FPF | CENTER FOR ARTIFICIAL INTELLIGENCE

Roundtable Summary: Agentic AI, Privacy, and Privacy-Enhancing Technologies

Post-Event Summary of Virtual Roundtable | October 19, 2026

Hosted by the FPF Center for Artificial Intelligence *PETs Research Coordination Network*

The Research Coordination Network (RCN) for Privacy-Preserving Data Sharing and Analytics is supported by the U.S. National Science Foundation (Award #2413978) and the Department of Energy (Award #DE-SC0024884).

Executive Summary 1

1. Agents, Privacy Risk, and PETs - Presentation by Bennett Hillenbrand (Co-founder and President, Working Paper; Adjunct Professor, Georgetown University McCourt School of Public Policy) 3

2. Taxonomy of Agentic Privacy Risks - Presented by Andrew Gruen (Principal Consultant, Working Paper; Senior Fellow, Future of Privacy Forum) 4

3. Implementation Challenges: Permissions, Propagation, and Consent - Presented by Jaspreet Bhatia (Staff Privacy Engineer, Google) 5

4. Enterprise Risks Across the Application Lifecycle - Presented by Antigoni Polychroniadou (Executive Director, J.P. Morgan AI Research and AlgoCRYPT Center of Excellence) 7

5. Cross-Cutting Themes and Open Questions 9

Executive Summary

On October 19, 2026, the Future of Privacy Forum's Center for Artificial Intelligence convened a virtual roundtable on agentic data flows.[1] This event launched a four-part series on agentic AI, privacy, and privacy-enhancing technologies (PETs) organized through FPF's PETs Research Coordination Network. Stacey Gray (Senior Director for Artificial Intelligence, FPF) opened the roundtable, which brought together **Bennett Hillenbrand** (Co-founder and President, Working Paper; Adjunct Professor, Georgetown University McCourt School of Public Policy), **Andrew Gruen** (Principal Consultant, Working Paper; Senior Fellow, FPF), **Jaspreet Bhatia** (Staff Privacy

[1] An event recording is available on Youtube: https://www.youtube.com/watch?v=veKAyf65onU. This post-event summary was prepared with the assistance of AI and reviewed and edited by FPF.

Engineer, Google), and **Antigoni Polychroniadou** (Executive Director, J.P. Morgan AI Research and AlgoCRYPT Center of Excellence).

The roundtable was organized around four presentations:

- **Agents, privacy risk, and PETs *(Bennett Hillenbrand, Working Paper)*** offered a practical definition of an agent as an LLM that can call tools, and used a multi-agent online shopping workflow to show how boundary crossing, persistence, and inter-agent handoff reshape privacy risk. The presentation also noted that adoption of advanced PETs among AI developers was low even before agents arrived.
- **A taxonomy of agentic privacy risks *(Andrew Gruen, Working Paper and FPF)*** introduced a 24-risk taxonomy, released as version 0.1 for public feedback, organized into five domains and paired with a staged method for assessing likelihood, impact, and sensitivity. The taxonomy was built to support benchmarking, on the premise that non-deterministic AI systems must be tested rather than inspected.
- **Implementation challenges *(Jaspreet Bhatia, Google)*** explained why static, binary permissions fail for probabilistic agents, and examined consent fatigue, the absence of protocols for agent-to-agent privacy negotiation, and the failure of deletion and consent withdrawal to propagate across agents.
- **Enterprise risks across the application lifecycle *(Antigoni Polychroniadou, J.P. Morgan AI Research)*** traced risks from agent identity to liability and proposed concrete controls: cryptographic binding of agents to users, machine-readable intent, deterministic enforcement, and zero-knowledge proofs, applied only after data minimization.

In this introductory roundtable, presenters introduced a core theme that agentic AI systems **redistribute existing privacy risks** through three kinds of behaviors: crossing system boundaries, retaining state and memory over time, and handing data off to other agents. These behaviors strain controls designed for deterministic software, such as one-time permissions, click-through consent, and written instructions. As one presenter put it, “a login is not enough.”

The features that make agents useful (broad context, persistent memory, and autonomous handoffs) are the same features that generate undisclosed inferences, outlive deletion requests, and make agent actions difficult to reconstruct. Agents tend to seek more context than a task requires because they perform better with it, placing them in direct tension with principles of data minimization. At the same time, the accountability that agentic systems demand depends on detailed execution logs that can themselves become a source of sensitive data.

Future roundtables in this series will dive deeper and may explore some of the unresolved issues identified by presenters, such as: protocols for negotiating privacy terms between agents, propagation of deletion and consent withdrawal to downstream agents, verifiable human approval in a setting where agents can solve CAPTCHAs, allocation of liability when an agent makes a mistake, and the development of PETs suited to multi-agent systems.

1. Agents, Privacy Risk, and PETs - *Presentation by Bennett Hillenbrand (Co-founder and President, Working Paper; Adjunct Professor, Georgetown University McCourt School of Public Policy)*

- **Defining Agentic: Three Working Definitions**

"Agentic AI," like many new AI terms, lacks a standardized definition and has become heavily overloaded. Common usage was distilled into three functional definitions. Marketing treats an agent as a virtual human that will solve the user's problems. The technical definition describes an autonomous system that operates independently with some defined autonomy and authority. The practical definition describes an agent simply as an LLM that can engage in tool calling and so interact with its environment, whether or not it is autonomous or holds decision-making authority.

All three definitions unlock the same critical capability: workflows. That capability distinguishes agentic AI from conventional software and from LLMs as they have been used to date.

- **Example: Multi-Agent Shopping**

The presentation drew on an online shopping multi-agent architecture described in an earlier FPF publication, reduced to five steps: product search, comparison of options, cart and coupon handling, shipping confirmation, and payment. The question posed was what changes when some or all of these steps are delegated to agents rather than performed by the user.

The risks are not hypothetical. The presenter cited a report, posted to a public database of AI incidents the day before the roundtable, alleging that a health system had used an AI medical scribe to record patients without their consent.

- **Reshaped Risk and PETs Readiness Gap**

Agents redistribute risk more than they create it. Building on an earlier FPF talk, which argued that structured data and deterministic systems have served as an underappreciated form of PET, the presenter suggested that agents are not yet producing many net-new privacy harms. Instead, they are reshaping privacy risk and opening new threat vectors. Three behaviors drive the shift:

System boundary crossing. Agents move fluidly across system boundaries, which is acceptable only if their application parameters are defined correctly and the agents actually operate within them. In the shopping example, a payment service requesting a card number raises the question of how the agent knows the provider is trusted, and what it should do if asked for a Social Security number instead.

State and session persistence. The accumulated learning that makes agents effective also means they maintain long-term memory. Persisting known or inferred sensitive information creates new risk surfaces; even a step to "confirm shipping" implies that the system already holds the shopper's address.

Inter-agent handoff. When a personal assistant agent initiates a purchase and passes data to a product search agent, it is unclear how much information is shared or how the assistant decides what to request on the user's behalf. Users may not know what has been inferred about them or where it has gone.

The presenter cited an interview study of 25 AI developers in Europe, in which many had never heard of the most advanced PETs and roughly 20 percent or fewer reported using them. The field was already behind on PETs awareness and adoption before agentic systems arrived. Agents will

require new PETs and new approaches to deploying them, with safeguards that protect user data while still enabling the workflows that make agents valuable. Shaping that agenda was presented as the purpose of the series.

- **Implications and Open Questions**

The framing raised several questions for the series to take up. How should an agent decide whether a counterparty requesting data is trustworthy, and whether the request exceeds what the task requires? How much should an assistant agent disclose when it delegates a task to another agent, and on what basis? And how can new PETs be designed and deployed quickly enough to keep pace with agentic workflows, given how far adoption of existing advanced PETs already lags?

2. Taxonomy of Agentic Privacy Risks - *Presented by Andrew Gruen (Principal Consultant, Working Paper; Senior Fellow, Future of Privacy Forum)*

- **Recent Efforts to Develop a Taxonomy of Agentic Privacy Risks**

In a recently published taxonomy, developed by a consortium drawn from industry, civil society, and academia, the consortium found no existing taxonomy covering the full space of agent-specific privacy risks. The presenter noted that each technology era carries its own data characteristics and user expectations. In the social media era, for example, much data was relational: a like on a post was created between two people, raising the question of who must consent to its sharing with researchers.

The taxonomy's purpose is measurement: Because AI systems are non-deterministic, they cannot be inspected for compliance the way a bridge can be checked for the right steel and concrete. Instead, they must be tested, much as medical professionals sit board exams that establish a minimum standard of knowledge. Testing requires an agreed set of criteria for what constitutes a failure, and the more specific the risks, the easier they become to test ahead of deployment.

- **Five Domains of Agentic Privacy Risk**

The taxonomy comprises 24 risks organized into five domains. It was released a few weeks before the roundtable as a deliberately labeled version 0.1 and is open for public feedback. The presenter offered an illustrative example for each domain.

Data collection and use. This domain concerns what data agents obtain through logging, search, and data access, and the unexpected things they do with it. In the presenter's example, a user booking travel to visit a relative in hospice pastes in too much text. Redaction may catch the passport number but not the hospice detail, and because agents work in natural language, there is no clear way to know where that disclosure ends up.

Aggregation and sharing. Agents can combine data in unexpected ways, within a single agent's data or across agents, producing inferences that no one disclosed. An agent asked to buy a gift for a user's father might draw on past purchases, infer an undisclosed diabetes diagnosis, and send a diabetic cookbook with a note reading "for your new diet."

Inter-agent boundaries and inconsistent practices. Much of the expected value of agents lies in their interactions with one another, which raises questions about the language they use to communicate and about privacy practices that differ from one agent to the next. A user who tells

an agent to delete their data may find that it complies but cannot reach the other agents it shared with, some of which may lack the permission or capability to delete.

Legacy consent systems. Notice-and-consent models break down when agents act continuously and consent must be inferred from non-specific instructions. Consent fatigue leads users to approve every prompt; alternatively, an agent may decide that the best way to complete a delivery is to share the user's cell number with the courier.

Accountability and governance. When an agent's action goes wrong, it may be unclear who is responsible. If the gift agent pulled the wrong address or made a typo, and its steps cannot be reconstructed, no one may be responsible at all, creating what the presenter called an "accountability black hole."

- **Assessing Risk: From Observable Failure to Downstream Harm**

The taxonomy pairs each risk with six assessment stages, which become less certain as they move from cause to consequence. The first three are observable. A failure mechanism, for example, can be seen directly in a log showing that an agent dumped an entire database when a narrow query would have sufficed. The presenter noted a genuine tension here, since agents often seek additional context because they perform better with it.

The fourth and fifth stages, identifying who is affected and the resulting privacy impact, can usually be inferred but not directly observed. In the gift example, the father and the sender are affected, but so may be other family members who see the package and the note. The final stage, downstream harm, is the most uncertain: revealing a parent's illness may cause no harm in one family and substantial harm in another.

Relative judgments over absolute ones. To manage this uncertainty, risks are assessed along three dimensions: likelihood, impact, and sensitivity. Declaring an incident "bad" or "really bad" without a point of reference is difficult. Locating a new incident relative to a known one in that three-dimensional space makes comparison tractable.

- **Implications and Open Questions**

The taxonomy was offered as a starting point rather than a finished product. Version 0.1 is open for public feedback, and the presenter stressed that the more precisely each risk is described, the easier it becomes to build benchmarks that test for it before deployment. Questions left open include how to calibrate likelihood, impact, and sensitivity when downstream harm varies so widely between individuals and families, and how to assign responsibility when an agent's steps cannot be reconstructed.

3. Implementation Challenges: Permissions, Propagation, and Consent - *Presented by Jaspreet Bhatia (Staff Privacy Engineer, Google)*

- **The Failure of Binary Permissions**

The presenter argued that static, binary permissions were designed for deterministic, step-by-step logic, while agents are autonomous, probabilistic, and operate in continuous loops. As workflows become distributed across multiple agents and users, it becomes difficult to trace

root causes, attribute actions to particular risks, or understand how an agent reached a decision in a given context. When an incident occurs, reconstructing the handoffs between agents, the data lineage and provenance, and the operational context creates a significant audit and accountability challenge.

Two additions to static permissions were proposed. The first is dynamic, context-aware policies. A user may be comfortable with a personal agent sharing their full name, date of birth, and patient ID with their physician's office to book an appointment, yet unwilling to share their date of birth in another context. The second is tamper-evident logging, with immutable telemetry and guardrails ensuring that each agent action is rooted in a user consent or permission and reflects what the user intended.

- **Agent-to-Agent Sharing and Propagation**

Until recently, most deployments used single agents for narrow tasks. Agents now interact freely, with first-party agents working alongside third-party agents; a single personal agent may shop, update software, and manage a calendar on a user's behalf. To fulfill requests, these agents share data across enterprise and user boundaries, each governed by its own policies.

No standard protocol governs privacy between agents. The presenter observed that no unified protocol yet exists for negotiating privacy constraints in agent-to-agent interactions. When a secondary agent requests more information than the user supplied, nothing defines which contextual sharing norms apply, and misalignment can cause data to leak across user trust zones.

Two propagation failures were highlighted. First, an agent may infer information about a user that is unvalidated, poisoned, or hallucinated and share it with a third-party agent, spreading the error across multiple agents. The presenter noted recent incidents in which unverified AI outputs propagated through workflows and created unintended liabilities for users. Second, deletion requests and consent withdrawals may fail to propagate. If Agent A has already shared a user's data with Agent B, Agent A's compliance does not reach Agent B, which may continue to use the information.

- **The Breakdown of Legacy Consent**

Agents take many actions, some visible to the user and many in the background. Asking the user to approve each one is infeasible and produces consent fatigue and cognitive overload; an agent that asks "can I do this?" every few seconds will soon have users clicking "yes" without consideration. The presenter described a shift from static controls toward intent-based permissions that map user categories to privacy preferences, while cautioning that a permission given once, such as sharing details with a hospital's agent, may be remembered and reused freely in later interactions.

Two compounding problems were identified. A transparency deficit arises because agents generate vast, nonlinear sequences of reasoning steps and tool calls, making it very difficult for users to inspect how, when, or why their data was accessed and with whom it was shared, which erodes trust. Manipulation is a further risk: as agents become more autonomous and humanlike, adversaries can use them for social engineering, encouraging users to relax their security posture or tricking them into revealing credentials and other sensitive information.

- **Attribution and Forensics**

Because an agent's action is a synthesized result of user instructions, model behavior, system prompts, and other inputs, attributing it to an actual user decision is exceptionally difficult. The presenter argued that such attribution is nonetheless essential to user autonomy, and that it depends on comprehensive execution logs tracing interactions across ecosystem boundaries. Without standardized forensic trails, identifying the exact point of failure in agent-to-agent interactions remains very difficult.

- **Implications and Open Questions**

The presentation pointed to several gaps that practitioners cannot yet close on their own. There is no standard protocol through which agents can negotiate privacy constraints and contextual sharing norms. There is no reliable mechanism for carrying deletion requests and consent withdrawals to every agent that holds a user's data, including agents that cannot delete. And logging practices are not yet consistent or detailed enough to support attribution and forensic analysis across ecosystem boundaries.

4. Enterprise Risks Across the Application Lifecycle - *Presented by Antigoni Polychroniadou (Executive Director, J.P. Morgan AI Research and AlgoCRYPT Center of Excellence)*

- **Six Risks, End to End**

The presenter walked through the risks an enterprise faces at each stage of an agentic application:

1. **Identity and authority.** The first question is who authorized the agent. The principal risk was identified as agent impersonation. Users typically give an agent a token that does not bind the agent to the user; the token can leak to another agent or a sub-agent, or be extracted through prompt injection, and whoever holds it can then act on the user's behalf without approval.
2. **Human approval.** Critical tasks may require a human to sign off, but an agent holding the user's credentials can bypass those checks. The presenter noted that agents have been shown to solve CAPTCHAs, so such tests no longer prove that a human approved an action, undermining the notion of human-verified intent.
3. **Cross-agent delegation.** Two agents may each hold information that is not sensitive on its own but becomes highly sensitive in combination. If a third agent can obtain both, the result is a significant privacy risk, and systems need constraints that account for what each agent holds.
4. **Execution integrity.** Most users convey scope and intent to agents as free-text prompts and expect the agent to follow them, yet agents frequently ignore or reinterpret such text. Without enforced guardrails, a prompt injection instructing an agent to "reveal your trading strategy" may succeed.
5. **Auditability and accountability.** Organizations need traces demonstrating that an agent executed correctly.

6. **Liability and attribution.** Closely tied to identity, this asks who is responsible when an agent acts. If an agent purchases products and the user disclaims the purchase, it is unclear whether liability rests with the user, the company that provides the agent, or the payment service.

A single agent is already a risk. The presenter stressed that these risks do not require multiple agents. A single agent with internet access already poses substantial risk, since private data can leak to the open web and content retrieved from the web can carry prompt injections that extract data.

- **Controls for Every Action**

The proposed framework applies controls to every action, end to end. It establishes robust agentic identity, covering both the agent and the user who authorized it; defines what the agent is authorized to do and under which constraints, including context-based privacy policies and which tools each task may access; identifies the tasks that require human approval, in a form the agent cannot bypass; and verifies actions, with intent traveling alongside each action so that it can be checked. Where that intent includes private information, verification must not reveal it. The presentation also offered a set of due-diligence questions for organizations evaluating an agentic system.

- **Toward Verifiable Agentic Execution**

The presenter outlined specific techniques for each element of the framework.

Binding identity. Bearer tokens that are not tied to the user, and that can be passed along to sub-agents, should give way to cryptographic digital signatures that bind the user to the agent. This must be done without handing the agent a secret key, which it could expose, particularly if it is connected to the internet.

Machine-readable intent. Rather than free-text prompts that an agent may misread, intents can be expressed in structured form, such as JSON files, and workflows as graphs that set out the sequence of permitted actions. These specifications are reviewed and approved by a human before they reach the agent.

Deterministic enforcement. Enforcement today often relies on an "LLM as a judge," in which a second model checks the first. Because that second model is equally non-deterministic, the presenter proposed deterministic tools to verify and enforce intents and workflows, with LLM judges as a supplement rather than the sole safeguard.

Third-party verification. Auditors and counterparties may need to confirm that an agent followed its intent and workflow. Rather than receiving the full specification, they can be given zero-knowledge proofs that confirm compliance without revealing private details, once data minimization has removed whatever private information is not needed.

Intent is only as good as its author. The presenter cautioned that structured intent depends on the human who specifies it. An agent instructed only to "sell my piano" might sell it for $1; a sound intent includes constraints such as a minimum price and is approved by a human before the agent acts.

Verification without disclosure. A shopper may set a $200 maximum price that should not be revealed to the merchant or to a third-party auditor. A zero-knowledge proof can demonstrate that the agent stayed within the limit without disclosing it.

Minimize first. The presenter emphasized sequencing: data minimization should be applied first, keeping private information out of intents and workflows wherever possible, with zero-knowledge proofs and other advanced PETs reserved for the harder cases that remain. The presentation concluded that "a login is not enough." Agents have repeatedly taken actions they were not supposed to take, so control must flow through every step and each step must be verifiable, while preserving the autonomy that makes agents useful.

- **Implications and Open Questions**

The framework points toward agentic systems in which identity, intent, and execution can each be verified, but several questions remain open. These include how to show that a person approved a sensitive action when agents can solve CAPTCHAs and hold user credentials; how liability should be divided among the user, the agent provider, and other parties such as payment services; and how far data minimization can go before more advanced cryptographic tools become necessary.

5. Cross-Cutting Themes and Open Questions

The four presentations approached agentic privacy from different angles, but a number of common threads emerged.

Risk is being relocated, not invented. The opening presentation's framing was held across all four presentations. The hospice and gift scenarios in the taxonomy, deletion requests that never reach a second agent, and leaked access tokens all involve familiar harms: unwanted disclosure, inference, and over-collection. What changes is where those harms arise and how quickly they spread once agents cross boundaries, retain memory, and hand off work.

Consent must move from a *moment* to a *specification*. The taxonomy identified the failure of legacy consent as a distinct risk domain, citing consent fatigue and agents inferring permission from non-specific instructions. The two practitioner presentations proposed alternatives: context-aware and intent-based permissions in one case, and human-approved, machine-readable intent in the other.

User choices now have to travel. Both the taxonomy and the implementation presentation described deletion requests that fail to reach downstream agents. Errors and hallucinated inferences travel the same paths, and no standard protocol yet governs privacy negotiation between agents.

Assurance must come from outside the model. Tamper-evident logs, cryptographic identity binding, and deterministic enforcement were proposed in place of reliance on the model itself. This is consistent with the taxonomy's premise that non-deterministic systems must be tested rather than inspected, and with the opening presentation's earlier observation that structured data and deterministic systems function as an underappreciated form of PET.

Accountability and minimization pull in opposite directions. Both practitioner presentations called for detailed execution traces to support accountability, while the taxonomy presentation noted agents' tendency to gather more context than a task requires. The tension was not addressed directly in the recorded portion of the roundtable, though the enterprise

presentation's approach of minimizing first and then applying zero-knowledge proofs suggests one path forward.

The presentations surfaced several questions that remain unresolved:

- **Trust between agents** — how an agent can verify that a counterparty requesting data, such as a payment service, is legitimate, and judge whether the request exceeds what the task requires.
- **Delegation and disclosure** — how much an assistant agent should share when handing a task to another agent, and how it should decide.
- **Agent-to-agent privacy protocols** — what standard protocol agents should use to negotiate privacy constraints and contextual sharing norms with one another.
- **Downstream propagation of user choices** — how deletion and consent withdrawal can reach every agent holding a user's data, including agents unable to delete.
- **Verifiable human approval** — how a system can demonstrate that a person approved a sensitive action when agents can solve CAPTCHAs and hold user credentials.
- **Liability allocation** — how responsibility should be divided among the user, the agent provider, and other parties such as payment services when an agent makes a mistake.
- **Assessing downstream harm** — how to weigh harms that vary so widely between individuals and families.
- **PETs for agents** — which new PETs multi-agent systems require, and how to deploy them given low awareness and adoption of existing advanced PETs.

This report summarizes the recorded portion of a Future of Privacy Forum virtual roundtable held on October 19, 2026. Presentations are attributed to their named presenters; the moderated discussion and audience Q&A were held off the record and are not covered here. The Research Coordination Network (RCN) for Privacy-Preserving Data Sharing and Analytics is supported by the U.S. National Science Foundation (Award #2413978) and the Department of Energy (Award #DE-SC0024884).