DARKTRACE
Evolving threats call for evolved thinking

Managing Autonomous Decision-Making

The Relationship Between Humans and AI

CONTENTS

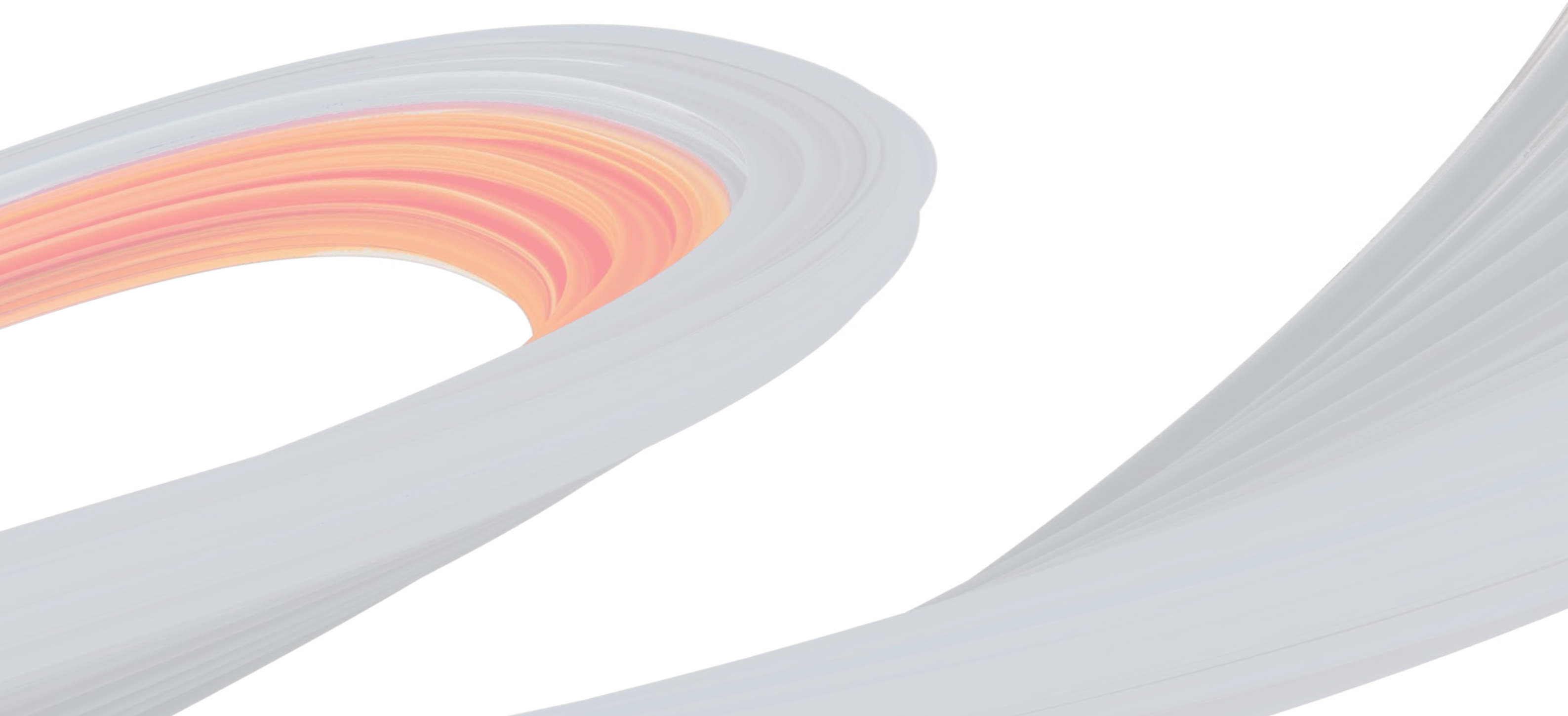
/ Abstract 1
/ Factors Driving the Move to Autonomous Systems 2
/ Raising Decision-Making to a New Level 3
/ Managing Autonomous Security: Four Models 3
/ Human-in-the-Loop (HITL) 4
/ Human in the Loop for Exceptions (HITLFE) 5
/ Human on the Loop (HOTL) 6
/ Human out of the Loop (HOOTL) 6
/ Cyber AI Loop 7
/ Implications of These Models Beyond Cyber 7
/ Autonomous Response 8

/ Abstract

Autonomous systems are becoming a critical component of cyber security, sparking vital conversations about the relationship between human security teams and advanced technology. What level of trust should be granted to an AI system taking autonomous action to stop cyber-attacks? At what point do security teams intervene in its decision-making?

Thousands of organizations now run Cyber AI technology in fully autonomous mode, yet the question of how human beings effectively manage AI technology remains critical. AI and machine learning combined with human insight can, and often does, augment the value of both sides of that equation.

The journey towards autonomous security is likely to differ according to company size, industry,and other considerations. Regardless, organizations have a number of possibilities in making the move to autonomous systems, which give human operators varying degrees of control and oversight. This paper explores four models underscoring the options available: Human in the Loop, Human in the Loop for Exceptions, Human on the Loop, and Human out of the Loop.



/ Factors Driving the Move to Autonomous Systems

As cyber-criminals exploit new opportunities presented by more complex digital infrastructure and wider attack surfaces, they can more effectively get creative and even shorten their attack time. We've seen this with the growing proliferation of sophisticated targeted phishing emails through to the attacker's successes with ransomware, a multistage attack that almost always represents a failure (or lack of) of existing defenses at many different points in the kill chain.

With attackers growing bolder with their ransom demands, and more easily able to cause disruption that often dwarfs the ransom payment itself in terms of costs, the costs of incurring a successful cyber-attack are soaring – the latest report from the Ponemon Institute found that the average cost of a data breach is now $4.24 million[1] – up 10% from 2019.

Even as threat actors innovate, the cyber security industry has broadly continued to take the same approach – with security teams inundated with crafting rules and policies in an attempt to predict future techniques of attackers, usually based on what they've done in the past.

When an attack is detected, either an automated system would issue a pre-programmed action or a human operator would run a series of pre-planned playbooks to 'undo' the attack step by step. These typically take too long, prove inadequate and miss part of the attacker's movements. Blanket response mechanisms fail to contain real-world attacks, which are constantly being tweaked and improved by determined and creative attackers.

Due to the complexity of modern digital infrastructure, thousands of micro-decisions now need to be made daily to match an attacker's spontaneous and erratic behavior to stand a fighting chance at avoiding cyber disruption. Business leaders are recognizing that this far exceeds what can typically be expected from even large teams of human operators, and this has led to a growing conversation around looking beyond automation and towards autonomous systems which can independently assess a cyber-attack and calculate the best possible action to take in any new threat scenario.

Autonomous decision-making based on AI and machine learning was introduced to the cyber security market in the form of Darktrace RESPOND in 2017, and early successes in containing never-before-seen threats – impossible to neutralize with manual pre-programming – led to its rapid adoption across all industries and all corners of the globe. This system took a new approach with Self-Learning AI, based on understanding the unique business as a bespoke entity and responding to subtle deviations indicative of a cyber threat without disrupting the day-to-day business. The technology, as a result, autonomously identifies and neutralizes both known threats and those unknown and unpredictable attacks not covered by blanket policies and deny-lists. Darktrace RESPOND has since been expanded to counter threats in applications, email, the cloud, SaaS, industrial environments, and endpoints – offering coverage of the digital infrastructure, regardless of where data and digital assets are located.

[1] https://www.ibm.com/security/data-breach

/ Raising Decision-Making to a New Level

With autonomous systems in cyber security, human operators are raising their decision- making to another level. Instead of struggling to make an increasingly unmanageable number of 'micro-decisions' themselves, they now preside over the logic, rules, and constraints that AI machines should adhere to when making millions of granular 'micro- decisions' at scale. By establishing the constraints and zones in which the algorithms may operate independently, organizations can become comfortable letting the system run on its own within those parameters. Human operators are no longer setting the rules and policies for specific cyber threats, but are now plotting out business priorities and setting guiderails for the AI system to act.

Once human operators are happy with the boundaries established – perhaps following a period of the AI being set to passive or 'human confirmation' mode – they find life is suddenly different in many ways. They no longer manage at a micro-level but at a macro-level: their day-to-day tasks become higher-level and more strategic, and they are brought in only for the most essential requests for input or action.

With autonomous technology, the role of the human security team, and the profile for a single security team member's day-to-day, has shifted from mundane, hand-to-hand combat, to strategic macro-decisions. In essence, it achieves for security what Steve Jobs saw the personal computer achieving for the average person:

> People are freed to think about the conceptual issues involved and the creative issues involved and use the computer actually to plow through the drudgery. And we're actually changing job descriptions based on allowing people to do more creative work, rather than more work-work

/ Managing Autonomous Security: Four Models

When assessing models for managing AI decision-making to stop cyber-attacks, it is important to bear in mind two facts. First, not every AI detection will have a corresponding response action. When a deviation from normal occurs, Darktrace RESPOND will assess whether this is indicative of an attack, or simply something that is 'unusual but benign'. It does this by drawing on another AI engine – one Darktrace has branded 'AI Analyst' for its ability to replicate human analyst thought processes.

When an unusual event occurs, this AI system asks additional questions around the incident to determine whether this was part of a larger nefarious incident. It then generates an incident summary for security teams to review and action if necessary. But this extra layer of investigation also informs Darktrace RESPOND's micro-decision-making framework to ensure it does not take action for unusual but benign events.

Second, not every AI response action is drastic – a large portion, in fact, are surgical interventions (blocking a certain specific connection over a certain port, or intelligently re-writing an unusual link in an email, for example). These are actions or events that the user may not even notice. AI responses based on micro-decisions are proportionate and incremental; as the impeded attacker gets creative and attempts new ways to progress, these actions may become more aggressive.

Four management scenarios set forth possibilities for varied interaction between humans and machines. These scenarios correspond with categories explored in the Harvard Business Review article, "Managing AI Decision-Making Tools" [2]. The article explains how the rise of AI in the digital world has enabled organizations to operate at scale and make millions of decisions every day, but that phenomenon has also "required a complete paradigm shift, a move from making decisions to making 'decisions about decisions'".

We take this thinking a step further by analyzing how they translate to Autonomous Response in cyber security and provide important insight into the best ways to use Darktrace RESPOND.

In the process we explore implications for the human operator and operations of the business.

○ Human in the Loop (HITL)

○ Human in the Loop for Exceptions (HITLFE)

○ Human on the Loop (HOTL)

○ Human out of the Loop (HOOTL)

[2] Michael Ross and James Taylor，"Managing AI Decision-Making Tools，" Harvard Business Review，Nov 10，2021

/ Human-in-the-Loop (HITL)

In this scenario, the human is, in effect, doing the decision-making and the machine is providing only recommendations of actions, as well as the context and supporting evidence behind those decisions to reduce time-to-meaning and time-to-action for that human operator.

In the cyber world, we see this as the equivalent of our Autonomous Response technology being deployed in passive mode. In this set up, the AI does all the micro-analysis and formulates the appropriate response before handing the action, with context for why it would take such an action, for the human to activate.

What this means for the security team

Under this configuration the human operator – likely a member of the security team – is in full control. They have complete autonomy over how the machine does and does not act. For this approach to be effective in the long-term, sufficient human resources are required. Often this would far exceed what is realistic for an organization.

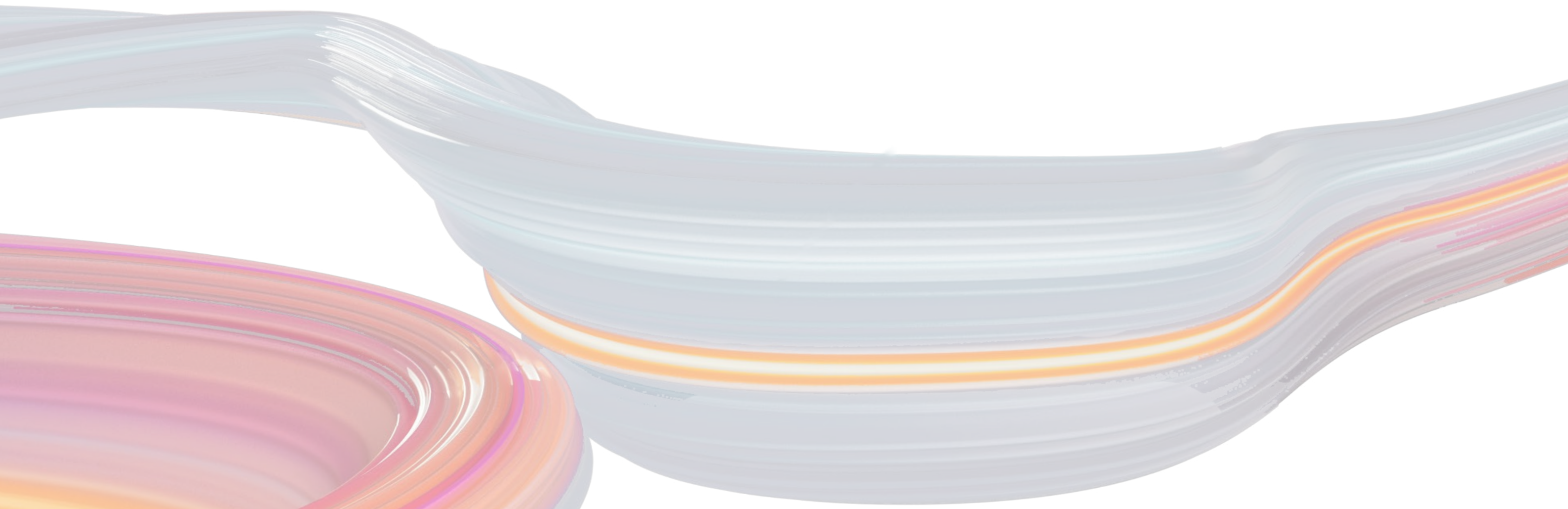
This 'passive mode' is the default setting for new Darktrace RESPOND trials, as the AI very quickly becomes familiar with the digital estate and the user builds trust in the decision-making. In more cases than not, this is a temporary arrangement before the security team entrusts Darktrace RESPOND to make decisions on their behalf – a capability which can be crucial in the context of fast-moving ransomware striking overnight or over holidays.

The result of Human-in-the-Loop is that the organization continues to operate, but in a way which demands significantly more time and resources from the user and risks time-to-action being too slow.

In some cases, an action recommended by a machine might on the surface appear to be counterintuitive, when in reality, the AI Autonomous Response capability has come to this decision based on thousands of factors and metrics, and a deep analysis that goes far beyond that which can be expected of a human, and in fact is the most appropriate action to take given the wider context of the incident.

In these cases, an unwise human intervention in the machine's decision-making may be the difference between ransomware being neutralized and being deployed, or the difference between industrial systems performing as they should and the failure of critical infrastructure services like oil pipelines or electricity grids.

Yet for organizations coming to grips with the technology, this stage represents an important steppingstone in building trust in the AI Autonomous Response engine.



/ Human in the Loop for Exceptions (HITLFE)

Most decisions are made autonomously in this model, and the human only handles exceptions. For the exceptions, the system requests some judgment or input from the human before it can make the decision. Humans control the logic to determine which exceptions are flagged for review.

With increasingly diverse and bespoke digital systems, different levels of autonomy can be set for different needs and use cases. Some teams may deploy Darktrace AI in fully autonomous mode in some areas of their digital estate, for example, while across Operational Technology (OT), a security team may give more consideration before handing over full control to a machine due to safety concerns.

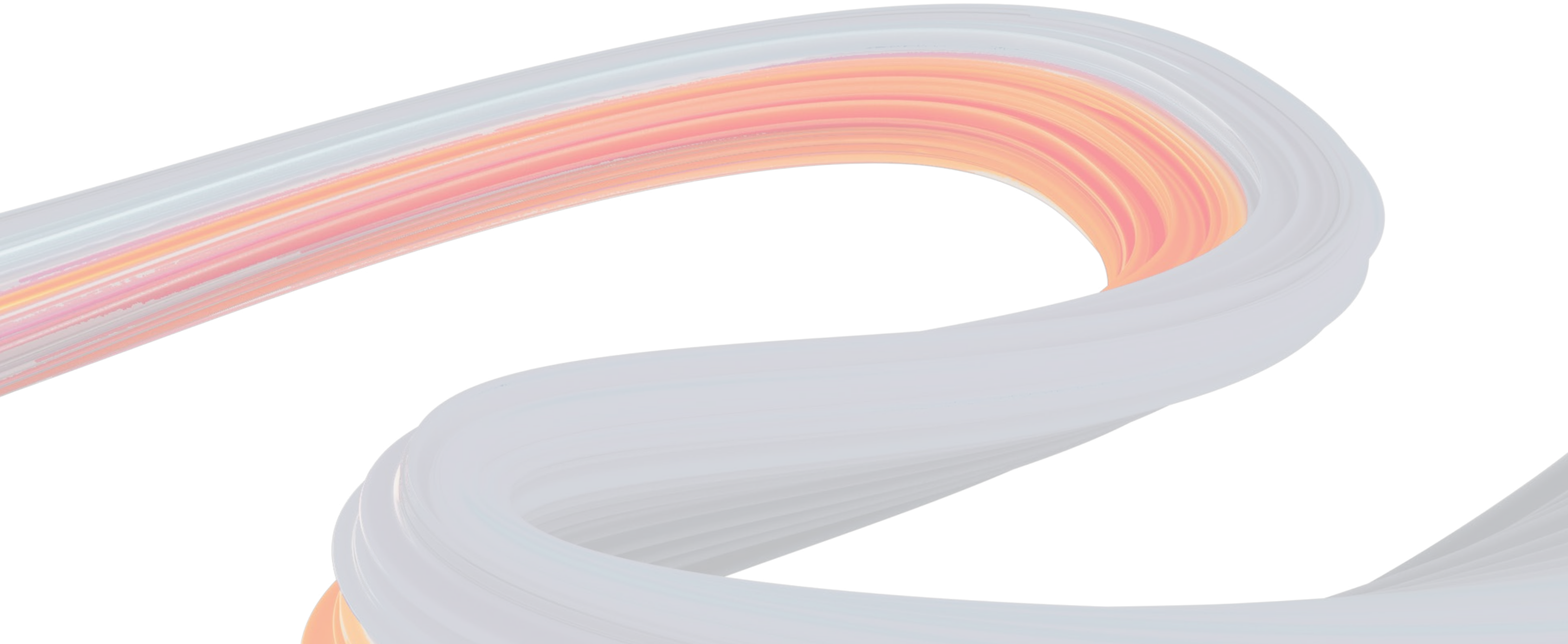
With Darktrace's constraint settings, a human operator can get as granular as they need to set the exceptions: which types of anomalous activity should an exception be considered, for which devices, users, or accounts, at which times (some organizations choose to set the system to be fully autonomous during nights, weekends, and holidays and like to be brought into the loop during the working day).

This results in what can be thought of as 'zoning', where no action would ever be taken unless it's in an 'enforcement zone', and actions may require a human's input depending on the configuration of the zone. The bulk of this 'zoning' is usually done in the initial deployment, with the system able to autonomously expand and contract the zones as environments scale and shift over time to remain in line with business priorities or demands.

What this means for the security team

This means that the majority of events will be actioned autonomously and immediately by the AI-powered Autonomous Response but the organization stays 'in the loop' for special cases, with flexibility over when and where those special cases arise. They can intervene, as necessary, but will want to remain cautious in overriding or declining the AI's recommended action without careful review.

As digital infrastructure and business priorities change, and new initiatives like zero trust become mainstream, the user remains in full control of the zones in which autonomous action is taken, and those which need to be monitored by a human.



/ Human on the Loop (HOTL)

In this case, the machine makes the micro-decisions and takes all actions, and the human operator can review the outcomes of those actions to understand the source of the anomalous, yet contained, behavior.

This is the equivalent of Darktrace RESPOND being set up in autonomous mode – the most common and ideal configuration for the technology. The AI engine is left to make decisions and carry out actions, but at any point, the human can review already-made decisions.

In the case of an emerging security incident, this arrangement allows autonomous actions to stun a threat actor in place, while indicating to a human operator that a device or account needs support, and this is where they are brought in to remediate the incident, with the work of Autonomous Response more or less complete. Additional forensic work may be required, and if the compromise was in multiple places, Autonomous Response may escalate or broaden its response.

What this means for the security team

For many, this represents the optimal security arrangement. Given the complexity of data and scale of decisions that need to be made, it is simply not practical to have the human in the loop (HITL) for every event and every potential vulnerability. This is particularly the case given the speed, volume, frequency and sophistication of cyber threats. With this arrangement, humans retain full control over when, where, and to what level the system acts, but when events do occur, these millions of 'micro-decisions' are left to the machine.

This arrangement also highlights the beauty of Autonomous Response: as its actions are proportionate to the detected activity, initially a light action may be taken, but as an attacker gets creative and attempts a new movement, increasingly progressive actions are taken against that device or account.

This is the power of autonomous mode: never-before-seen activity is immediately met with an ideal counter-response that keeps a system operating as intended, and only as required increasing the severity of its actions.

/ Human out of the Loop (HOOTL)

In this model, the machine makes every decision, and the process of improvement is also an automated closed loop. This results in a self-healing, self-improving feedback loop where each component of the AI feeds into and improves the next, elevating the optimal security state.

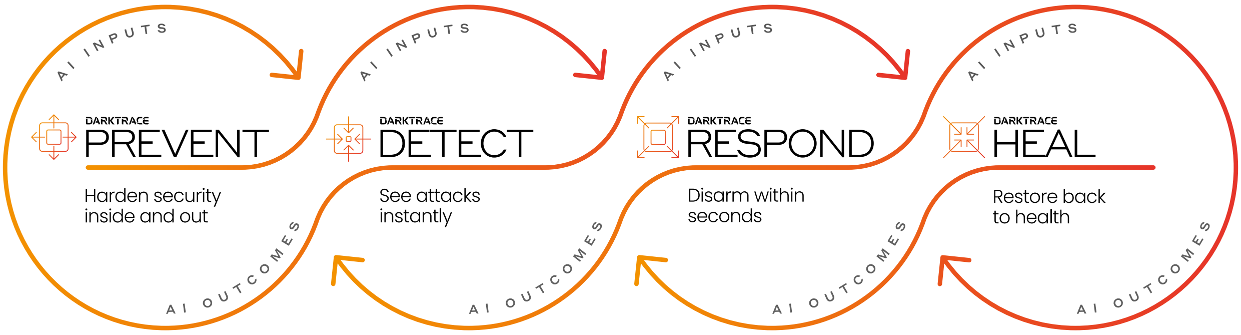
What this means for the security team

This represents the ultimate 'hands off' approach to security. It is unlikely human security operators will ever want autonomous systems to be a 'black box' – operating entirely independently, without the ability for security teams to even have an overview of the actions it's taking, or why. Even if a human is confident that they will never have to intervene with the system, they will still always want oversight.

This is why even as autonomous systems improve over time, an emphasis on transparency will be important. This has led to a recent drive in Explainable Artificial Intelligence (XAI) – in the form of technology like the AI Analyst – that uses natural language processing to explain to a human operator, in basic everyday language, why the machine has taken the action it has.

/ Cyber AI Loop

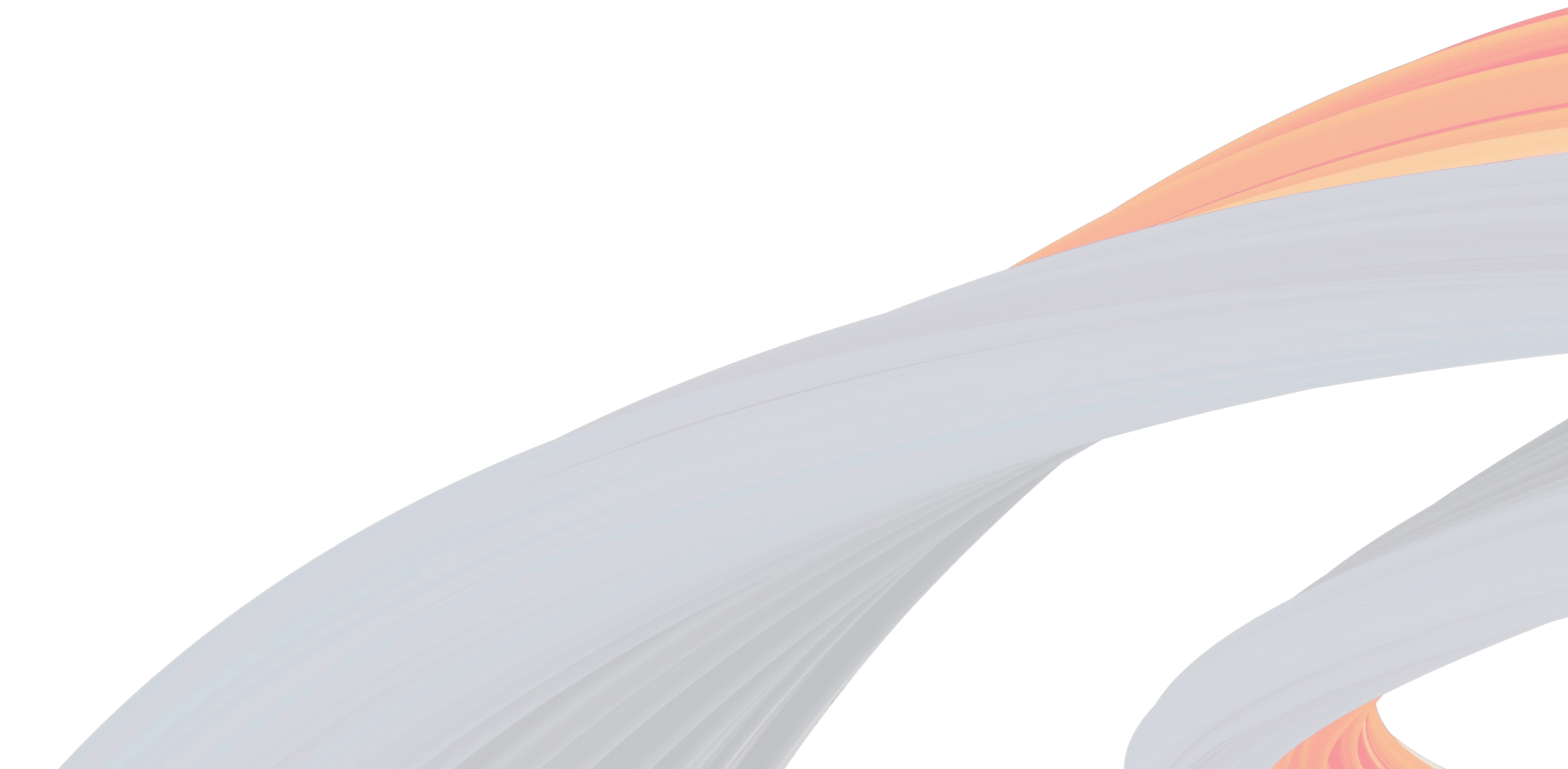
Darktrace RESPOND forms part of Darktrace's technology vision of a Cyber AI Loop, which empowers defenders to reduce cyber risk and disruption at every stage of the attack life cycle – from proactive measures taken to harden security before an attack gets in, to detecting and containing an attack, through to ultimately healing in the aftermath of a breach.



At each of these stages, it is vital that insights are shared with the wider technology ecosystem, continuously improving the state of cyber security for the organization. But it is equally important for a human operator to understand, at every step, what the AI found, what action (if any) it chose to take, and why. To this end, Explainable AI is hugely valuable in generating natural-language reports that can be quickly and easily understood by anyone – from a new IT starter to a board member.

/ Implications of These Models Beyond Cyber

Effective and practical applications of AI have improved productivity and efficiency across many walks of life, from healthcare to traffic light sensors and waste management systems, and cyber is just one area where the question of how humans and AI operate and interact is relevant.



/ Autonomous Response

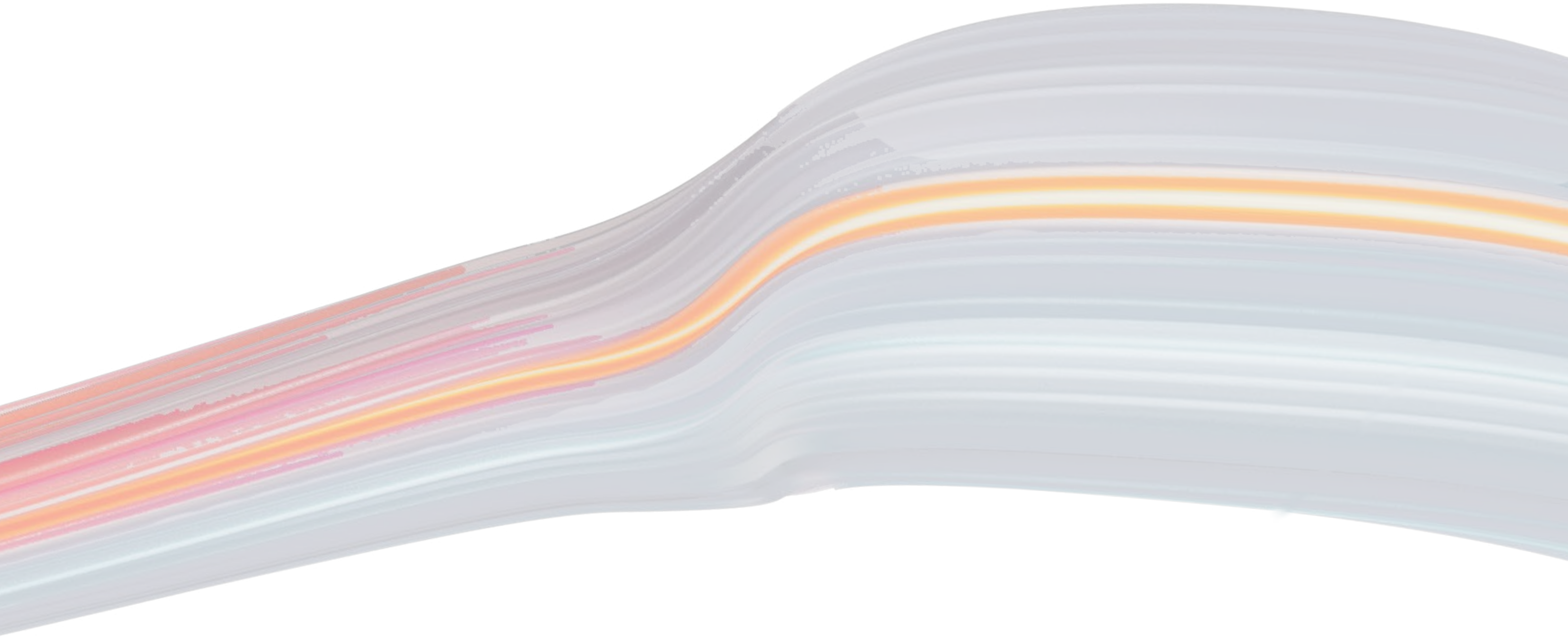
In the realm of cyber, accurate and proportionate machine-led decision-making is only possible with an intimate and evolving knowledge of the unique digital infrastructure the AI is tasked with defending.

Only when that system has this understanding of 'self' can it make effective micro-decisions regarding ever-changing data streams. Only with complete visibility into the entire digital ecosystem can it determine the best action (if any) to contain the threat in the fastest and most appropriate manner, without disrupting regular business operations.

Not all AI is created equal, and this is what makes Autonomous Response unique from others who claim detect and responds capabilities. It is not looking at historical data sets, with an additional layer of machine learning bolted on. It takes never-before-seen events that couldn't possibly have pre-programmed responses and self-determines whether these new events are indicative of a fast-changing, dynamic business or of cyber-threat that needs neutralizing.

It is then designed to enforce normal operations, taking the minimal action required at each phase of the attack to contain the attack. It can also be used to enforce security behavior that an organization wants to have happen (i.e., never let this device talk to this other device; don't let a human do this; helpstay in compliance with a regulation).

These four models described in this paper all have their own unique use cases, so no matter what a company's security maturity is, the CISO and the security team can feel confident leveraging a system's recommendations, knowing it makes these recommendations and decisions based on micro-analysis that goes far beyond the scale any single individual or team can expect of a human in the hours they have available. In this way, organizations of any type and size, with any use case or business need, will be able to leverage AI decision-making in a way that suits them, while autonomously detecting and responding to cyber-attacks and preventing the disruption they cause.



About Darktrace

Darktrace (DARK.L), a global leader in cyber security AI, delivers complete AI-powered solutions in our mission to free the world of cyber disruption. We protect more than 7,400 customers from the world's most complex threats, including ransomware, cloud, and SaaS attacks. Darktrace is delivering the first-ever Cyber AI Loop, fuelling a continuous security capability that can autonomously spot and respond to novel in-progress threats within seconds. Darktrace has 115+ patent applications filed. Darktrace was named one of TIME magazine's "Most Influential Companies" in 2021.

Scan to
LEARN MORE

DARKTRACE
Evolving threats call for evolved thinking

North America: +1 (415) 229 9100
Europe: +44 (0) 1223 394 100
Asia-Pacific: +65 6804 5010
Latin America: +55 11 97242 2011
info@darktrace.com

darktrace.com