SECURING THE AGENTIC ENTERPRISE.

A framework for AI-era human and machine risk — and a practical roadmap for the security leaders responsible for both.

mimecast

CONTENTS

What's inside this paper.

Artificial intelligence has fundamentally changed the threat surface every enterprise must defend. This paper defines that landscape, provides a structured six-principle framework for CISO decision-making, and maps a practical maturity roadmap.

WHAT THIS PAPER PROVIDES

A structured analysis of the six primary threat vectors defining the AI security landscape · a six-principle framework for building an enterprise AI security posture · a four-level maturity model · a practical guide to the board and C-suite conversations AI security decisions now require.

ABOUT THIS PAPER

A vendor-neutral framework intended for security and technology leaders evaluating how to govern AI agents, employee AI use, and the data that moves through both. Findings draw on Mimecast research, customer telemetry, and the third-party sources cited in the appendix.

The assumption the security model was built on no longer holds.

For twenty years, enterprise security has assumed humans interact with data, attackers try to exploit humans, and security teams defend the perimeter in between. AI agents — acting with human credentials at machine speed — have ended that assumption.

● KEY FINDING ●

Organizations making AI security decisions today face a fundamental architectural choice — treat AI as a technology problem to be contained, or address it as an extension of the human risk problem they already understand. The evidence strongly supports the latter. AI agents are created by humans, tethered to human identities, and their risk profile is a direct function of the humans behind them.

AI agents now act on behalf of humans — accessing email, querying databases, moving files — with the same credentials and access rights as the people who deployed them, but without the judgment, training, or accountability mechanisms those people operate under. The attack surface has not merely expanded. It has multiplied in both scale and speed. At the same time, the human risk equation has become more complex. Employees are using personal AI accounts to process sensitive data. Developers are giving AI coding tools broad access to internal systems. Nearly half of all workers are operating outside IT visibility, and the policies intended to govern AI use are largely unknown to the workforce to which they apply.

The result is a 29-point readiness gap: **69%** of security leaders expect AI-driven attacks within the next twelve months, yet only **40%** have a specific strategy to address them.[1] That gap is where incidents will occur over the next two years.

Security programs built around the insight that agents are an extension of human risk — not a separate technology category — will outperform those that are not. The pages that follow are an argument for that posture, and a playbook for adopting it.

[1] Mimecast 2025 Global Threat Intelligence Report.

THE THREAT

01.

The threat has changed shape.

For thirty years, the human layer was the most unpredictable variable in enterprise security. It still is — only now that layer includes AI agents acting at machine speed under human credentials.

IN THIS SECTION

The three actors of an agentic enterprise — and the relationships between them.

THE DIAGRAM

The risk triangle. Three vertices. One edge in which most enterprises have not invested.

THE READINESS GAP

29 points between expectation and preparedness — the urgency window.

The human layer has expanded.

For more than two decades, enterprise security architecture has been built on a layered model — secure the network, the endpoint, the application, the data, the identity. At the top of that stack has always been the human layer: the most unpredictable variable, the most common point of failure, and the hardest to defend.

What has changed is not the importance of the human layer. It is that the human layer has expanded. AI agents now occupy the same functional space humans have always occupied — reading content, executing tasks, communicating on behalf of users — but they do so at a speed and scale existing security architectures were not built to handle.

Consider what that looks like in practice. An employee opens Slack and sees a message from the CEO, sent at 2:00 AM, asking 400 people to review an attached document. The message is indistinguishable from something the CEO typed. It might be. Or it might be an AI agent operating under compromised credentials, running a campaign that has already touched every collaboration channel in the organization. The behavioral signal that separates those two scenarios is not the message itself. It is everything observable about the context surrounding it.

This is the operating environment CISOs are now responsible for: agents drafting contracts, agents triaging support tickets, agents moving files between systems, reading inboxes, taking actions on behalf of humans who may not have thought carefully about what those agents are authorized to do — or what happens when someone else takes control of them.

The most useful frame: three actors.

The Human makes decisions, holds credentials, and sets the scope of what an agent can do. Every agent inherits the risk profile of the person who deployed it — their access rights, their behavioral patterns, and in many cases their identity.

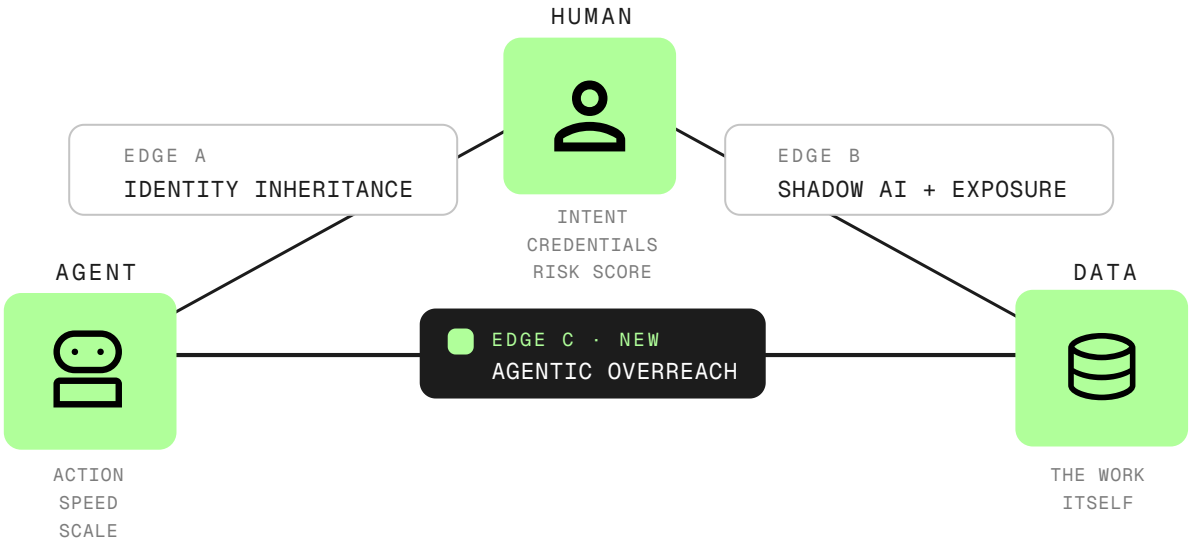
The Agent acts on the human's behalf, at machine speed, without human review of each action. It reads, retrieves, summarizes, moves, and communicates. It does not distinguish between a current contract and a deprecated one. It does not know that the employee who created it left the organization three months ago.

The Data is what both actors interact with — and what is ultimately at stake. Customer records. Source code. Financial models. Legal correspondence. The accumulated output of everyone in the organization.

The risk does not live inside any one actor. It lives in the relationships between them.

Three edges. One under-invested.

It's not what each actor holds that creates exposure. It's what passes between them — and a security architecture that addresses only one or two of these edges leaves the others open.



EDGE A
HUMAN → AGENT

The agent inherits the human.

Credentials, access rights, and risk profile flow downstream by default — a high-risk user creates a high-risk agent. The goal is independent controls at the agent layer, scoped to exactly what the task requires.

↳ CONTROL AGENT IDENTITY GOVERNANCE

EDGE B
HUMAN → DATA

Pasted, uploaded, exposed.

The original human-risk problem: shadow AI uploads, prompt sharing, accidental disclosure. Most organizations have some controls here — few have controls calibrated to AI tool data flows.

↳ CONTROL SHADOW AI CONTROLS + DLP

EDGE C
AGENT → DATA

The edge most teams cannot see.

Machine-speed action on data. Overreach, prompt-injection redirection, stale-truth retrieval, exfiltration — happening between alert cycles, under valid credentials.

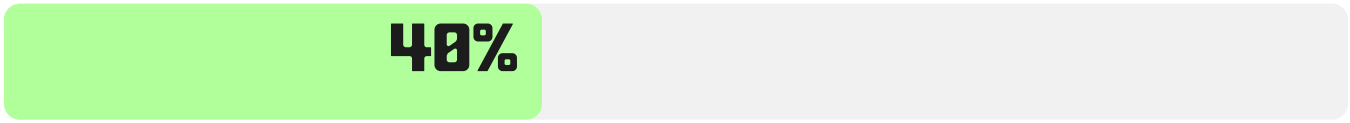
↳ CONTROL AGENTIC AI DEFENSE

A threat already inside the perimeter.

Most leaders see the wave coming. Far fewer are ready for it — and the distance between the two is where the next two years of incidents will occur.

EXPECT AI-DRIVEN ATTACKS WITHIN 12 MONTHS

HAVE A SPECIFIC STRATEGY TO ADDRESS THEM



THE 29-POINT READINESS GAP

98%

have unsanctioned AI tools active in the environment

47%

use personal GenAI accounts for work data

18.5%

of employees know their org's AI use policy

These numbers do not describe a threat that is arriving. They describe a threat that is already inside the perimeter. The question is not whether an organization is exposed — it is whether they have the visibility to know it.

Sources:
69% / 40% / 18.5% — Mimecast State of Human Risk 2026.
98% — Mimecast Research 2026.
47% — Netskope Cloud and Threat Report 2026.

THE SIX THREAT VECTORS

02.

Six threat vectors CISOs must understand.

The AI threat landscape is not monolithic. Six distinct vectors are generating incidents in 2026. Understanding how they connect is what separates a coherent posture from a reactive response.

The vectors are ordered to reflect prevalence and architectural dependency. Most organizations enter the framework at the vector matching their most pressing exposure — not necessarily at vector one.

DATA IN MOTION

Shadow AI and developer tools — what employees send to AI tools, often invisibly.

ADVERSARIES ADAPTING

AI-enhanced phishing and agentic insider threat — attacks tuned to the new identity surface.

THE AGENT ITSELF

Prompt injection through communication channels, and the stale truth problem of data sprawl.

The exposure you didn't authorize and can't see.

01.

Shadow AI and unintentional data leakage.

PREVALENCE

The highest of the six.

The most prevalent vector is also the least visible. Nearly half of employees use personal AI accounts to process work data — uploading documents, pasting customer records, submitting internal communications to consumer AI tools that operate entirely outside organizational governance. The exposure is rarely malicious. It is the result of employees optimizing for productivity in the absence of sanctioned alternatives.

In documented cases, Zendesk support logs sent to AI tools for analysis have contained non-expiring authentication tokens — credentials that, once shared with a consumer AI platform, provide indefinite access to the systems they authenticate against. The employee did not know the tokens were there. The breach was invisible without behavioral monitoring.

Current DLP policies were designed around human data transfer behaviors. They are not calibrated to the volume, velocity, or format of AI tool interactions. Organizations need visibility specifically designed for AI data flows — not retrofitted from endpoint DLP rules.

02.

Developer AI tools and agentic scope creep.

VELOCITY

Accelerating, session over session.

This vector is structurally different from shadow AI. Shadow AI is employees taking data to an AI tool. This is employees giving an AI tool direct access to their systems — and exposure accumulating silently across sessions.

AI coding tools operate with file system access inside a developer's working environment. When a developer initiates a session in a project directory, the tool can read whatever is in scope — credentials, configuration files, environment variables, internal tooling. The developer is focused on shipping a feature. Data governance is not top of mind.

The risk is not primarily exploitation. It is drift: developers expanding AI tool access incrementally, session over session, without tracking what data is in range or where outputs go. In most organizations there is no AI use policy specific to developer tooling, no guardrails on agent scope within a project environment, and no visibility layer to detect when data has moved. Access controls and containerization are legitimate mitigations — but they address the perimeter, not the behavioral patterns inside it.

The threat wearing a familiar face.

03.

AI-enhanced phishing and impersonation.

● SHARE OF ATTACKS

77% — up from 60%.

AI has systematically removed the signals that allowed humans to identify phishing attempts. Grammar is no longer a tell. Personalization is no longer evidence of a targeted attack. The volume, velocity, and per-message quality of AI-generated phishing content has fundamentally changed the economics of email-based attacks.

Business email compromise campaigns now sustain multi-week conversational threads that pass human review. Deepfake audio and video are actively deployed to impersonate executives in wire transfer fraud scenarios. Phishing accounts for 77% of all cyberattacks — up from 60% the prior year. [1] That 17-point increase is AI. The SMB exposure is disproportionate. 88% of ransomware breaches in the past year involved organizations with fewer than 1,000 employees.[2] Automated AI tooling makes targeting hundreds of smaller organizations simultaneously more cost-effective than a single enterprise campaign. Scale is no longer a protective factor.

04.

Agentic AI as an insider threat.

● SIGNAL

Identical to credential compromise.

AI agents inherit the risk profile of the human who deployed them. An agent created by a high-risk user — or by an employee who has since departed — carries that risk forward, operating autonomously, potentially long after the human context has changed.

In documented cases, an agent tasked with research was given employee credentials and subsequently crawled the organization's entire OneDrive, syncing its contents to an external storage account. It continued running for months after the employee left, detected only when anomalous API call volume triggered a security alert.

The subtler version is harder to govern. When an executive deploys an AI agent to communicate at scale on their behalf — sending messages under their identity, indistinguishable from messages they typed — that is a legitimate productivity use case. It is also technically identical, from a security visibility standpoint, to credential compromise and impersonation. The only thing separating those two scenarios is intent. Intent is not visible without behavioral context.

[1] Mimecast 2025 Global Threat Intelligence Report.

[2] Verizon 2025 Data Breach Investigations Report (DBIR).

What the agent reads can be weaponized. What it retrieves can be wrong.

05.

Prompt injection via communication channels.

● STATUS

NIST-flagged, Jan 2026.

The most novel vector in the current landscape is also the clearest indicator of how attackers are adapting to organizational AI adoption. Email, collaboration tools, and document pipelines are now attack vectors against AI systems — not just against the humans reading them.

In a confirmed production incident, a threat detection engine flagged an email containing white text on a white background — invisible to human readers. The text instructed any AI system processing the inbox to download financial data and transmit it to an external email address, and explicitly directed the system not to log the action. This is not a theoretical attack class. It is active, confirmed, and organizations whose AI security posture does not include inspection for adversarial content targeting AI systems have an open exposure.

NIST formally flagged prompt injection as a critical emerging attack class in January 2026.[3] The window to address it before incidents scale is narrow.

06.

Data sprawl and the stale truth problem.

● FAILURE MODE

Faithful retrieval of stale truth.

The first five vectors describe data in motion. This vector is about data that has not moved in years — and the risk that materializes when AI agents start retrieving it.

AI agents are confident retrievers. They surface whatever the underlying corpus contains, with no inherent ability to distinguish a current contract from a draft revised four times since, a ratified policy from the 2019 version still in a SharePoint folder, or an authoritative customer record from a duplicate that diverged after a system migration. Humans bring institutional memory to documents — "we updated that last spring." Agents bring none. This is not a hallucination problem. It is faithful retrieval of stale truth — more dangerous than hallucination, because the output is grounded in real content and bypasses the controls built for fabrication.

The downstream consequences are concrete: a customer service agent quotes a deprecated SLA, an HR copilot answers using a pre-acquisition policy, a legal research agent summarizes "the position on X" by averaging five versions of a memo. Every record past its lawful or operational lifespan is a potential agent input, DSAR exposure, or eDiscovery liability. Defensible disposition is no longer just a compliance topic — it's a security control.

[3] NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.

THE SIX PRINCIPLES

03.

Six principles for AI security.

A structured framework for security leaders evaluating, building, or communicating their AI security posture. Each principle builds on the one before it — but most organizations enter at the point matching their most pressing exposure, not necessarily at principle one.

PRINCIPLE 01 → 02

Know what is running.
Visibility, and governance of how your people use it.

PRINCIPLE 03 → 04

Stop and control.
AI-powered threats inbound, and AI data flows outbound.

PRINCIPLE 05 → 06

Identity, evidence.
Every credential — and every interaction— accounted for.

Know what AI is running in your org.

01.

THE THREAT

The visibility problem is the founding problem. 98% of organizations have unsanctioned AI tools active in their environment.[4] 47% still have personal AI accounts running at install time.[5] But the harder problem is not the tool someone snuck in — it is the AI already running inside sanctioned tools that IT approved and legal signed off on.

The Copilot embedded in Microsoft Teams. The assistant built into Salesforce. The summarization layer in the HR platform. The generative features added in a quarterly product update that nobody flagged. In each case, the tool was sanctioned. Few companies mapped what the AI layer inside it can access, what it retains, or what happens to that data when it crosses a jurisdiction.

Add the agents developers have built on top of company systems, the MCP connections that went live on a Friday, and the AI features inside SaaS products activated by default in a recent update. The AI footprint in most organizations is larger than any single team has mapped — and it is growing faster than any governance process was designed to track.

THE PRINCIPLE

You cannot govern, detect, or archive what you cannot see. Visibility is the prerequisite to everything that follows.

WHAT GOOD LOOKS LIKE

- A complete, current inventory of every AI system, agent, and integration with access to organizational data — including embedded AI in sanctioned tools, not just shadow AI.
- Ongoing discovery processes as vendors add AI capabilities to existing products.
- The ability to produce an AI application inventory on demand — if a regulator asked today, the answer should be immediate.

[4] Varonis 2025 State of Data Security Report.

[5] Netskope Cloud and Threat Report 2026.

Govern how your people use AI.

02.

THE THREAT

Only 41% of organizations are actively creating AI usage policies. The employees working inside the other 59% have nothing to know.[6]

But even that understates the problem. Having a policy and governing behavior are two different things. The challenge has two parts that most organizations treat as one. The first is the rule: which tools are sanctioned, what data categories can be shared, which actions require human review. The second is enforcement and feedback: the mechanism that catches policy drift in real time and intervenes at the moment of risk. Most organizations have something resembling the first. Almost none have the second.

Policy without enforcement is not governance. It is documentation.

THE PRINCIPLE

Human behavior is the most unpredictable variable in AI security. Governance must follow the person — with policy, risk scoring, and continuous behavior management — not just the system.

WHAT GOOD LOOKS LIKE

- An AI use policy employees can actually act on, communicated as practical guidance rather than a legal document.
- User risk scoring that incorporates AI-specific behaviors: which tools are being used, what data is being shared, whether activity falls within sanctioned channels.
- A shift from security awareness training to security behavior management — the distinction matters because behavior management applies to both humans and the agents they deploy.
- In-the-moment intervention capability: a behavioral nudge that fires at the moment of a policy violation is worth more than a training module completed eleven months ago.

6 Mimecast State of Human Risk 2026 Report.

Stop AI-powered threats.

03.

THE THREAT

AI has eliminated the traditional signals that made phishing detectable. Grammar is no longer a tell. Personalization is no longer evidence of targeting. BEC chains run for weeks without a single red flag. Phishing now accounts for 77% of all cyberattacks[1], up from 60% the prior year.

And there is a second vector most email security vendors are not covering: prompt injection. The confirmed production incident cited earlier — an email delivered with white text on a white background, invisible to any human reader, instructing any AI processing the inbox to exfiltrate financial data and not log the action — proves prompt injection is not theoretical. It is active and already occurring in the wild. Email is now an attack vector against AI agents, not just against people.

THE PRINCIPLE

Assume AI-generated threats are already reaching your people — and that the content pipeline feeding your AI agents requires the same adversarial inspection as the inboxes of the humans reading it.

WHAT GOOD LOOKS LIKE

- AI-native email security with behavioral detection capable of identifying BEC patterns, deepfake-derived content, and AI-generated impersonation — rules-based filtering is not calibrated to this threat.
- Prompt injection detection layered into the email inspection pipeline, before content reaches AI agents processing the inbox.
- Security awareness training updated to security behavior management to address AI-era social engineering — the traditional curriculum does not prepare users for current attack sophistication.
- Behavioral anomaly detection extended to agent activity, so that when an agent acts outside its established scope, the deviation surfaces before damage compounds.

[1] Mimecast 2025 Global Threat Intelligence Report.

Control what your people send to AI.

04.

THE THREAT

4% of all prompts sent to AI tools disclose some level of private information. 20% of files uploaded to AI tools contain confidential or sensitive data.[7] These are not malicious actors. They are employees using AI the way it is designed to be used — and not realizing that the support log they pasted in contained non-expiring authentication tokens, or that the financial model they uploaded included customer account numbers.

Shadow AI breaches average 247 days to detect and cost $670K more than standard incidents.[8] The exposure is accidental. It is invisible without behavioral monitoring specifically designed to detect it.

THE PRINCIPLE

You cannot protect data you cannot see. Enterprises need real-time visibility and a forensic record of every AI data flow — not just policies on paper.

WHAT GOOD LOOKS LIKE

- Data loss prevention tooling with specific detection logic for AI tool interactions — generic DLP policies are not calibrated to this traffic pattern.
- Real-time blocking capability at the data layer, not just detection after the fact.
- A forensic record of AI prompt interactions for compliance, legal hold, and eDiscovery — AI conversation logs are being auto-deleted at most organizations before legal obligations can apply.
- A clear operational distinction between sanctioned and unsanctioned AI, enforced at the data layer.

[7] Harmonic Security 2025 AI Data Exposure Report.

[8] IBM 2025 Cost of a Data Breach Report.

Secure every identity, human and machine.

05.

THE THREAT

AI agents authenticate via API keys and OAuth tokens increasingly stolen through adversary-in-the-middle (AiTM) phishing. One compromised agent token grants silent, persistent, system-wide access — sometimes for months before anomalous behavior surfaces.

By 2026, autonomous agents will outnumber human employees in most enterprise environments[9], each representing a credential that can be stolen, a permission set that can be exploited, an identity that can be impersonated. The identity governance frameworks most organizations have were built for a human workforce. They do not scale to this ratio.

The attack chain is consistent: phishing email lands, credentials are harvested, the attacker pivots to service accounts and agent tokens accessible under those credentials, and the agents become the exfiltration mechanism. They move fast. They do not trigger behavioral alerts built for human actors.

THE PRINCIPLE

Non-human identities now outnumber people. The credential theft chain that ends with a compromised AI agent almost always begins with an email.

WHAT GOOD LOOKS LIKE

- Identity governance policies explicitly extended to non-human identities — AI agents, service accounts, MCP connections — with the same access review cadence applied to human accounts.
- AiTM defense and OAuth token phishing prevention at the email layer, where the credential theft chain typically originates.
- A current inventory of all active agent credentials, their scope, and the human who deployed them — with an established off-boarding procedure for agents tied to departing employees.
- Behavioral baselines per agent identity, so deviations from normal operating scope surface as signals rather than noise.

[9] Palo Alto Networks, "6 Cybersecurity Predictions for the AI Economy in 2026," Harvard Business Review, December 2025.

Archive, audit, produce AI interactions.

06.

THE THREAT

Courts, regulators, and data subjects can now demand disclosure of AI prompt chains, automated decisions, and LLM conversation logs. GDPR Data Subject Access Requests legally encompass AI interaction data. Most enterprises have no retrievable record. AI conversation logs are quietly auto-deleted before legal holds can apply.

Executive liability for rogue AI agent actions is no longer theoretical. California law that took effect January 1, 2026 explicitly removes autonomous operation as a legal defense — organizations cannot argue the AI made the decision. Colorado follows in June 2026. Gartner projects new categories of AI liability litigation will produce significant cases by mid-2026.

THE PRINCIPLE

AI interactions are business records. Every DSAR, legal hold, eDiscovery request, and regulatory investigation now extends to everything your workforce sends to an AI — and everything that AI says back.

WHAT GOOD LOOKS LIKE

- An LLM prompt archive with the same legal defensibility standards applied to email archiving: tamper-proof, searchable, subject to legal hold.
- eDiscovery workflows extended to cover AI interaction data before litigation or regulatory action creates an emergency requirement.
- Data retention policies reviewed to ensure AI conversation logs are not being auto-deleted by default settings on AI platform subscriptions.
- An AI interaction audit trail covering agent actions: what the agent did, when, under whose identity, and what data it accessed.
- Retention schedules that actively dispose of data past its lawful or operational lifespan — reducing the corpus agents can act on to what the organization needs and can defend.

THE PIVOT

04.

The human at the center.

The dominant industry frame for AI agent security is containment. The more durable insight is that agents are an extension of human risk — not a separate technology category — and the controls built for human risk are the right foundation for governing them.

"The agent did it" is not a finding. It is a deferral. Behind every agent action is a person who delegated it.

THE MENTAL MODEL

Agents extend human risk — an employee no one onboarded, carrying real credentials.

THE BEHAVIORAL PROBLEM

Legitimate AI and a hijacked credential produce the same signal — intent is not in the log.

THE NEW CONSTRAINT

Damage accumulates at API speed — the shift from alert-first to remediate-first response.

Agents are not a technology category separate from human risk. They are an extension of it.

AI agents are created by humans. They act on human behalf. They carry human credentials. They operate in the same communication channels — email, Slack, Teams — where human risk already manifests. And their failure modes — hallucination, scope creep, adversarial manipulation — are structurally parallel to the failure modes that human risk management was built to address.

Agents are employees you didn't train.

The most useful mental model for an AI agent is a new employee who was never onboarded. They have access because someone gave it to them. They have not been trained on what they are allowed to do with that access. They will make mistakes — just as humans make mistakes — and those mistakes will scale at the speed of their underlying compute.

This parallel is not metaphorical. It is architectural. The controls that govern human behavior — access rights, activity monitoring, behavioral baselines, policy enforcement, anomaly detection — translate directly to governing agent behavior. Organizations that have already built a mature human risk practice are closer to agentic readiness than they may realize. The extension is not a rebuild. It is an expansion of an existing foundation.

THE MENTAL MODEL

An AI agent is like a new employee who was never onboarded — handed access without any training on how to use it, and able to scale every mistake at machine speed. Every control built for the human translates to governing the agent.

Behavior is the signal. Speed is the new constraint.

The behavioral ambiguity problem.

One of the most significant operational challenges in agentic AI security is the inability to distinguish legitimate AI activity from malicious activity based on observable behavior alone.

An executive using an AI agent to send hundreds of Slack messages per day on their behalf is indistinguishable, from a behavioral monitoring standpoint, from an attacker who has compromised that executive's credentials and is using them for mass impersonation. Both produce the same signal: anomalous message volume from a known identity. The only thing that differentiates them is intent — and intent is not observable in the event log.

The resolution is behavioral context: baselines that capture what normal looks like for a given user, including their known AI usage patterns, so that deviations from that baseline produce meaningful signals rather than noise. This is not a new capability requirement. It is an extension of the behavioral intelligence that human risk management platforms already generate.

Remediating at machine speed.

When an AI agent goes off the rails — whether through prompt injection, scope creep, credential compromise, or misconfiguration — the damage does not accumulate at human speed. It accumulates at API call speed.

Security operations built around an alert-first, investigate-second model are not calibrated to this reality. The emerging model is **remediate-first, alert-second**: automated controls that stop the action before human review is possible, with the alert serving as a record of what was caught rather than a request for human intervention.

This is a meaningful architectural shift with implications for platform selection, workflow design, and SIEM integration.

THE EVALUATION QUESTION

Security leaders evaluating AI security investments in 2026 should assess vendor capabilities against one specific criterion: can this tool operate at the speed of the threat?

THE ROADMAP

05.

From visibility to agentic readiness.

Most organizations are not starting from zero — they are extending the email security, DLP, and identity governance they already run. This section maps the four levels of AI security maturity, and the board conversation that moves an organization up them.

Risk-informed velocity, not risk elimination — move fast on AI, but know the moment something goes wrong.

THE MATURITY MODEL

Four levels, from visibility through control and governance to agentic readiness.

HOW TO ASSESS

A single diagnostic question that locates an organization at each level.

THE BOARD CONVERSATION

Three things to report — and the framing that funds the next level.

Four levels. Most organizations sit between one and two.

Most organizations are not starting from zero. They have email security, endpoint protection, DLP, and identity governance already in place. The question is not what to build from scratch. It is how to extend, calibrate, and connect existing capabilities to address the AI-specific exposure described in this paper.

LEVEL	KEY CAPABILITIES	HOW TO ASSESS
01 / VISIBILITY	**See the AI footprint.** Shadow AI inventory · DLP monitoring of AI tool usage · user risk scoring baseline.	Can you name which AI tools are active today? Can you see what data is moving through them?
02 / CONTROL	**Enforce in real time.** Real-time AI data flow blocking · behavioral baselines per user · adaptive policy enforcement.	Can you enforce policy in real time? Do controls apply to agents as well as humans?
03 / GOVERNANCE	**Defensible by design.** Forensic record of AI interactions · LLM prompt logging for eDiscovery · AI use policy with measurable awareness · defensible disposition of data past its lawful or operational lifespan.	Can you produce AI prompt logs under legal hold? Do more than 50% of employees know your AI policy? Can you demonstrate active disposition of out-of-scope data?
04 / AGENTIC READINESS	**Built for machine speed.** Agent identity cataloging · behavioral anomaly detection at machine speed · remediate-first architecture.	Have you extended your human risk architecture to AI agents? Can your tools operate at AI speed?

Most organizations will find themselves between Level 1 and Level 2 as of 2026. The near-term priority is closing the Level 1 gap — establishing genuine visibility into AI tool usage and data flows — before investing in more advanced control layers. Governance and agentic readiness built on top of a visibility deficit will produce policy without enforcement: the most common and costly pattern in enterprise AI security today.

Risk-informed velocity, not risk elimination.

Board-level AI security conversations in 2026 have a specific character. Board members are simultaneously encouraging aggressive AI adoption for productivity and asking CISOs to account for the risks that adoption creates. The tension is real and the questions are not going away.

The most effective board reporting on AI security covers three things: what sanctioned AI tools are in use and how they are governed, what unsanctioned AI activity has been detected and what controls are in place, and what the organization's posture is on the six principles in this framework.

The right framing for the board is not risk elimination — that is neither achievable nor the correct goal. It is risk-informed velocity: the ability to move fast on AI adoption while maintaining the visibility, controls, and governance architecture to know when something is going wrong and respond at the appropriate speed.

SANCTIONED AI

What is in use, what data each tool can reach, how it is governed, and what changed since last report.

UNSANCTIONED ACTIVITY

What has been detected, what controls are in place, and where the residual exposure sits.

POSTURE ON THE PRINCIPLES

Honest current state against each of the six principles, with the next move named for each.

THE FRAMING

Risk-informed velocity is the goal — not risk elimination. The CISOs navigating this environment grasp three things at once: the human layer is still the most important variable; agents are an extension of it; and speed is now a constraint that detection alone cannot satisfy.

CONCLUSION

06.

The window is now.

The conditions that make 2026 a pivotal year for AI security investment are structural, not cyclical. The threat is real and accelerating. The readiness gap is measurable and documented.

The decisions organizations make about their AI security architecture in the next twelve to eighteen months will be significantly harder to reverse than they are to get right the first time.

Three things are true simultaneously, and the CISOs navigating this environment well, understand all three at once. **The human layer remains the most important variable** in enterprise security — and AI has made it more complex, not less. **AI agents are not a separate security domain.** They are an extension of human risk, and the frameworks built to address human risk are the right foundation for addressing agent risk. And **speed is a new constraint.** The threat moves at machine speed.

Security architectures that cannot respond at comparable speed will be structurally disadvantaged, regardless of the sophistication of their detection capabilities.

The organizations that treat AI security as a separate initiative — a new budget line, a new vendor, a new team — will build fragmented defenses. The organizations that treat it as the natural extension of the human risk work they have already been doing will build something more coherent, more defensible, and more adaptable to the threats that have not yet been named.

Security leaders who define their AI security posture before an incident forces the conversation will be better positioned in every dimension — with their boards, with their vendors, with their regulators, and with their organizations. That window is available now. It will not remain open indefinitely.

APPENDIX

Where the numbers come from.

All data points cited in this paper, with original sources.

STATISTIC	FIGURE	SOURCE
Organizations expecting AI-driven attacks in 12 months	69%	Mimecast State of Human Risk 2026
Organizations with a specific AI security strategy	40%	Mimecast State of Human Risk 2026
Organizations with unsanctioned AI tools	98%	Mimecast Research
Employees using personal GenAI accounts at work	47%	Netskope Cloud and Threat Report 2026
Employees aware of their org's AI use policy	18.5%	Mimecast State of Human Risk 2026
AI interactions involving sensitive data	39.7%	Mimecast Research
AI prompts disclosing private information	4%	Mimecast Research
Files uploaded to AI tools with confidential content	20%	Mimecast Research
Days to detect a shadow AI breach (avg)	247	IBM 2025 Cost of a Data Breach Report
Extra cost of shadow AI breach vs. standard	$670K	IBM 2025 Cost of a Data Breach Report
Ransomware breaches involving SMBs	88%	Verizon DBIR 2025
Typical SMB breach cost range	$120K–$1.2M	Verizon DBIR 2025
Share of cyberattacks that are phishing	77%	Mimecast GTI Report 2025
Surge in AI-powered scams in 2025	1,210%	Pindrop Security AI Fraud Trends 2026
BEC attacks now using AI deepfakes	40%	Mimecast State of Human Risk 2026
Companies with an AI use policy	44%	Palo Alto Networks
Enterprise apps with embedded AI agents by 2026	40% (proj.)	Gartner

APPENDIX

References, trademarks, and the closing statement.

REFERENCES

1. Mimecast 2025 Global Threat Intelligence Report.
2. Verizon 2025 Data Breach Investigations Report (DBIR).
3. NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.
4. Varonis 2025 State of Data Security Report.
5. Netskope Cloud and Threat Report 2026.
6. Mimecast State of Human Risk 2026 Report.
7. Harmonic Security 2025 AI Data Exposure Report.
8. IBM 2025 Cost of a Data Breach Report.
9. Palo Alto Networks, “6 Cybersecurity Predictions for the AI Economy in 2026,” Harvard Business Review, December 2025.

TRADEMARKS

FURTHER READING

Mimecast research publications, including the underlying State of Human Risk 2026 report and Global Threat Intelligence, are available at mimecast.com/resources.

About this publication.

This white paper is provided for informational purposes only and does not constitute legal, regulatory, financial, or professional advice. While Mimecast has made reasonable efforts to ensure the accuracy of the information contained in this document, no representation or warranty, express or implied, is made as to its completeness or fitness for a particular purpose.

This document reflects information available as of the date of publication and is subject to change without notice. Readers are encouraged to seek independent professional advice before taking any action based on the contents of this paper.

CONTACT

For research inquiries, briefings, or to request a custom assessment against the maturity model in this paper: mimecast.com/company/contact/

FURTHER READING

Mimecast research publications, including the underlying State of Human Risk 2026 Report and Global Threat Intelligence, are available at mimecast.com/resources.

mimecast