DARKTRACE

Darktrace / SECURE AI

Feature & Capability Guide

Behavioral Security for AI

AI requires a different approach to security. Traditional security methods were designed to protect deterministic systems, where a control can be written in advance for a condition someone anticipated. But AI is different. It operates in natural language, can take multiple paths to the same outcome, and increasingly acts on its own. There is no artifact to catalogue in a prompt, no CVE for an agent persuaded to exceed its purpose, and no fixed set of guardrails that can account for every adversarial interaction. Security policies and controls remain essential for defining what AI systems and agents are allowed to do, but they cannot determine whether a sequence of seemingly legitimate actions signal a jailbreak, prompt injection, or an agent gradually drifting beyond its intended purpose. As AI becomes more autonomous and more widely adopted across the enterprise, securing it requires understanding behavior, context, and intent, not simply enforcing predefined rules.

Darktrace / SECURE AI applies behavioral security to provide visibility into how prompts, agents, and AI systems operate within the context of each organization, helping security teams identify risks before they escalate into security, compliance, or operational issues. Powered by Adaptive AI, Darktrace's proprietary technology, it learns how AI is used across the organization to identify subtle signs of risk, including agent drift, over-permissioning, misconfiguration, policy violations, and unexpected behavior. By understanding how prompts, agents, and AI systems normally operate within the business, it can uncover sensitive data exposure, prompt injection attacks, agent misuse, and other emerging threats that static controls and predefined rules may miss. With oversight across sanctioned AI, autonomous agents, development environments, and Shadow AI, Darktrace helps organizations strengthen governance, support compliance, proactively manage AI risk, and prevent threats from escalating into incidents as AI adoption scales across the enterprise.

Business Value Overview

Outcome	What / SECURE AI Delivers
Adopt AI with confidence	Turn AI security from a barrier to adoption into an enabler, with the visibility, behavioral understanding, and governance needed to scale AI while maintaining control over risk.
Stop threats before they escalate	Behavioral analysis identifies subtle deviations in AI activity that may signal prompt injection attempts, malicious chaining, or agent misuse, helping teams investigate and manage risk before it becomes an incident.
Protect sensitive data	Detection of sensitive information appearing in prompts and sessions, with risk-ranked sessions that show where exposure is concentrated.
See the AI you did not know about	Shadow AI discovery across users, devices, and web traffic, distinguishing misuse of approved tools from genuinely unsanctioned services.
Govern in business terms	Upload your own AI policy and have it mapped against real prompt activity, with alignment tracked over time and legislation such as the EU AI Act optionally tagged alongside it.
Fast time to value	A SaaS offering delivered through the Darktrace Platform Portal, connected through APIs with no appliances and no endpoint agents of its own.

Core Capabilities

Darktrace / SECURE AI brings together four core capabilities spanning AI prompts, agent identities and actions, agent development, and Shadow AI. Each addresses a distinct area of AI risk, while Policy Manager provides a consistent governance layer across all four by connecting organizational policy with real-world AI activity.

Governance you can actually evidence

Policy Manager. Upload your organization's AI policy and have it automatically interpreted into policy-aligned rules that are mapped against real prompt activity. Rather than manually translating governance documents into technical controls, Policy Manager operationalizes existing policies, continuously evaluates AI activity against them, and highlights where behavior aligns with expectations or may require attention. Each flagged session is linked back to the relevant policy and rule, giving security teams context for investigation, a clear audit trail, and a measurable way to track governance across the AI estate.

Note: *does not constitute legal advice or a determination of policy or regulatory compliance.*

AI Prompt Analysis

Prompts are the control plane for enterprise AI. Real-time inspection of prompts, sessions, and responses lets teams understand intent, trace what data is entering model interactions, and separate business-aligned activity from risky deviation, including conversational prompt attacks and malicious chaining.

Why it matters: sensitive data now leaves the business through prompts long before it leaves through a file share. Seeing intent as it is expressed lets teams contain exposure at the moment it happens, without switching off the AI tools people depend on.

Feature	Description	Outcome For Your Team
Prompt visibility	Monitoring and analysis of prompts across approved third-party SaaS providers to detect anomalous behavior by both humans and agents, including the ability to view raw prompts – subject to compliance with applicable privacy and employment laws, role-based access, and data minimization.	Investigate AI incidents with the right evidence at your fingertips, while configurable role-based access and user management settings ensure sensitive prompt content is visible only to the teams that need it.
Prompt categorization	Prompts grouped by function, such as coding, HR, market research, and sensitive categorizations.	See which parts of the business are driving AI usage, and focus review effort where the most sensitive work is happening.
Prompt detections	Alerting on attempted jailbreaks, sensitive data in a prompt, and possible indirect prompt injection attempts, surfaced in the SOC dashboard.	AI-specific attacks arrive as alerts in the workflow analysts already use, rather than going unnoticed in a separate console.
Prompt policy alignment	Organizational policy uploaded through the Policy Manager is mapped to prompts to highlight instances of a violation.	The AI policy the business wrote is measured against what employees actually do, so violations surface in days rather than at audit.
Session risk categorization	Sessions ranked Critical, High, Medium, or Trivial, the presence of sensitive information which may include PII, and harmful activity.	A lean team starts with the handful of sessions that genuinely matter instead of reading everything.

AI Agent Identities and Actions

Prompts define intent, but an agent's identity, permissions, and tool access determine the impact of that intent. Over-permissioned or misconfigured agents turn routine interactions into data exposure. / SECURE AI gives identity-centric oversight of agents across suppliers and environments, so teams can see what an agent can reach, which systems it interacts with, and how that changes over time.

Why it matters: when an agent is compromised or simply over-permissioned, the blast radius is whatever that agent can reach. Knowing that scope in advance turns an open-ended investigation into a bounded one and shortens the time it takes to act.

Feature	Description	Outcome For Your Team
Identity audit	Detection and audit of all agents and human users associated with AI activity and applications, including their sessions.	Every agent and user touching AI is accounted for, so nothing operates outside the security team's view.
Context of users and agents	Visibility and visualization of groups, permissions, and roles for both agents and users.	Tell privileged agents from routine ones before an incident, and right-size permissions on the ones that carry real reach.
Real-time insights	Live insight, including alerts, into the AI activity associated with a given user or agent.	Suspicious agent activity is visible while it is happening, not discovered later in a log review.
Low-code agent security	Visibility into prompts associated with an agent, and into low-code agents being created in environments such as Copilot Studio, with monitoring to confirm activity aligns with intended purpose.	Business teams keep building their own agents, and security keeps them accountable, so citizen development does not become an unmanaged risk.
Workflow visualization	Mapping of agent workflow, permissions, and access, so teams can see the scope of an agent's reach and understand the potential impact of a compromise.	The blast radius of any agent is a known answer before an incident, which makes scoping and containment faster.

AI Agent Development Risk Management

Many of the highest-impact AI risks are introduced before an agent ever goes live, during build and configuration. / SECURE AI provides visibility into AI agent identities and build activity across high-code cloud AI platforms and low-code builders, surfacing misconfigurations, over-permissioning, and anomalous development events early enough to remediate them.

Why it matters: correcting an over-permissioned agent during development costs a configuration change, while correcting it in production costs an incident response. Build-time visibility keeps AI projects moving instead of stalling in security review.

Feature	Description	Outcome For Your Team
High-code agent security	Visibility into prompts associated with agents built on cloud AI infrastructure used by developers, such as Amazon Bedrock.	Developer-built agents fall under the same prompt oversight as enterprise assistants, with no separate tooling to buy or run.
Development audit	High-code agents being created are included in identity audits and monitored to confirm activity aligns with intended purpose.	Misconfiguration and over-permissioning are caught while they are still a configuration change, not an incident.
Cloud AI architecture	Architecture views for agents developed and deployed through cloud AI services, generated through the Darktrace / HYBRID NETWORK (/ CLOUD) integration.	Security and cloud teams work from one picture of how AI is assembled, which shortens design review and hand-offs.

Shadow AI Management

Shadow AI appears across every part of the digital estate, from unsanctioned SaaS AI usage to locally deployed MCP servers. / SECURE AI surfaces unapproved AI activity where it appears, distinguishes misuse of legitimate tools from genuinely unsanctioned services, and applies policy to contain data exposure while guiding users toward sanctioned options.

Why it matters: teams cannot govern what they cannot see. Surfacing unsanctioned AI turns a blind spot into a short, prioritized list of services to control and conversations to have, so the business keeps the productivity without absorbing the exposure.

Feature	Description	Outcome For Your Team
Shadow AI detection	Detection of unsanctioned AI services identified through Darktrace / HYBRID NETWORK telemetry and third-party SASE integrations, including detection of local MCP servers as an indicator of potential shadow AI activity.	Unapproved AI use is found where it actually happens, so it can be contained before data leaves the business.
Shadow AI tracking	An overview of all known shadow AI services in internal use, with visibility into trends and usage over time.	Leaders can see whether shadow AI is growing or shrinking and gain the visibility needed to make informed decisions about AI governance and adoption.

Supported Integrations

Coverage of sanctioned AI services is delivered through API integrations with the providers rather than by sitting inline, so there is nothing to place in the path of production traffic. The list below reflects general availability and continues to expand.

Integration Category	Supported Services And Coverage
SaaS AI services	Microsoft Copilot, Microsoft Copilot Studio, Microsoft 365, Salesforce, Claude, and ChatGPT Enterprise, for prompt security, session visibility, identity audit, and low-code agent security.
Cloud AI platforms	Amazon Bedrock, for prompt security, session visibility, identity audit, and cloud AI architecture views.
SASE providers	Zscaler and Microsoft Global Secure Access, expanding shadow AI visibility across users, devices, and web traffic.
Darktrace / HYBRID NETWORK (/ NETWORK, / OT, / ENDPOINT)	Shadow AI and MCP use detection, visibility into AI-related processes on devices, detection of AI activity affecting OT systems, and network-based response actions such as blocking unsanctioned AI usage or quarantining an affected device.
Darktrace / HYBRID NETWORK (/ CLOUD)	Visibility into agents deployed in cloud environments, insight into agent permissions and access paths, detection of over-permissive agents and misconfiguration, and cloud AI architecture views.
Darktrace / EMAIL	Analysis of emails consumed by AI agents as prompts, prompt injection detection, and visibility into agent use of email accounts.
Forensic Acquisition & Investigation	Automated decision-chain reconstruction across prompts, tools, models, and external data sources, for organizations with regulatory, legal, audit, or board-level forensic readiness requirements.

Please note: All three Microsoft Copilot variants are covered, including Chat, Microsoft 365 Copilot, and Copilot Studio.

Delivery and Technical Requirements*

Feature	Description
Delivery model	Darktrace/ SECURE AI is a SaaS product that integrates via API and is accessed through the Darktrace Platform Portal. Portal access is a requirement, and existing Darktrace customers who already have it are the fastest path to value.
Infrastructure	No endpoint agents and no on-premises hardware of its own. Coverage of sanctioned AI services is delivered through provider APIs rather than inline.
Authorization	Each monitored platform is authorized once by an administrator using OAuth 2.0. Microsoft 365 tenants require audit logging to be enabled; it is on by default in most tenancies.
Shadow AI prerequisites	Shadow AI detection draws on network telemetry or a SASE integration. Existing Darktrace / NETWORK and / ENDPOINT customers see these detections populate automatically once portal access is configured; other organizations should discuss coverage options with their Darktrace representative.
Learning period	As with other Darktrace products, the system builds a pattern of life before it reliably surfaces what does not belong. Seven to ten days is typical, and the more actively AI services are used, the shorter that period.
Data handling	Prompt data, user and agent identity information, audit logs, and associated metadata are held in Darktrace cloud-hosted infrastructure. Prompt data is retained for up to twelve weeks. Role-based user management controls who can view prompt content and full policy documents. Current hosting regions are confirmed as part of the commercial engagement, governed by the applicable order and DPA.

* Specifications are current and subject to change. Governed by terms of applicable order, MSA and DPA.

Why Darktrace

Darktrace has spent more than a decade learning how real users, systems, and machine-driven processes behave in dynamic, cross-domain environments, which is precisely what AI security now demands. Using Darktrace allows an organization to build a unique behavioral profile of that organization from its own activity rather than importing a model trained on another organization's, so normal is defined by your business and not by an external threat feed.

Governance frameworks and provider guardrails are necessary but not sufficient. Governance cannot surface the operational risks that arise from real-world AI usage without visibility into prompts, the behaviors those prompts drive, and the identity of the agent acting; guardrails are static while AI is adaptive. / SECURE AI complements the AI investments an organization has already made by verifying how AI actually behaves at runtime and surfacing risk that policy and access controls alone cannot detect.

© 2026 Darktrace Holdings Limited. All rights reserved.