of these clusters to determine if the cluster should be a campaign, or if they are just generally similar messages.

Each campaign has a certain Time-to-Live (TTL), and once this duration expires, the campaign closes. When a campaign is closed, retrospective actions are applied to all emails in the campaign. The chosen action will be the same as the one applied to the email in the cluster with the highest impact (i.e. blocking an email would be higher impact than re-writing a link).

Self-Correction

It's expected that each email within a campaign will be assigned a similar anomaly score (though slight differences are expected based on an individual's Pattern of Life). An additional check is performed to see how much the risk score fluctuates between the different emails within a campaign. [5]

The variance of anomaly scores is allowed within a certain range, but if this exceeds an allowed threshold, then this suggests the emails have been clustered based on a coincidental similarity and are not actually part of a campaign. This additional check is used to self-correct campaign groupings. If self-correction happens, the campaign will be moved into a "breakout" status where they will be hidden from the campaigns dashboard.

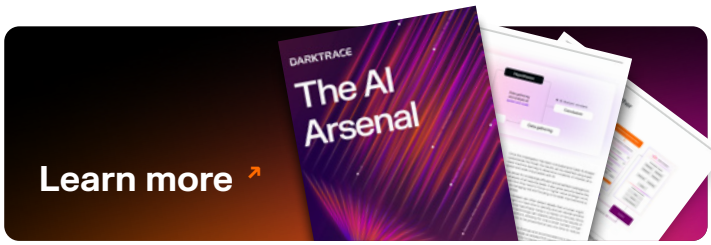
The campaign cluster will still exist, and emails may get linked to this campaign despite it being in the breakout status. If this happens, the self-correction algorithm will be re-run, and the campaign may be re-opened if the anomaly score fluctuation returns to a reasonable level.

LEVEL 07

Cyber AI Analyst

Darktrace's Cyber AI Analyst can utilize data from the / EMAIL application to enhance its investigation of incidents. When a particular domain is referenced in a hypothesis from the Cyber AI Analyst, / EMAIL provides insights into the email behaviour of the domain and the patterns of sending between mailboxes.

The Cyber AI Analyst itself is a separate multi-layered AI system which is a key component of the Darktrace Active AI Security Platform. More information about the Cyber AI Analyst is available in the Darktrace AI Arsenal. [6]



5 https://www.darktrace.com/research/a-real-time-self-correcting-similarity-classifier-for-emails
6 https://www.darktrace.com/resources/the-ai-arsenal

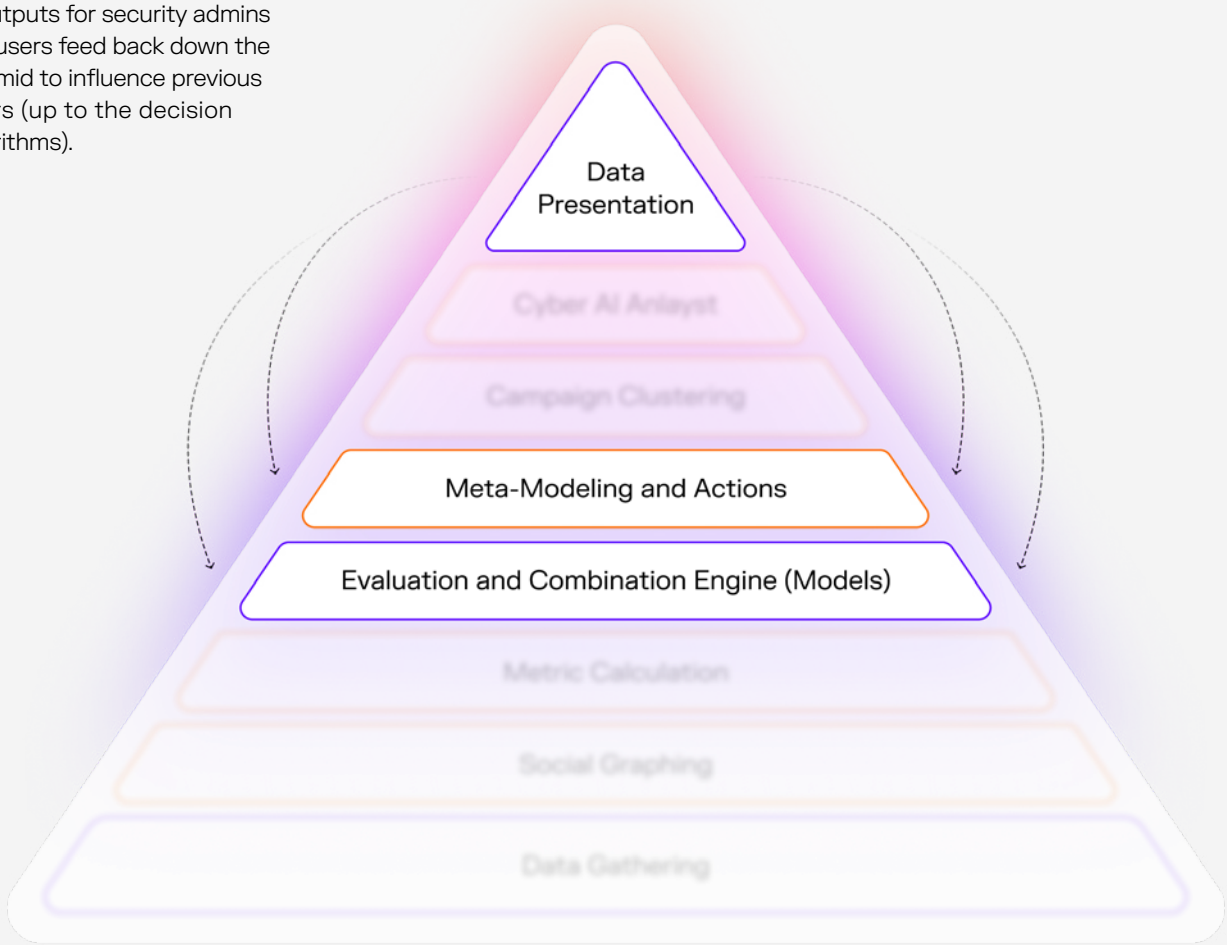
"In less than one month, **Darktrace saved the company 264 analyst hours** spent on investigations, only escalating 8% of suspected threats for further review."

■ **Darktrace Customer**
Utility Management

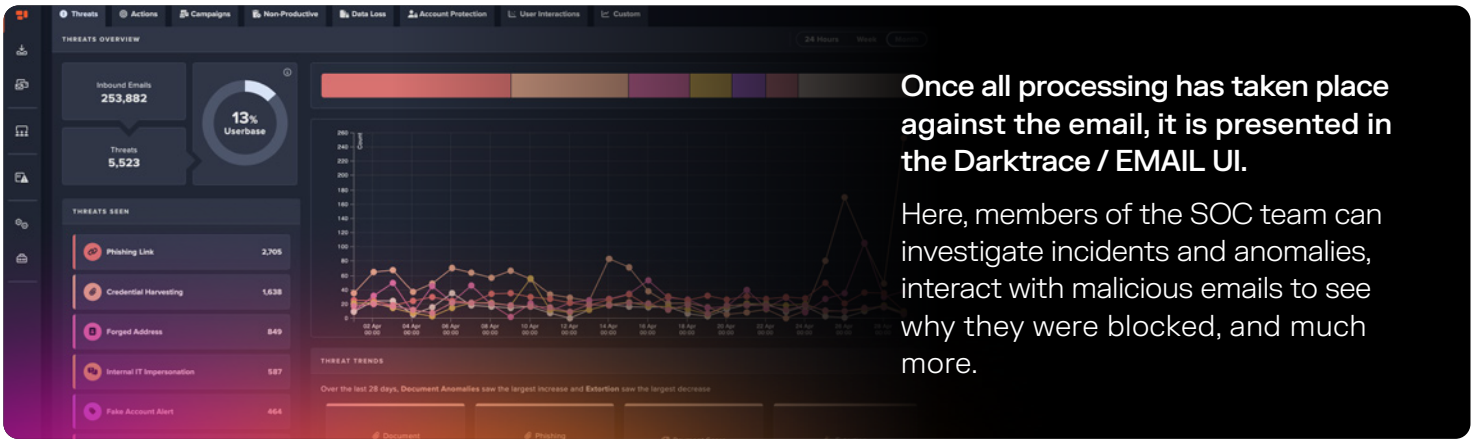
LEVEL 08

Data Presentation

At the top of the pyramid, AI outputs for security admins and users feed back down the pyramid to influence previous layers (up to the decision algorithms).



AI LAYERS ALGORITHMIC LAYERS (NON-AI) --- USER/ADMIN FEEDBACK



Once all processing has taken place against the email, it is presented in the Darktrace / EMAIL UI.

Here, members of the SOC team can investigate incidents and anomalies, interact with malicious emails to see why they were blocked, and much more.

AI Augmented SOC and Tunability

Darktrace / EMAIL provides a multitude of functions for members of an organization's Security Operations Centre (SOC) which allow for system tunability. These functions are intended to help with organizational efficiency and allow Darktrace to process emails more accurately.

Dynamic Lists

Darktrace / EMAIL maintains four dynamic mailbox groups which each have their own classifier. These groups are "High Profile Users" or VIPs, Public-Facing Mailboxes, External Sources of Internal Domain Mail, and User Account Subscriptions.

Membership within these lists can affect a variety of metric and model outcomes. Members of a SOC team have full access to these lists and can make manual adjustments as required to tune the system

Darktrace / EMAIL can save time and effort for SOC teams by automatically identifying and grouping important or high-risk groups of users together. Additionally, / EMAIL can utilize these groupings to calculate more specialized metrics which are tailored to the different types of emails each group is expected to receive.

VIPs

VIP users are detected using a novel system developed by Darktrace. The system takes a series of metrics such as the mailbox exposure score, frequency, source, and destination of inbound and outbound emails, sending behaviours, and participation in groups containing other VIPs to determine the likelihood that a given user is a VIP. If available, Darktrace / EMAIL can also utilize job titles and descriptions for users provided through integrations, factoring this information into the calculations (i.e. C-Suite positions).

Members of an organization's SOC team can confirm or deny if users on the generated list are actual VIPs using the Darktrace console. By confirming if a flagged user is a VIP, / EMAIL will refine its calculations and continue to adapt better to an organization's structure.

Similar classifiers are used to identify mailboxes in the other dynamic lists which are available within the Darktrace console.

VIPs are targeted x5 more often than other email users

(Darktrace Annual Threat Report 2026)

Learning Exceptions

Every organization has different behaviors, expectations, and processes. Email addresses sending content which one organization may find disruptive or inappropriate, may be seen as another organization as important or essential for business function. False positives like this can impact an organisation's efficiency, so Darktrace allows for a certain level of fine-tuning.

Learning exceptions can be configured by SOC teams to allow certain emails to bypass the actions which / EMAIL would typically perform. When applied, these take effect on detection behaviour instantly. Learning exceptions are configured with an anomaly threshold and other parameters that are automatically identified and suggested by EMAIL based on the characteristics of the specific email in question, meaning that the email will be exempt from actions unless the anomaly score is above a certain value and other conditions are unmet.

All regular processing occurs on emails with a learning exception, including metric and anomaly score calculation and the granular nature of the learning exception ensures that organizations stay safe from BEC and vendor compromise, even if they have a learning exception configured for a specific email address.

Security Mailbox Assistant

Darktrace also provides functionality to automate analysis of reported emails. A user can submit an email for analysis through the Inbox Analysis Add-In (mentioned below), forwarding the email to a security mailbox, or through any other 3rd party integrations.

When an email is sent for analysis, Darktrace / EMAIL will automatically produce a report about the email and send it to the user. This report provides the end user with information about email authentication, anomaly indicators, and information about the sending domain. Additionally, requesting analysis of an email will automatically perform a second stage triage on any links, as described in section 06 Actions.

By enabling this functionality, SOC teams can save a lot of effort by automating their first-line processes and reducing the time spent analyzing emails their employees submit.

End-users can also use the Inbox Analysis Add-In to report an email as safe or as undesirable. When a user reports an email, the email will appear in the console with a "user reported" tag (positive or negative tag depending on report type). Darktrace / EMAIL will keep track of manually reported emails on a per-user and per-domain basis. If it's found that a particular address or domain is getting reported frequently and consistently, / EMAIL automatically uses this data to influence decision making on future emails.

"Within just one month of using Darktrace / EMAIL, the volume of suspicious emails requiring analyst attention dropped by 75%, **saving analysts 45 hours per month** on analysis and investigation."

- **Darktrace Customer**
  Global Tech Provider

AI for End Users

Darktrace / EMAIL doesn't just provide AI in its detection systems and autonomous response, but also provides AI functionality which directly assists end-users. This functionality is intended to increase efficiency and reduce the load of an organization's SOC team.

Non-Productive Email Classification

During processing, Darktrace may detect certain emails as "graymail." This refers to benign and non-harmful emails, typically coming from a legitimate sender, which is still spam-like in nature. These types of emails include, for example, newsletters, external announcements, or advertisements.

Darktrace / EMAIL integrates into employee mailboxes and can move emails tagged as graymail into their junk folder. Reducing clutter in employee inboxes improves productivity and reduces the amount of time employees spend manually filtering their emails. Darktrace identifies emails which are consistently sent to junk by multiple members of an organization and uses this information to update its graymail detection models. This ensures that the system is tailored to an organization but requires consistency from multiple users to ensure that important emails aren't sent to the junk incorrectly.

60%

- **Reduction**
  in end-user phishing reports in organizations using the inbox-analysis add-in (Darktrace Internal Research)

Misdirected Email Analysis

Misdirected Email Analysis is triggered every time a user attempts to send an outbound email. By identifying the typical sending behaviours of an individual, and which users and domains they typically contact, Darktrace / EMAIL - DLP can identify emails which are about to be sent to the incorrect address.

A user may accidentally mistype an address or domain. If this happens, Darktrace will identify the deviation from the user's pattern of life and will be able to identify the similar addresses the user typically would send to.

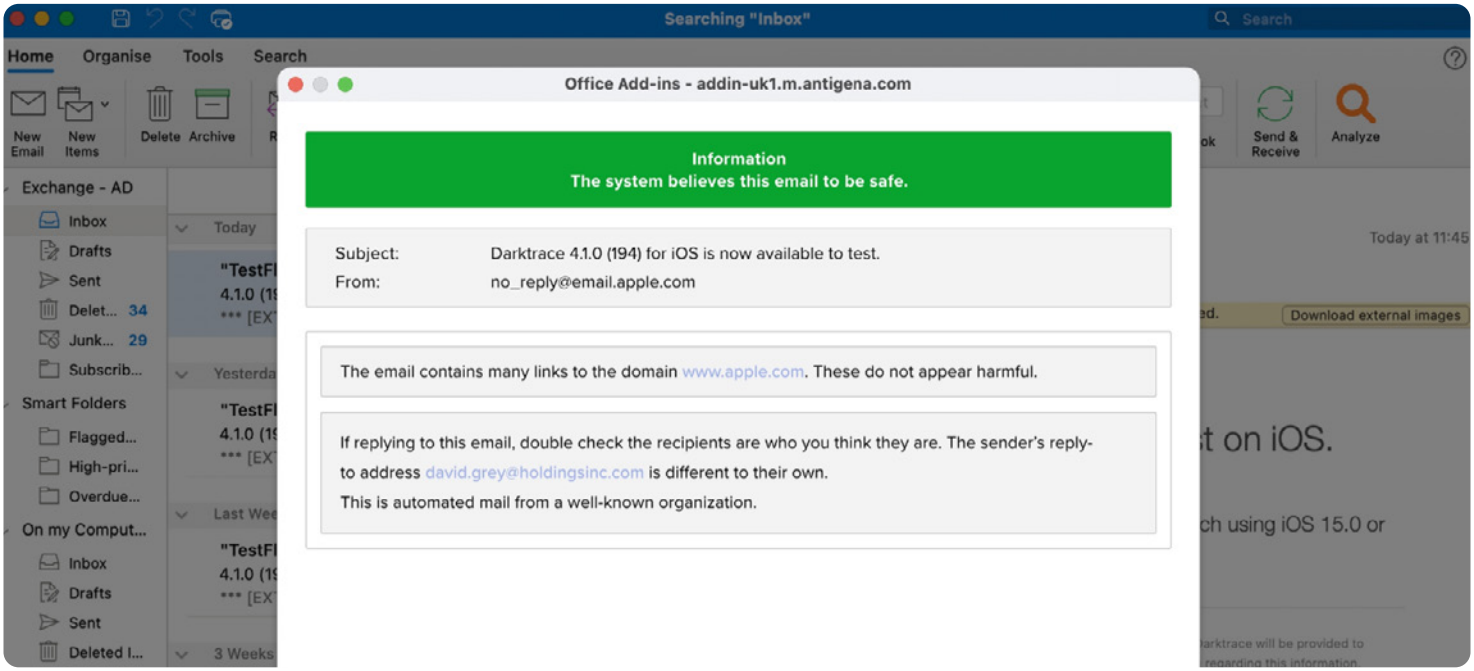
If / EMAIL suspects that the user has misdirected an email, their email will be directly held or a notification will be sent to them (depending on admin configuration).

Inbox Analysis Add-In

Darktrace / EMAIL end users can submit emails for additional analysis if they believe they are suspicious, or if they want to receive a report about a given email. This is performed using an Add-In available for Microsoft 365 (Outlook) mailboxes.

Teams which aren't using Outlook will still be able to use this functionality by instead forwarding emails to a security mailbox for analysis.

If a user sends an email for analysis, Darktrace / EMAIL will produce a report, based on the AI Analyst overview, and send it to the user. The report allows users to receive information about why an email is or isn't detected as malicious. This can help educate end-users about mailbox risks and reduce the chance of email-driven breaches by allowing end-users to more easily stay informed about risks.



Users can click "Analyze" to view an AI narrative explaining why an email has been flagged, building their confidence and improving reporting quality.

Conclusion: Unifying intelligence across the enterprise

Email remains the most pervasive and effective attack vector for cyber-attacks, a reality reflected consistently across industry reporting. In 2025 alone, 83% of malware-related breaches were driven by email, while 54% of ransomware attacks originated from phishing. [7]

These figures underscore a critical truth for defenders. Email threats are no longer isolated, static, or purely signature-based; they are adaptive, socially engineered, and often unfold across users and identities. As a result, effective defense cannot rely on a single model or detection technique. It requires a multi-layered approach that combines diverse and complementary AI methods, contextual understanding, and automated response to address threats at every stage of the attack lifecycle.

As explored in this paper, multi-layered AI enables deeper visibility, more accurate decisioning, and faster containment across email, messaging, and account activity. In an environment where email continues to be the primary delivery mechanism for cyber risk, multi-layered email defense is no longer optional. It is a foundational requirement for resilient security operations.

From there, connecting AI insights across the wider digital estate is the next step in delivering a comprehensive security system. Unifying email security with endpoints, identities, cloud services, and network activity delivers a more resilient security posture that protects organizations from the initial entry vector to their IP, customer data, critical systems, and key personnel.

Ready to see what Darktrace's multi-layered AI could find in **your email environment?**

Get a demo ↗

Or calculate your annual ROI potential and security benefits with Darktrace / EMAIL

Calculate ↗

7 https://www.verizon.com/business/resources/reports/dbir/

About Darktrace

Darktrace is a global leader in AI cybersecurity that keeps organizations ahead of the changing threat landscape every day. Founded in 2013 in Cambridge, UK, Darktrace provides the essential cybersecurity platform to protect organizations from unknown threats using AI that learns from each business in real-time. Darktrace's platform and services are supported by 2,700+ employees who protect nearly 10,000 customers globally. To learn more, visit www.darktrace.com.

North America: +1 (415) 229 9100 | Europe: +44 (0) 1223 394 100 | Asia-Pacific: +65 6804 5010 | Latin America: +55 11 4949 7696

darktrace.com | info@darktrace.com

© 2026 Darktrace Holdings Limited. All rights reserved.