Marionette and ghost? When AI agents join you in the workplace.

Cory Kramer
Sr. Director and Global Head
Microsoft COE for Digital Workplace Services
NTT DATA

Sandeep Kaushik
Sr. Director and Strategic Advisor
Microsoft COE for Digital Workplace Services
NTT DATA

Sujay Bhattacharya
Sr. Managing Director and Global Head
Digital Workplace Services
NTT DATA

Executive summary

We are on the cusp of a workforce revolution. Autonomous AI agents — what we call "digital workers" — are reshaping the modern workplace. These agents already process invoices, manage cloud infrastructure and write code. Soon, they will execute complex, multistep business processes autonomously. This new productivity engine introduces a profound security crisis that strikes at the heart of our most established security model: role-based access control (RBAC).

The problem is one of identity. RBAC was built for humans, who have clear roles, predictable patterns and persistent identities. AI agents operate differently. They work at machine speed, executing thousands of tasks per second. While their runtime instances may be short-lived, their identity objects must persist for lifecycle management, policy enforcement and audit — a basic mismatch with traditional security models.

This paper explores a fundamental question: Is an AI agent a tool or a worker? Is it a **"marionette,"** acting as an extension of its human creator, inheriting all their permissions? Or is it a **"ghost,"** a distinct entity with its own tenant-governed identity and specific roles? The path we choose will dictate the future of workplace security — its architecture, auditing practices and risk management frameworks.

The scale of the challenge

The numbers reveal the magnitude of this identity crisis. The World Economic Forum reported that nonhuman identities — including AI agents, service accounts, application programming interfaces (APIs) keys and bots — would have exceeded **45 billion by the end of last year.**[1] That's more than 12 times the global human workforce. In most organizations, nonhuman identities already outnumber human identities by **40 to 1**, and in cloud-native enterprises, that ratio can reach **5,000 to 1**.[2]

Gartner predicts that by 2028, up to 15% of daily work decisions will be made autonomously by AI agents — approving financial transactions, granting system access, modifying production code and responding to security incidents.[3]

The question is no longer if your organization will deploy agentic AI at scale — it's how quickly you can secure agentic AI before it becomes your largest attack surface.

The identity dilemma: Marionette or ghost?

Traditional security answers a simple question: who is doing what? RBAC assigns permissions (what) to a role (who). But who, exactly, is an AI agent?

The marionette (delegated identity)

In this model, the agent acts purely on behalf of its human creator with no independent identity. When "Agent-Fin-01" processes an invoice, the audit log shows: John Doe (acting via Agent-Fin-01) approved payment.

The appeal: Simple and familiar. Accountability rolls up to John, requiring no changes to existing identity and access management (IAM) systems.

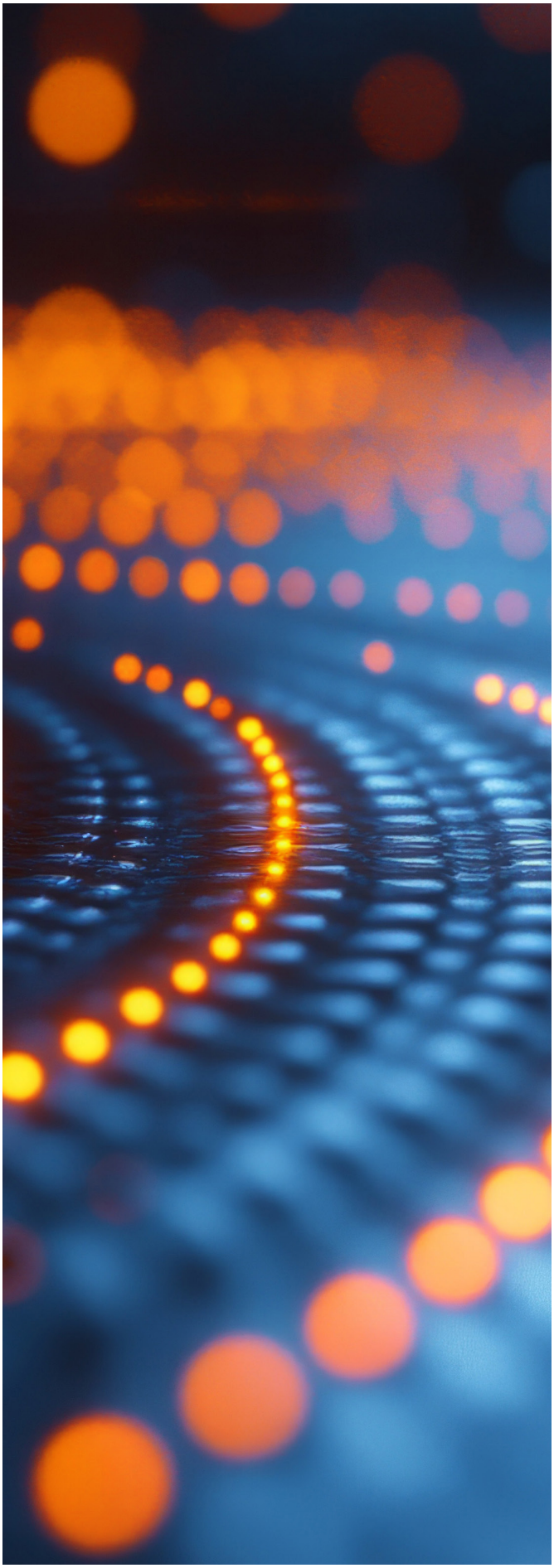
The "sorcerer's apprentice" flaw: The agent inherits all of John's permissions. If John has access to HR portals, financial summaries and administrative settings, his invoice-processing agent can potentially access everything. This creates massively over-privileged automation operating at machine speed.

The lifecycle problem: What happens when John goes on vacation or leaves the company? Does the agent's access suspend automatically, or does it continue running with orphaned credentials?

The ghost (tenant-governed identity)

In this model, the agent is a "digital worker" with its own unique identity (agent-fin-01-uuid-xyz) in the organization's tenant directory. It has an owner or supervisor rather than a creator and receives specific role assignments through RBAC. All access is centrally governed, auditable and revocable.

The appeal: This enables true zero trust and the principle of least privilege. The agent's permissions can be hyperspecific: CAN READ from 'invoices-pending' bucket and CAN WRITE to 'invoices-approved' database. Nothing more.



The identity sprawl challenge: This creates an explosion of identities to manage. With agents potentially outnumbering humans **5,000 to 1**, organizations face massive new operational overhead: provisioning, deprovisioning, access reviews and continuous auditing.

The accountability question: Who is legally and ethically responsible when an agent makes a catastrophic error? Traditional employment frameworks don't apply to autonomous software entities.

The emerging framework: Tenant-governed identity in practice.

The "ghost" model is already being implemented at enterprise scale. Microsoft Entra Agent ID, discussed at Build 2025, automatically issues each agent its own identity object in the organization's tenant directory, enabling conditional access policies, least-privilege role assignment and full audit logging.[4]

This approach provides distinct agent identities while maintaining centralized control. Organizations manage agent identities through the same IAM infrastructure used for human employees — creating, monitoring and revoking access as needed. This ensures regulatory compliance, comprehensive audit trails and clear accountability while delivering fine-grained, purpose-bound access control.

While an agent's runtime instance may be short-lived (spinning up to process invoices, then terminating), its identity object persists in the tenant directory. This persistent identity enables lifecycle management, policy enforcement, audit trails and ownership accountability. Current enterprise IAM systems require this persistence — the identity object remains in the directory even when the agent isn't actively running.

Building agentic security from existing enterprise controls

Organizations can secure agentic AI today by orchestrating existing enterprise security tools. The Microsoft ecosystem demonstrates this practical approach:

- **Microsoft Entra Agent ID** provides the foundational identity layer, giving each agent its own identity object with RBAC assignments, conditional access policies and comprehensive auditability.
- **Microsoft Agent 365** addresses identity sprawl through centralized inventory, automated lifecycle management, real-time observability and sprawl control.

- **Microsoft Purview** extends data governance to agents through sensitivity labels, data loss prevention policies and retention controls that govern what agents can access and exfiltrate.
- **Microsoft Defender** provides AI-specific security posture management, detecting configuration weaknesses, prompt injection attempts and anomalous agent behavior.
- **Microsoft Sentinel** correlates agent and human activity in a unified security information and event management platform, enabling human-supervised, agent-augmented security operations centers (SOCs) that can operate at the scale and speed that agent workforces demand.

The rise of the agent swarm

The true power of digital workers emerges when specialized agents collaborate. Consider a "product launch" swarm: Market agent scans social media for trends, passes findings to copy agent to draft content, which sends the draft to legal agent for compliance review, which then flags ops agent to publish.

This introduces new security challenges: How does market agent prove to copy agent it's authorized to request work? Does the swarm need a collective identity with temporary roles? If legal agent makes a mistake, was it the agent's fault or was it misled by upstream data?

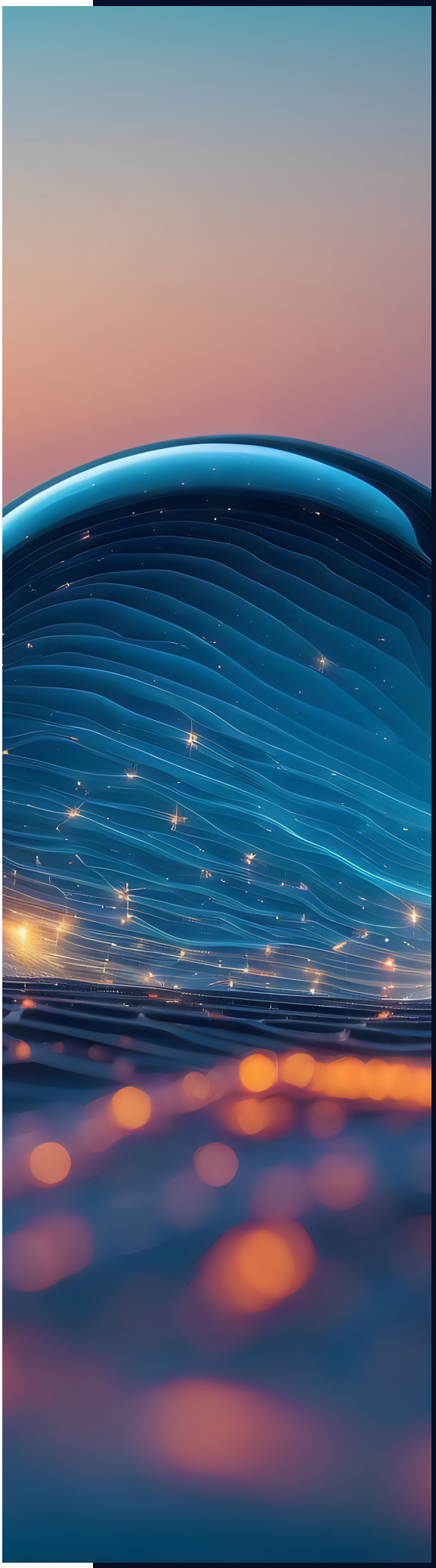
The future workforce includes **human orchestrators** managing entire "factories" of digital workers. A single "conductor" might supervise 100 agents, becoming both the most powerful and most vulnerable person in the organization — their credentials a single point of failure for the entire digital workforce.

Multiagent orchestration and communication

Organizations deploy agent swarms using three patterns: centralized control through a supervisor agent, decentralized peer-to-peer coordination and hybrid approaches — each balancing control, scalability and complexity differently.

Collaborations from **NTT DATA and Microsoft** reveals a critical trade-off: security controls that prevent malicious instructions from spreading also reduce collaborative efficiency. Strict sandboxing degrades the system's ability to accomplish complex tasks — the "security tax" that forces organizations to balance speed and safety.

For agent-to-agent authentication, emerging architectural patterns are being explored: mutual Transport Layer Security, JSON web tokens for capability attestation and event-driven architectures for coordination. However, these remain research prototypes. Current enterprise implementations rely on established protocols like API keys, OAuth 2.0 and service principals.



The audit trail: Who did it, and why?

Traditional audit logs record what happened: [10:30:41] user:'jdoe' action:'UPDATE' table:'finance' id:'inv-123'

For agents, we need an immutable log of intent and reasoning:

```
plaintext
[10:30:41] agent:'fin-bot-042' (supervisor:'jdoe' goal:'Process Q3 invoices')
[10:30:42] reasoning:'Invoice 123 from Acme is 30 days past due'
[10:30:43] reasoning:'Cross-referenced PO 456. Match confirmed'
[10:30:44] reasoning:'Confidence: 99.8%. No human review required'
[10:30:45] action:'UPDATE' table:'finance' id:'inv-123' status:'APPROVED'
```

Without this explainability, we cannot debug agents that go rogue or make compounding errors.

From logs to observability

True explainability requires comprehensive observability. This means instrumenting agents with standardized telemetry using frameworks like OpenTelemetry, where each action becomes a traceable "span" with metadata about inputs, outputs and decision factors.[5] Organizations must also maintain "model cards" documenting training data, limitations and performance characteristics.[6]

Explainability is now a regulatory requirement. The EU Artificial Intelligence Act classifies many agentic AI applications as "high-risk," mandating transparency and oversight. GDPR's "right to explanation" extends to automated decisions.[7] The NIST AI Risk Management Framework provides structured governance through four functions: govern, map, measure and manage.[8]

When agents are hijacked

Agents represent a new attack vector. Three primary hijacking scenarios illustrate the risks:

The marionette hijack: Compromising John Doe's credentials gives attackers not just his access, but a high-speed autonomous agent to wield it. They can exfiltrate years of data in seconds, with logs showing only that "John Doe did it."

The ghost hijack: Compromising an agent's tenant-governed identity (its API key or token) grants attackers a "loyal employee" inside the network. If the agent was built with least privilege, damage is contained. But centrally governed identities enable immediate response: security teams can revoke credentials, apply conditional access policies and audit all actions through standard IAM infrastructure.

The "social engineering" hijack: Attackers can "trick" agents through **prompt injection**. By feeding a poisoned invoice or malicious email to an agent, they can convince it to break its own rules, execute unauthorized commands or reveal sensitive access tokens.

Prompt injection: The #1 threat to agentic AI

Prompt injection is the top threat in the OWASP Top 10 for Large Language Model Applications.[9] Unlike SQL injection, which exploits code parsing flaws, prompt injection exploits how language models parse language itself — the very medium through which they reason. There's no "syntax" to sanitize, no malicious pattern to detect.

Direct attacks override system instructions through crafted prompts. Indirect prompt injection is more insidious: malicious instructions embedded in external content — poisoned documents, web pages with invisible text and email signatures. When agents process this content, they execute the embedded commands unknowingly.

Second-order prompt injection turns compromised agents into trojan horses in multiagent systems. A low-privilege agent embeds malicious instructions in data it controls. When a more powerful agent queries that data during legitimate workflows, it inherits and executes the malicious instructions — turning collaboration into privilege escalation.

As agents become **multimodal** (processing text, images, audio and video), the attack surface expands. Adversarial techniques embed instructions in invisible pixel perturbations, ultrasonic frequencies or steganographic video encoding. The most alarming evolution: AI worms that self-replicate through agent networks via email chains and document workflows, exploiting semantic understanding rather than binary code.

Building the agentic security team

Human security teams cannot monitor thousands of agents executing millions of actions per hour. The scale, speed and nature of the task demand a different approach: building security teams made of agents.

This "agentic SecOps" approach deploys specialized security agents as digital antibodies. Human supervisors act as systems managers, setting policies and tuning their digital security force:

The "auditor" agent consumes reasoning logs in real-time, hunting for anomalies in intent: "Why did the finance bot suddenly query the employee database?"

The "provisioner" agent manages identities with just-in-time access, automatically creating and destroying permissions when tasks complete.

The "red team" agent runs continuous simulated attacks against production agents, proactively finding vulnerabilities.

The "responder" agent acts instantly when threats are detected — revoking credentials, isolating environments and providing forensic traces.

This agent-on-agent security model enables trust at scale: humans set policy; agents enforce it in real time.

Real-world implementation: Microsoft and industry leaders

Agentic security is already operational. The Microsoft multiagent architecture demonstrates how AI transforms reactive security into autonomous defense.[10] Specialized agents operate in parallel: threat detection agents analyze telemetry continuously, investigation agents enrich alerts with context, response agents execute containment within seconds and communication agents generate real-time reports.

These systems achieve massively parallel execution — processing hundreds of alerts simultaneously where human analysts might handle five per hour. This is human-supervised autonomous action: security engineers define policies and review high-confidence decisions while agents handle continuous monitoring.

Beyond security, **NTT DATA's Workplace Smart AI Agent™ Suite** demonstrates transformation across industries. Airlines using coordinated flight operations, maintenance, customer service and baggage-handling agents achieved 20% cost reductions. A bank deployed risk-assessment, customer engagement and compliance agents, reducing loan processing times with 30% cost savings. A consumer-goods company used market-analytics and supply chain agents to accelerate time to market and cut costs by 25%.

The market imperative

The stakes couldn't be higher. The AI governance market is projected to grow from $890 million to $5.8 billion over the next seven years, reflecting the critical importance organizations place on controlling these systems.[11] Yet many companies rush to deploy agents without adequate security frameworks, creating what security experts call "autonomy debt" — a technical and organizational liability that compounds over time.

The momentum is undeniable. The Microsoft announcement of Entra Agent ID at Build 2025 signals that major platforms are committing to tenant-governed agent identity. Gartner's formal recognition of machine identities as its own category validates that this is not a niche concern but a fundamental shift in how we architect systems.[12]

But we must be clear-eyed about the costs. Every security control imposes what researchers call the "security tax" — a trade-off between protection and operational efficiency.[13] Sandboxing agents reduces collaboration speed. Strict access controls slow task execution. Comprehensive logging consumes storage and compute resources.

This reality has led many organizations to adopt a pragmatic middle path: rather than treating agents as fully autonomous "ghost" entities, they implement what might be called a "governed marionette" approach — one that retains human oversight while avoiding the pitfalls of naive credential delegation.

1.1. The marionette: Governed agentic AI

The alternative path treats AI agents as **powerful tools requiring careful orchestration** — marionettes that perform complex tasks under human direction and oversight.

This approach doesn't reject autonomy; it contextualizes it within robust governance frameworks that recognize traditional security models fall short when dealing with autonomous systems.

Organizations implementing this approach deploy AI agents with:

- **Identity-based access controls** that treat each agent as a unique entity with specific permissions
- **Just-in-time access provisioning** that grants capabilities only when needed for specific tasks
- **Comprehensive audit trails** that log every agent action for review and compliance
- **Human-in-the-loop checkpoints** for high-stakes decisions

The results speak to effectiveness: Organizations implementing AI agents report a 35% reduction in incident response time while maintaining security standards that would be impossible with purely human teams.

Navigating the security tax

The art of agentic security is finding the balance — implementing controls strong enough to prevent catastrophic failures but flexible enough to enable the innovation and productivity that agents promise. This requires continuous tuning, measurement and adaptation as agent capabilities evolve and threats become more sophisticated.

Organizations that treat security as a binary choice — either open or locked down — will fail. The successful path is dynamic risk management: security policies that adapt to context, agents that prove their trustworthiness through behavior and governance systems that enable both protection and productivity.

Conclusion: The new rules for a new workforce

The security models of today are not ready for the workforce of tomorrow. Relying on traditional RBAC for digital workers is like using a horse-and-buggy rulebook for a fleet of self-driving cars.

The paradigm must shift:

1. **From static RBAC to dynamic access:** We must move beyond "roles" and toward dynamic, context-aware access. We need just-in-time permissions that grant an agent access for a specific task, for a specific duration and for a specific reason.
2. **From human identity to purpose identity:** The "ghost" model — tenant-governed, purpose-bound identities with centralized oversight — is the only scalable path forward. Microsoft Entra Agent ID demonstrates this is achievable at enterprise scale today, providing agents with distinct, persistent identity objects while maintaining organizational control, auditability and compliance.
3. **From action audits to reasoning audits:** We must demand and build "glass box" agents. If we cannot audit their internal reasoning, we cannot trust them. This requires comprehensive observability and standardized instrumentation.
4. **From human SOC to agentic SecOps:** We must automate security at the same speed as the workforce. We must build security agents to police, audit and protect our workforce agents. This isn't about replacing human judgment — it's about augmenting it with autonomous systems that can operate at the scale and speed that threats now demand.
5. **From reactive to proactive defense:** AI-powered security systems will increasingly predict noncompliance before breaches occur, shifting security from forensic analysis to preventive action.

The debate between the "marionette" and the "ghost" isn't just philosophical; it's the central security question of the next decade. Choosing the naive "marionette" model — where agents simply inherit their creator's full permissions — is easy, and it is catastrophically wrong. It leads to a future of unmanageable, over-privileged "sorcerer's apprentices." The path forward requires tenant-governed ghost identities — distinct agent identities that remain centrally managed, auditable and revocable — or carefully governed marionette frameworks with granular, dynamic access controls.

References

1. World Economic Forum. Unsecured AI Agents Expose Businesses to New Cyberthreats. World Economic Forum. September 2025.
2. CyberArk. AI Surge Drives a 40–1 Ratio of Machine to Human Identities. CyberArk.
3. Gartner. Gartner Predicts Over 40 Percent of Agentic AI Projects Will Be Canceled by End of 2027 Gartner Press Release. June 25, 2025.
4. Microsoft. Microsoft Entra Agent ID Documentation. Microsoft Learn
5. OpenTelemetry Authors. OpenTelemetry. Website
6. NVIDIA. Enhancing AI Transparency and Ethical Considerations with Model Cards. NVIDIA Developer Blog.
7. GDPR Info.eu. Art. 22 GDPR – Automated Individual Decision Making, Including Profiling. Website.
8. National Institute of Standards and Technology (NIST). AI Risk Management Framework. NIST.
9. OWASP Foundation. OWASP Top 10 for Large Language Model Applications. OWASP.
10. Microsoft. Designing Multi Agent Intelligence Microsoft Developer Blog. 2025.
11. David Prosser. Governance Start Ups Boom in the Battle to Keep AI Honest Forbes. May 29, 2025.
12. Gartner. Top Strategic Technology Trends for 2026. Gartner.
13. Peigne Lefebvre, Pierre, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher Obermaier, Christian Schroeder de Witt, and Esben Kran. Multi Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi Agent Systems. arXiv. February 2025.


NTT DATA