Hypatos

Securing Agentic AI at Scale: A Practical Guide for GBS and IT Leaders

Ensuring Agentic AI is Secure and Compliant Event When Scaling

hypatos.ai

Executive Summary:

Agentic AI is moving rapidly from pilot to widespread implementation, creating new security risks that traditional AI and automation controls were not designed to handle. Unlike previous iterations of automation, Agentic systems can autonomously execute actions, making compromised agents equivalent to high-privilege “digital-insiders.” As adoption accelerates in 2026, organizations must shift security thinking from data-exposure to the risk of unauthorized actions.

This report details how a zero-trust approach (which includes giving Agents unique-identifiers, least-privileged access and full traceability) is essential. Security, compliance, and data governance must be embedded by design, rather than retrofitted to prevent operational, financial, and reputational damage. Whilst the aim of implementation is straight-through-processing, human leadership of agents remains critical.

Secure Agentic AI requires the right balance of autonomy, oversight and explainability to maintain trust and reliability. Organizations that act now and take concrete steps to Agentic AI security will be able to unlock the productivity and scalability benefits of this revolutionary technology, without compromising security, compliance and control.

Who is this guide for?

GBS Leaders & Professionals:

- Understand and mitigate the security and data protection risks of using Agentic AI.
- Learn the right balance of Human-in-the-Loop for effective governance, whilst still enabling increased productivity.
- Uncover how Agentic AI is a “glass-box” and not a “black-box,” and can create an immutable audit-ready trail.

IT and Security Professionals:

- Understand how Agentic AI security differs from traditional security.
- Establish best practice techniques such as Identify Access Management, defense in-depth and zero-trust principles.
- Establish what to look for when assessing your Agentic AI solution complies with security regulations.

Contents

Introduction: The Urgency of Security in Agentic AI

According to data collected by SSON, a staggering **81% of Global Business Services are considering investing in Agentic AI in the next 12 months.**

Expectations around Agentic are huge, and rightly so. It has the power to create a step-change in efficiency, better scalability, unleash critical data analytics, and elevate service experience; all without adding headcount. It's unsurprising that with this level of opportunity, adoption will be rapid.

Whilst 2025 was the year of pilots, **2026 represents widespread scaling of Agentic AI**, which is being let loose on mission-critical processes. Without security at the heart of implementation, vulnerabilities will be quickly exposed, and the consequences will be dramatic.

Whilst most projects are being driven by the efficiency opportunities Agentic AI offers, there are definite security benefits. Larger, especially multinational, organizations are a patchwork of different solutions, unconnected processes, and multiple ERPs.

Every time a workaround is required, or data must be moved from one point solution to another, there is an increased risk of compromise. Added to this, the substantial risks created by out-of-date solutions mean that security is just as powerful an incentive for implementing Agentic AI as productivity is.

"In just a couple of years, this will not be experimentation, this will be GBS."

Uli Erxleben

CEO and Co-Founder of Hypatos, on Agentic GBS

The New Security Paradigm

How does Agentic AI Security Differ from Traditional Security?

Traditional AI generates text. Agentic AI performs tasks. That means an attacker who compromises an agent doesn't just gain access to information; they gain access to capabilities.

This means an exploited agent could:

- Submit forms or execute transactions
- Access internal systems
- Trigger workflows
- Move or modify data
- Interact with third-party services.

> In an Oct 2025 article, McKinsey wrote, ***"Already, 80 percent of organizations say they have encountered risky behaviours from AI agents, including improper data exposure and access to systems without authorization,"*** so the threat is very real.

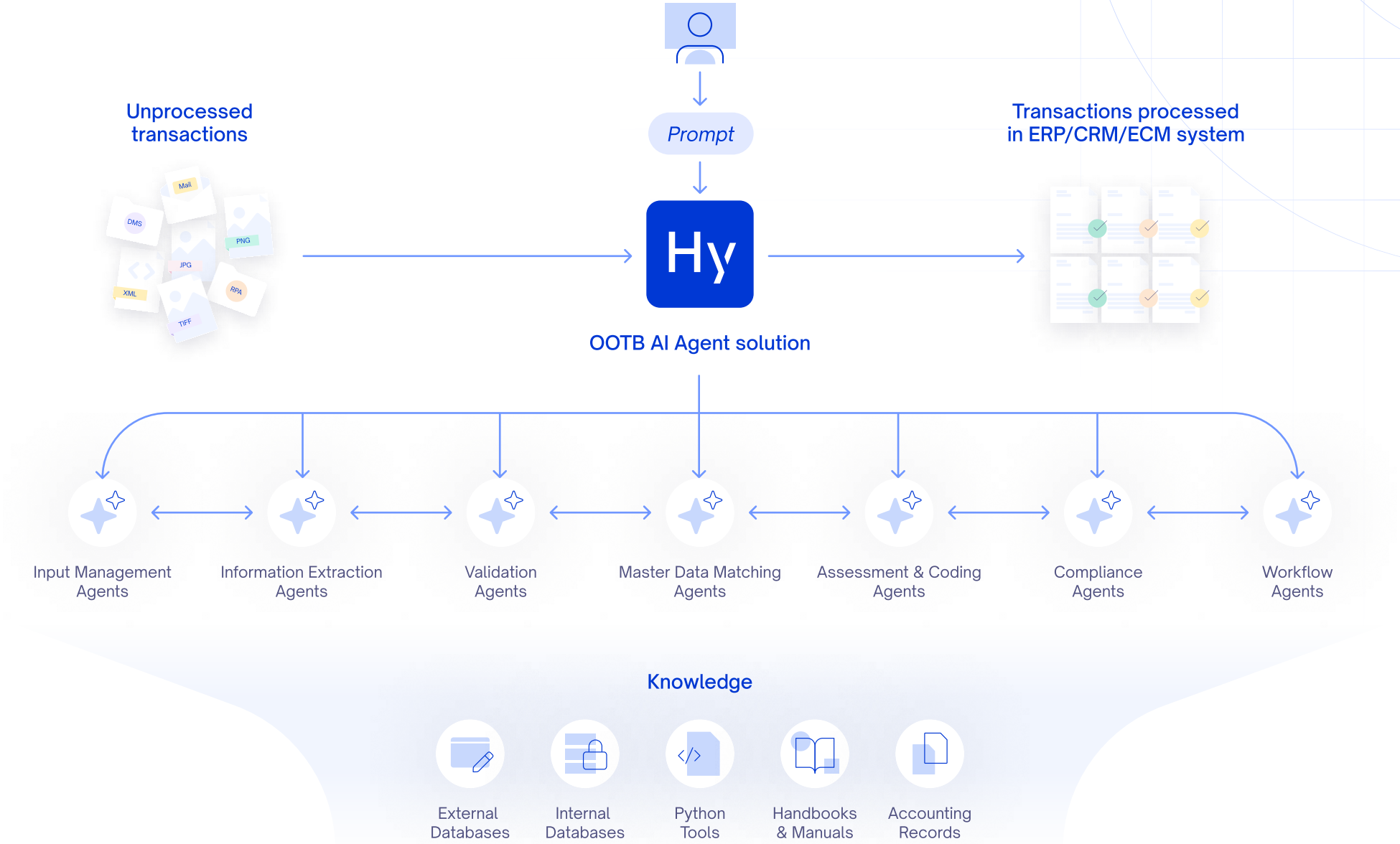
In security terms, an autonomous AI agent is like an insider with privileged credentials to your cloud or on-premise architecture, a "digital insider" if you will. This means an expanded attack surface where every AI agent is a potential entry point for attackers.

Unlike automation solutions that came before (which follow pre-defined rules), Agentic AI can behave unpredictably. This might mean hallucinating actions or misinterpreting a prompt. Luckily, with the right controls and a robust security framework, these risks can be mitigated.

As Hypatos, **Head of Information Security, Vasil Sultanov** explains it:

> *"We need to stop thinking about AI as just a tool that reads and start treating it as an entity that acts. In traditional software security, we worry about data leaking out. With Agentic AI, we have to worry about unauthorized actions happening within.*
>
> *We treat AI agents effectively as "digital employees." Just like you wouldn't give a new intern the master keys to your bank account on day one, you shouldn't give an AI agent unrestricted permission to execute transactions."*

Figure 1: Example of a Hypatos AI Agent Framework.



Solving Key Security Risks for Agentic AI

Identity and Access Management for AI Agents

A cornerstone of security in any environment is Identity and Access Management (IAM), ensuring that each actor in a system has a unique identity and only the minimum access necessary (least privilege). When it comes to Agentic AI, this policy must extend to AI Agents, and zero-trust principles must be applied.

Unique Credentials:

Just like with human employees, **avoid any shared or hard-coded credentials for AI systems** and instead issue unique, trackable identities for each agent. Experts recommend approaches such as short-lived certificates or tokens, so the agents authenticate just like human users (avoiding long-term passwords).

This means that not only is it easy to trace the rogue agent if something goes wrong, but you can revoke permissions for specific agents without damaging the whole workflow.

Least Privilege Access:

Experts recommend **roles-based or policy-based access based on Agent skills and tasks.** This should be defined narrowly; for example, an agent tasked with reading customer invoices shouldn't also be allowed to delete or modify them, unless absolutely required.

Combine this with just-in-time permissions to ensure even stronger security. In this case, Agents should request a temporary token for a specific task, which expires immediately upon completion.

The IBM 2025 Cost of Data Breach Report revealed that **97% of organizations that reported an AI-related security incident lacked proper AI access controls.** Underscoring the need to get IAM right.

Prompt Injection & Data Poisoning

What is prompt injection?

Prompt injection is a type of attack in which an attacker manipulates an AI system by inserting malicious or misleading instructions. This can happen via emails, documents, tickets, or corrupted historical data. Many organizations mitigate that by simply sanitizing and validating inputs and stripping hidden text before data is entered into the system.

However, **Hypatos Head of Information Security, Vasil Sultanov,** recommends the more secure method of structural segregation.

> *"Do not rely solely on stripping text. Instead, strictly separate the "System Prompt" (instructions) from "User Data" (content). Furthermore, enforce rigid output schemas to significantly reduce the risk of the model outputting malicious code."*

In addition, rely only on trusted sources for input, such as your ERP, and secure data provenance and lineage by implementing tracking to know exactly where and when data was added or modified.

Data Access, Storage & Privacy

How secure is my data when using Agentic AI, and will it be used to train models?

Data collected at a Hypatos webinar in Nov 2025 revealed that the biggest perceived barrier to implementing Agentic AI in Global Business Services is **"Data quality, access, or governance constraints."** 58% of respondents cited this as a problem over and above integration challenges, technical skills, and talent shortages.

Chart 1: *What do you see as the biggest barriers to implementing agentic AI in Global Business Services?*

At a basic level, you must ensure your Agentic AI solution provider has all the necessary security certificates. Hypatos have their own dedicated "Trust Center" and maintain rigorous compliance posture, including SOC 2 Type 2, ISO 27001, ISO 27017, ISO 27018, C5, and HIPAA certifications.

This ensures that the agentic framework is built upon a verified, secure foundation rather than just theoretical guidelines. It's also important to ensure your Agentic AI provider isn't keeping your data unnecessarily long or using non-reputable LLM providers.

Another frequent question asked by enterprise leaders is *"Will my most sensitive documents be used to train AI models?"* **Sultanov addresses this:**

> *"The answer lies in architecture, not just policy. You must demand "Zero Data Retention" for inference. At Hypatos, we clearly distinguish between "processing" and "learning."*
>
> *When our agents process a document using third-party foundational models, that data is sent via stateless APIs, meaning the model provider processes the request and immediately discards the data. It is never stored or used to train their base models. We combine this with strict tenant isolation and encryption, ensuring that one customer's data never influences another's results."*

From a data security and privacy point of view, it's important to ensure that your Agentic AI complies with regulations such as GDPR. As an AI that acts autonomously, makes decisions, and initiates actions, Agentic AI naturally raises questions around accountability and control (fundamentals on which GDPR was built).

Under GDPR, a "data controller" decides why and how personal data is processed, but legally, AI can't be the controller. Therefore, responsibility always needs to be traced back to a human.

Traceability and explainability are also incredibly important to GDPR compliance, as you are required to explain what data is processed and why. This means having a solution that can create an immutable path and explain its workings in plain-English is paramount.

Defence-in-Depth: Ongoing Security Maintenance and Incident Response

In addition to the procedural controls necessary for regulatory compliance, technical safeguards are also important. **Sultanov recommends the following:**

> *"**Inventory Your Agents:** You can't secure what you don't know exists. Treat AI services like any other asset and maintain a strict inventory to prevent "Shadow AI" from growing unchecked in your environment." Also important are enforcing external guardrails, not just prompts; "Do not rely on the LLM to police itself. Build deterministic, code-based checks outside the model to validate outputs.*
>
> *For example, if an agent drafts external client emails, use a regex scanner to block the message if it detects sensitive PII before the email is sent."*

In addition, **Sultanov recommends** organizations implement '**Circuit Breakers**':

> *"Security is not just prevention; it is also containment. If an agent's transaction velocity spikes anomalously or error rates exceed a defined threshold, the system must automatically suspend the agent's permissions to prevent runaway errors or mass fraudulent actions."*

Furthermore, incorporate AI systems into your normal security testing regime. This might involve penetration testing the overall workflow (can a tester trick the AI into doing something harmful?), testing for prompt injections (malicious inputs that cause the AI to deviate from intended behavior), and checking that all oversight triggers and fail-safes work properly. **Hypatos performs third-party penetration testing annually.**

Lastly, to touch briefly on incident response and BCP: **Ensure both you and your Agentic AI provider have robust incident response procedures.** Hypatos has a documented security incident response plan that outlines the procedures for identifying, containing, eradicating, and recovering from security incidents.

The Importance of Human-on-the-Loop (But not Off!)

What level of human oversight is needed when using Agentic AI?

At the World Economic Forum 2026 in Davos, Accenture CEO Julie Sweet described a future where AI is used with “human in the lead”, rather than in-the-loop. This is an incredibly important distinction, and most AI leaders agree that critical business decisions should remain in the hands of humans.

As Hypatos **CEO and Co-Founder, Uli Erxleben** states:

> *“These are not productivity tools. These are co-workers that need to be managed by human experts.” When implementing the solution “You must curate the knowledge the AI uses; legal texts, master data, accounting rules.”*

However, having too many human approvals undermines the core purpose of Agentic AI, which is to enable fast, autonomous processes! So completely removing humans from the workflow isn't recommended. **The goal is to find the right balance.**

Define explicit checkpoints:

When does the Agentic AI need to stop and check? A good example in GBS is that an AI Agent can process an invoice, but **final approval to release payment should sit with a human.**

The AI Agent needs to provide clear contextual prompts that a non-expert understands, so when it comes to approvals, an employee can understand the decision they are making. Hypatos speaks in plain English, rather than script, to ensure the human can make a clear, informed choice and prevent rubber-stamping.

Ongoing security training:

Like with traditional technology, the biggest attack-risk comes from humans. In Agentic AI’s case, this might be due to over-trust in AI recommendations, rubber-stamping, and responsibility diffusion (i.e., the AI, did it?) To mitigate this, you can rotate AI approvals for high-risk categories between team members and provide ongoing training. As **Erxleben cautions:** *“The biggest mistake companies make is not investing enough in enabling people.”*

As your organization's use of Agentic AI evolves, so too can the role of HITL. The figure below represents how human and AI Agent work together in the “setup phase” where humans act as teachers and validators, and later during the “steady phase,” where human input shifts to supervisor and governor.

GBS Analyst

Always in control, role shifts over time

HITL during Setup Phase / *Teacher & Validator*

AI Agent *(Learning mode)*

Human defines rules

- Set business logic
- Defines thresholds
- Configure workflows
- Map exceptions
- Train with examples
- Validate every output

Human reviews 100%

- Check every decision
- Correct mistakes
- Provides feedback
- Refine prompts
- Build confidence
- Sign off on go-live

HOTL during Steady Phase / *Supervisor & Governor*

AI Agent *(Production 95% auto)*

Human monitors

- Real-time dashboards
- Track accuracy metrics
- Review audit logs
- Spot anomalies
- Compliance checks
- Performance trending

Human intervenes

- Handle exceptions (5%)
- Resolve edge cases
- Override when needed
- Update rules
- Investigate issues
- Continuous improvement

Maintaining Explainability, Auditability and Security Regulations

How explainable are your agents' decisions, and will auditors trust them?

We've already touched on this briefly, but it bears repeating. There is a misconception that Agentic AI acts as a black box. However, if you get the right solution, this is not the case.

As **Sultanov explains:**

> *"While the neural networks inside an LLM are complex, the actions of an agent are fully observable. A secure agentic platform logs every thought (chain-of-thought), every tool it requests, and every output it generates.*
>
> *With the right governance, an AI Agent actually provides more consistent auditability than a human employee, because it creates an immutable, timestamped record of exactly why it made a decision and what data it used."*

It is important to keep these logs secure as they will contain sensitive information about your operations and data, so protect and access control the logging systems. Also, decide on an appropriate level of retention for these logs, i.e. long enough to satisfy the auditors but not so long that data management becomes a risk of its own.

When implementing Agentic AI it's important to review your security policies (acceptable use, data handling, incident response, etc.) and make sure they cover scenarios involving AI. Traditional frameworks like ISO 27001 or NIST CSF focus on systems, processes, and people – they do not yet fully account for autonomous agents acting with discretion.

Lastly, **do not forget your vendors.** Most multinational organizations operate as extended networks of outsourcers, suppliers, consultants, and software providers.

A single weak control at a partner can create operational disruption, reputational risk, and financial loss. This means that all vendor-provided agents must be reviewed by security, incident notification of SLAs contractually defined, and vendor update processes understood and governed.

Conclusion

Agentic AI is fundamentally changing the security model, moving it from the risk of data exposure to capability-level risks where compromised agents can take real actions with significant business consequences. Therefore, security frameworks must be designed from day-one, not bolted on at scale. This means identity access control, and zero-trust principles applied to AI Agents just as rigorously as if they were human users.

Data governance, privacy and regulatory compliance depend as much on architecture as the solution itself so it's important to implement minimal data retention, strict tenant isolation (separating each customer's data, agents and workloads) and auditable decision trails.

Despite talks of "autonomous," the future of secure Agentic AI is human-led, not absent, and it is important that security leaders work closely with business partners to establish the right balance of autonomy and oversight. This will help ensure your use of Agentic AI enables scale without sacrificing quality and trust.

About Hypatos

Hypatos enables The Agentic GBS with AI agents that execute end-to-end processes across P2P, O2C, H2R, and R2R—making decisions within human-set boundaries and scaling operations without headcount growth.

We put process ownership in the hands of business leaders, elevating organizations from support provider to business partner. Results: 90%+ straight-through processing, 6-9 month ROI, and 20x productivity gains.

Trusted by enterprises in Europe and North America, with hubs in Berlin, Miami, and New York, and teams across 15+ countries.

Author

Sally Fletcher

Head of Thought Leadership & Community
at Hypatos

Contributor

Vasil Sultanov

Head of Information Security
at Hypatos

Contact us

+49 (0) 302 09 97

info@hypatos.ai

Corporate Headquarters:

Hypatos GmbH c/o Mindspace
Zimmerstraße 78, 10117
Berlin, Germany