Identity Salon January 2026.v

Recap and Insights

Table of Contents

Executive summary

The first virtual Identity Salon of 2026 brought together identity practitioners, standards contributors, and academic researchers to examine how AI deployments—especially LLM-powered agents—change expectations around authentication, authorization, accountability, and audit.

The participants selected a single starting question via poll, then explored adjacent issues that quickly surfaced as tightly coupled: whether agents can protect secrets and sensitive data, how intent can be expressed and constrained (especially across multi-agent workflows), and where responsibility and liability should sit when agents act on someone's behalf.

A recurring theme was that many challenges are not brand-new. They echo familiar problems in enterprise identity—delegation, constrained authorization, and auditability—while AI amplifies them through scale, unpredictability, and new human–machine interaction patterns. Participants also flagged an "assumption mismatch": many mitigation ideas implicitly assume that today's agent behavior will remain stable over the next five years, even though AI capabilities and integration patterns are changing at an extraordinary pace.

This report captures themes, points of tension, and ideas raised during discussion; it does not represent consensus or formal recommendations.

Meeting Overview

The session built on questions generated during a previous in-person Salon discussion on AI and identity. Participants reviewed a short list of candidate questions and voted to select a starting point.

- *Should all AI-related identity data be ephemeral?*
- *Should AI-driven transactions be labeled differently—and if so, at what layer?*
- *Should "context" have boundaries, and how do we define and enforce them?*
- *As AI increasingly automates access and decision-making, how do we ensure outcomes remain explainable to auditors, regulators, and affected users?*
- *Will there ever be enough data for AI systems to be both accurate and accountable?*
- *And in a world of continuous identity, how do we avoid surveillance creep becoming the default?*

The question that garnered the most interest was:

> ***"As AI increasingly automates access and decision-making, how do we ensure outcomes remain explainable to auditors, regulators, and affected users?"***

It seemed to fit well with an additional question raised during topic selection:

> *Can we trust AI agents to protect the secrets needed for authentication—and if not, is "agent identity" viable in practice?*

The selected question served as an umbrella topic; secrets management and "trusting agents with data" became the first deep thread of the conversation.

Deterministic Code vs Non-deterministic Agents

Enterprises already manage code that can leak secrets and data. One side of the debate suggested that the solution here is the same: don't let application code (or agents) hold or access high-value secrets directly.

The counterpoint was that much of enterprise security "prior art" assumes comparatively deterministic systems and constrained execution paths. LLM-based agents introduce non-deterministic behavior that is harder to reason about, potentially increasing the frequency and unpredictability of failure modes.

From Keys to Data

Participants repeatedly broadened the question: if an agent can be prompted into handing over a private key, the same dynamic may apply to *any* sensitive data it can access.

In practice, many agents will hold or access information on behalf of multiple entities or roles. That raises the risk that the most harmful failure is not "credential loss" per se, but—more broadly—the unauthorized disclosure of data, or the taking of actions using valid authorization under unintended circumstances.

Mitigation Patterns

Participants discussed architectural patterns that reduce reliance on trusting an agent with secrets:

- Keep signing out of the agent: Use a constrained, non-LLM component to hold keys and sign on the agent's behalf.
- Hardware-backed key boundaries: Use HSM-style boundaries so that keys cannot be extracted even if an agent is manipulated.
- Avoid over-centralization: Implement per-agent or per-group signing and credential

boundaries to avoid creating single points of failure.

Participants also noted the operational tradeoffs: centralizing "signing on behalf of agents" can introduce deployment complexity, new high-value targets, and performance or availability concerns, especially in distributed environments.

Intent Management and Constrained Authorization

A substantial thread focused on intent: how it can be expressed, constrained, delegated, and audited for AI agents.

Intent Profiles as an Enterprise Control

Participants explored the idea of agent intent profiles, analogous to conditional access policies for humans:

- Agents are registered and managed as identities within an enterprise identity system.
- An agent's allowed intentions are defined and enforced by the "agentic system," not by the LLM itself.
- An "intent firewall" prevents requests that fall outside the agent's designed purpose.
- Tool access relies on short-lived tokens that exist only long enough to perform a bounded action.

Intent as an Adoption and UX Problem

Participants noted that even if intent can be formally represented, sometimes in a credential-like form, success depends on people's ability to express constraints clearly.

A recurring observation was that humans are not naturally good at specifying comprehensive boundaries. This shifts part of the problem from architecture to human-

centered design:

- How do people review, confirm, and understand what an agent is allowed to do?
- How do systems help users express intent without creating brittle, overly complex policies?

Multi-agent Workflows: Reduced Intent and Privacy Minimization

Participants highlighted a particular challenge in multi-agent chains: an upstream agent may need to delegate to downstream agents without exposing full intent or full context.

This raises two coupled requirements:

- Constrained redelegation: downstream agents should receive only the scope necessary to complete their part of the task.
- Data minimization requirements: sharing full intent across the chain can create unnecessary exposure.

Participants discussed the need for mechanisms that allow an agent to mint or pass along “reduced intent” credentials or proof objects, potentially drawing on patterns from derived credentials, though no single approach emerged as dominant.

Referenced Patterns and Historical Precedent

Because the entities and systems discussed were not represented in the room, participants were comfortable naming concrete examples to ground the discussion.

Agents to Payments Protocol (AP2) / payment-oriented agent protocols

Participants referenced emerging payment-oriented agent protocols, particularly the Agents to Pay Protocol (AP2), that explicitly formalize *buyer intent* as part of the

transaction flow. These approaches typically:

- Treat intent as a first-class object that can be verified and enforced.
- Bind intent to a specific transaction or scope, rather than granting open-ended authority.
- Place enforcement responsibility with an intermediary (for example, the paying bank), rather than solely with the merchant or the end user.

Participants found this model compelling because liability and fraud controls are clear: the enforcing party has both visibility and incentives to reject transactions that fall outside declared intent. However, questions were raised about whether such tightly bounded, intermediary-enforced models generalize beyond payments to broader enterprise or consumer agent use cases.

IBM Tivoli Privacy Manager (historical example)

As historical precedent, participants cited IBM Tivoli Privacy Manager, an early (and failed) attempt to externalize authorization decisions and include *intent* as an explicit input to access control.

The system was viewed as conceptually strong but operationally unsuccessful. The primary failure mode identified was adoption friction:

- Application developers were required to explicitly model and pass intent as part of transactions.
- The additional design and integration burden proved too high for widespread uptake.

Participants noted that this example remains instructive: even well-designed intent architectures can fail if they demand too much precision or effort from developers and application teams.

Learning by Example

Across both contemporary and historical examples, participants observed a recurring

pattern:

- Formalizing intent can significantly improve control, auditability, and fraud prevention.
- Success depends less on cryptographic soundness alone and more on *who enforces intent*, *where enforcement sits in the ecosystem*, and *how easily humans and developers can work with it*.

These lessons were seen as directly applicable to agentic systems, where intent must often be delegated, reduced, and enforced across multiple technical and organizational boundaries.

Auditability, Responsibility, and Liability

The group repeatedly returned to the question of what it means to hold anyone accountable in agent-mediated actions.

Can an Agent Ever be the Responsible Party?

Participants debated whether an AI agent can be responsible for an action, or whether responsibility always traces back to the human, organization, or service that prompted or deployed it.

A useful analogy emerged: tracking an agent's identity can be like tracking a vehicle identifier in an incident—helpful for investigation, but not sufficient for assigning responsibility.

Class vs Instance: What Must be Tracked

From a security and compliance standpoint, participants emphasized the need to distinguish:

- Agent class: what kind of agent this is (capabilities, policies, intended role)
- Agent instance: the specific execution/session that took an action

Participants noted that identical prompts may not yield identical behaviors, suggesting that prompts and tool invocations may need to be treated as auditable transactions.

Distinguishability vs Autonomy

A key point of tension was whether services should be able to distinguish "human-initiated" activity from "agent-initiated" activity. One position argued that restricting agents in ways humans would not be restricted effectively limits human autonomy. Another position argued that distinguishing user error from agent error matters operationally, if an agent repeatedly produces unwanted outcomes, someone needs to know what to fix.

Participants suggested that this may not be resolvable as a purely technical debate. Instead, liability allocation may determine what architectures and ecosystem norms emerge.

Binding Agents to Governance and Trust Protocols

Participants raised concerns about what binds agentic systems to trust and accountability expectations, particularly when models and their training or governance are not transparent to relying parties.

Approaches such as internal "constitutional" constraints and feedback loops were referenced as one way to steer model behavior. However, participants were skeptical that rule-based constraints alone can reliably enforce accountability in deep neural networks.

This remained an open question: what does it mean to be "strongly bound" to protocols for trust and responsibility when model behavior can be unpredictable and difficult—or

impossible—to validate externally?

Candidate Security Mechanisms

Participants proposed and discussed mechanisms that could strengthen control and audit, especially for high-risk actions.

Presence Gating for Critical Operations

One proposal was to require a verified human presence for certain critical agent actions, using biometric-style authentication as a practical “human-in-the-loop” control.

This was framed less as a universal solution and more as a safeguard for high-impact operations where the cost of unintended action is high.

Monitoring Agent-to-Agent Communication to Mitigate Oversharing

Participants discussed how multi-turn, language-based interactions between agents can lead to progressive oversharing, mirroring familiar human conversational dynamics. This effect was seen as particularly relevant in complex, multi-agent workflows where no single agent has full situational awareness.

Rather than proposing a single control, participants outlined a set of complementary monitoring practices that could be combined depending on risk tolerance and context:

- Selective capture or logging of agent-to-agent exchanges, bounded by policy and proportionality considerations.
- Security analysis of conversational artifacts to identify patterns associated with oversharing or

unintended disclosure.
- Traceability and alerting mechanisms that allow humans to detect, investigate, and intervene when risk signals appear.

These practices were framed as supporting both security and auditability by enabling post-incident reconstruction while also providing earlier warning signals that could limit the blast radius of unintended disclosures.

Five-year outlook: the assumption mismatch

Participants noted that many proposed controls implicitly assume that today's agent behavior and failure modes remain consistent over the next five years.

At the same time, AI capabilities and integration patterns are changing rapidly enough that five years may represent multiple architectural generations. Participants suggested that:

- Some mitigations may be transitional "bridges," not endpoints.
- Acceptance thresholds may shift as capabilities improve.
- Organizations may face a choice between incremental hardening of existing systems and more fundamental redesign.

Resources referenced

The discussion referenced existing materials related to authorization and delegation:

- UMA Implementer's Guide – Authorization assessment
  https://kantara.atlassian.net/wiki/spaces/uma/pages/4849800/UMA+Implementer+s+Guide#UMAImplementer
  Referenced as prior art for thinking about authorization relationships and constraints across multiple parties.
- Delegation use cases paper
  https://alanhkarp.com/UseCases.pdf
  Referenced as background reading on delegation patterns that reappear in agentic systems.

2026 Salon continuity and upcoming events

Participants were reminded that the Salon continues as a mix of in-person and virtual discussions throughout 2026.

- March 3, 2026: UNC Charlotte, US
- May 18, 2026: Berlin, DE (proximate to EIC)
- June 14-15, 2026: Las Vegas, US (proximate to Identiverse)
- October 18, 2026: Carlsbad CA (proximate to Authenticate)

Virtual Salon Gatherings will also be scheduled throughout the year

Open questions to carry forward

The session surfaced a set of unresolved questions, some of which we intend to explore more thoroughly in future Salons:

- What technical and governance mechanisms best prevent unauthorized disclosure when agents can be socially engineered?
- How can intent be expressed and constrained in ways that are usable for humans and enforceable for systems?
- How should reduced intent and constrained redelegation work across multi-agent chains without creating privacy spillover?
- What should be recorded for audit without creating prohibitive overhead or new surveillance risk?
- Where should liability sit when agents act autonomously within defined constraints—and what does that imply for whether services should detect agents at all?

About The Identity Salon

The Identity Salon™ provides a unique, exclusive environment where seasoned digital identity architects, technical standards experts, and researchers can engage in meaningful, protected conversations. Limited in size to foster genuine connections, this gathering allows experienced professionals to dive into complex, long-term challenges with peers who understand the depth and breadth of identity's impact.

We host the Identity Salon under the Chatham House Rule, facilitating candid dialogue that often isn't possible in larger, more public settings. Participants have the rare opportunity to explore the '5-year problems' in identity, share leading practices, and discuss emerging approaches with like-minded experts. Our aim is to bridge the gap between academic and industry research and real-world practice, connecting public and private sectors to advance knowledge and drive practical solutions.

Heather Flanagan, is the Principal at Spherical Cow Consulting, where she helps organizations navigate the fast-moving world of digital identity and Internet standards. With more than 15 years of experience translating complex technical concepts into clear, actionable strategy, Heather is known for her ability to bridge communities, guide collaborative work, and make standards development a little less intimidating. Named to the 2025 Okta Identity 25 as one of the top thought leaders in digital identity, Heather is also a regular speaker and writer, focusing on standards, governance, and the real-world challenges of identity implementation.

Ian Glazer, is the Chief Customer and Strategy Officer at SGNL. His extensive career in identity includes founding the advisory firm Weave Identity, serving as Senior Vice President for Identity Product Management at Salesforce where he led product strategy, and acting as a research vice president at Gartner, overseeing identity and privacy research. A prominent figure in the industry, Ian is the co-founder and Board Emeritus of IDPro. He also co-founded and is a board member of the Digital Identity Advancement Foundation, which works to remove financial barriers to participation in the digital identity space. Throughout his career, he has co-authored a patent, contributed to user provisioning specifications, and is a noted blogger, speaker, and photographer of his socks.

Andrew Hindle, is an independent consultant focusing on digital identity, privacy, cyber security, and corporate governance. A co-founder of The Identity Salon, Andrew is also Conference Chair for both the Identiverse and Authenticate conferences, serves as a non-executive member of the board at Curity, and chairs the UK Advisory Board at the Kantara Initiative. Andrew has over 25 years' experience in the software industry in a range of technical sales, pre-sales, product marketing, business development and corporate governance roles. He maintains CIPP/E, CIPM and CIPT privacy certifications with the IAPP, a CIDPRO certification with IDPro; and holds a BA in Oriental Studies (Japanese) from Oxford University and an advanced professional diploma in corporate governance. Outside of the world of identity, Andrew holds several governance roles with the Scouts at both local and county level; rides regularly with a local road cycling group; and plays keyboard, guitar and bassoon (not at the same time) with more enthusiasm than skill, and for an audience of one. Andrew is based in the UK

Thanks to Our 2026 Event and Report Supporters!

...and Splendid

Hindle Consulting

Spherical Cow Consulting

Weave Identity

Acknowledgments

The Identity Salon thrives because of the generosity and openness of its participants. Each month, technologists, policymakers, researchers, and implementers step into a Chatham-House-rule space to test ideas, challenge assumptions, and share experience. Their candor and willingness to explore difficult questions make the Salon work.