trevonix

Securing AI Agents:

A Framework for Identity, Authorisation and Governance

Executive Summary

AI agents are moving rapidly from experimentation into day-to-day use. They now write code, retrieve and process data, support customers, automate workflows, and call APIs on behalf of users and systems.

As their influence grows, so does their potential to cause harm if not properly governed.

Two of the most difficult challenges in securing AI agents are:

Knowing they exist - AI agents appear in many places and rarely follow traditional onboarding routes.

Managing what they are allowed to do - authorisation models designed for humans don't yet fit dynamic, autonomous agents.

Around these issues, a new security framework is emerging. It is built on five pillars:

Discovery - identifying where agents operate and what they connect to.

Identity - giving agents unique, accountable identities with clear ownership.

Authorisation - granting the correct level of access based on context and intent.

Observability - understanding what agents did, and why.

Governance - ensuring controls work at runtime and behaviour matches policy.

This white paper outlines a practical approach for organisations wishing to adopt AI agents safely and at scale.

trevonix

Why AI Agents Change the Security Landscape

1.1 From Predictable Systems to Adaptive Actors

Conventional applications behave in largely predictable ways:

They follow fixed logic.

They rely on known integrations.

Their behaviour is controlled by developers.

AI agents behave differently. They:

Make **real-time decisions.**

Adjust their behaviour based on prompts, data and context.

Cross multiple systems in a single task.

May act on behalf of a user, or independently.

This shifts the core security challenge from asking:

"What can this application do?"
to:
"What might this agent decide to do, for whom, and under what circumstances?"

1.2 The Two Foundational Challenges

Organisations repeatedly encounter the same obstacles:

Discovery

Most AI agents never pass through IT or HR. They simply appear:

- in command-line tools
- as Slack or Teams bots
- as custom GPTs
- embedded within SaaS products
- built quietly by internal teams

Authorisation

Even when an agent is known, it often receives overly broad permissions because existing access controls are too coarse for autonomous behaviour.

Without a strategy for these two areas, everything else becomes reactive.

Discovery: Knowing Your Agents Exist

2.1 Why Discovery Is Difficult

Human identities are straightforward. HR creates a record, identity systems ingest it, and accounts are provisioned.

Agents have no such lifecycle.

They are created wherever someone can plug a model into a workflow. A developer, data scientist, product manager or support analyst can create an agent without informing anyone.

This leads to blind spots, and blind spots lead to risk.

2.2 Discovery as an Embedded Capability

Agent discovery is emerging as a **feature** across multiple markets rather than a standalone product:

Key capabilities include:

- Identity governance solutions scanning for non-human identities.
- Endpoint and EDR products detecting agent processes.
- Network and cloud security tools identify agent-driven traffic.
- AI security platforms scanning codebases, plugins and configurations.

The goal is simple:

Maintain a clear, continually updated inventory of every AI agent, its integrations, and the systems it can reach.

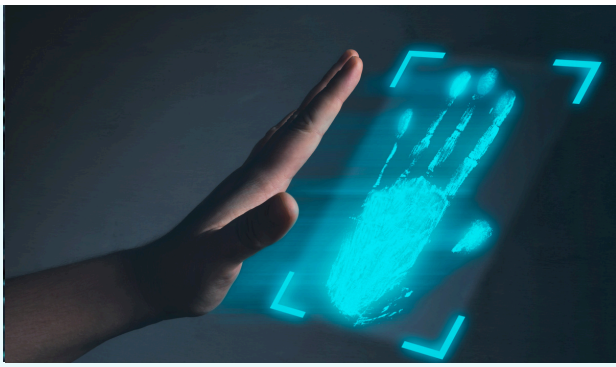
Identity: Who Is the Agent and Who Do They Represent?

3.1 From Anonymous Bots to Accountable Entities

Once discovered, agents require **unique and accountable identities.** Organisations must be able to answer:

Benefits include:

- Which agent is this?
- Who owns it?
- Which user (if any) is it acting on behalf of?
- How is it created, maintained and retired?



Shared service accounts or generic API keys remove accountability entirely. If an agent behaves incorrectly, attribution becomes almost impossible.

3.2 Where Agent Identity Is Emerging

Agent identity spans several existing and emerging markets:

Identity Governance (IGA) - extending lifecycle processes to non-human entities.

Non-Human Identity (NHI) solutions – specialised tools for machine and agent identities.

Workforce and customer identity platforms – treating agents as first-class identities.

Agent-first identity brokers – new platforms designed for agent-to-system access.

The direction of travel is clear:

Agents must be treated as proper identities with traceability, ownership and lifecycle controls.

trevonix

Authorisation: What Is the Agent Allowed to Do?

4.1 The Limitations of Human-Centric Access Controls

Legacy authorisation models: OAuth scopes, role-based access, and static API tokens were designed for humans.

Humans:

- Act slowly.
- Perform limited tasks.
- Use judgment.
- Trigger alarms when things feel wrong.

Agents can:

- Issue thousands of actions in seconds.
- Traverse multiple systems in a single execution.
- Change their behaviour based on prompts or retrieved context.

Current tooling cannot provide the precision needed for safe AI automation.

4.2 Dynamic, Context-Aware Authorisation

Agents require dynamic, task-aware and context-sensitive authorisation.

This includes:

Just-in-time permissions – only granting access when needed.

Continuous evaluation – revalidating access as the context changes.

Just-enough access – restricting permissions tightly to the task at hand.

Immediate intervention – allowing security tools to halt agents the moment they behave suspiciously.

This area requires the most reinvention. Progress is emerging from:

- Fine-grained authorisation platforms.
- Data access governance solutions.
- AI-specific control planes.
- Evolving capabilities in workforce identity tools.

The challenge is balancing productivity with safety.

trevonix

Observability: Understanding What Agents Actually Did

5.1 Capturing Decisions, Not Just Events

Identity and authorisation define boundaries.
Observability provides **evidence.**

Traditional monitoring focuses on:

- System logs
- Errors
- API calls
- Infrastructure behaviour

AI agents require deeper insight. They:

- Make decisions, not just calls.
- Sometimes represent a user, sometimes not.
- Act across many systems in one request.
- Change behaviour with different stimuli.

Observability must therefore capture:

- The decision path.
- The context behind decisions.
- The user-to-agent linkage.
- A trace from prompt to action to outcome.

trevonix

5.2 Market Movement

We are seeing:

Existing observability platforms are extending tracing to include agent behaviour.

AI-first observability tools are emerging.

Combined platforms offering monitoring, behavioural detection and real-time blocking of risky activity.

The end goal is:

A complete, auditable record of every agent action, linked to an accountable human.

Governance: Proving Controls Work in Practice

6.1 Beyond Model Risk into Runtime Behaviour

Much AI governance to date has focused on models:

- Bias
- Training data lineage
- Model risk frameworks

For agents, this is insufficient.

Agent governance must cover runtime behaviour and provide evidence that:

- Identity controls were applied.
- Authorisation checks executed as intended.
- Policies were followed at each step.
- Exceptions were logged, reviewed and approved.

6.2 Runtime Governance Requirements

Effective agent governance needs:


Logs showing policy enforcement at the moment of action.


Regular reviews of agent activity and access patterns.


Audit trails proving every sensitive action passed an authorisation check.


Clear mechanisms for approving and documenting exceptions.

New AI security platforms are appearing to deliver this: blending monitoring, behavioural detection, policy enforcement and automated revocation.

A Practical Adoption Framework for Organisations

Organisations can adopt AI agents safely by following a five-step approach:


Establish Agent Discovery - Build an accurate inventory of agents, their permissions and their integrations.


Define Agent Identity Standards - Assign unique identities, link them to accountable owners, and manage them through existing lifecycle processes.


EModernise Authorisation - Shift towards dynamic, context-aware, just-in-time access models.


Implement Agent Observability - Capture decisions, user linkage, and end-to-end action traces.


Strengthen Governance - Document expectations, gather runtime evidence and create feedback loops to refine controls.

Conclusion

AI agents introduce a fundamentally different security challenge. They are autonomous, fast-moving and capable of complex behaviour across multiple systems.

To secure them, organisations must:

- Discover where agents exist.
- Give them proper identities.
- Control their access dynamically.
- Observe their decisions.
- Govern their behaviour with evidence, not assumptions.

A new market is forming around these needs, bridging identity, authorisation, observability and AI security.

The organisations that succeed with AI won't merely focus on models; they will build robust identity and governance foundations for the agents acting on their behalf.