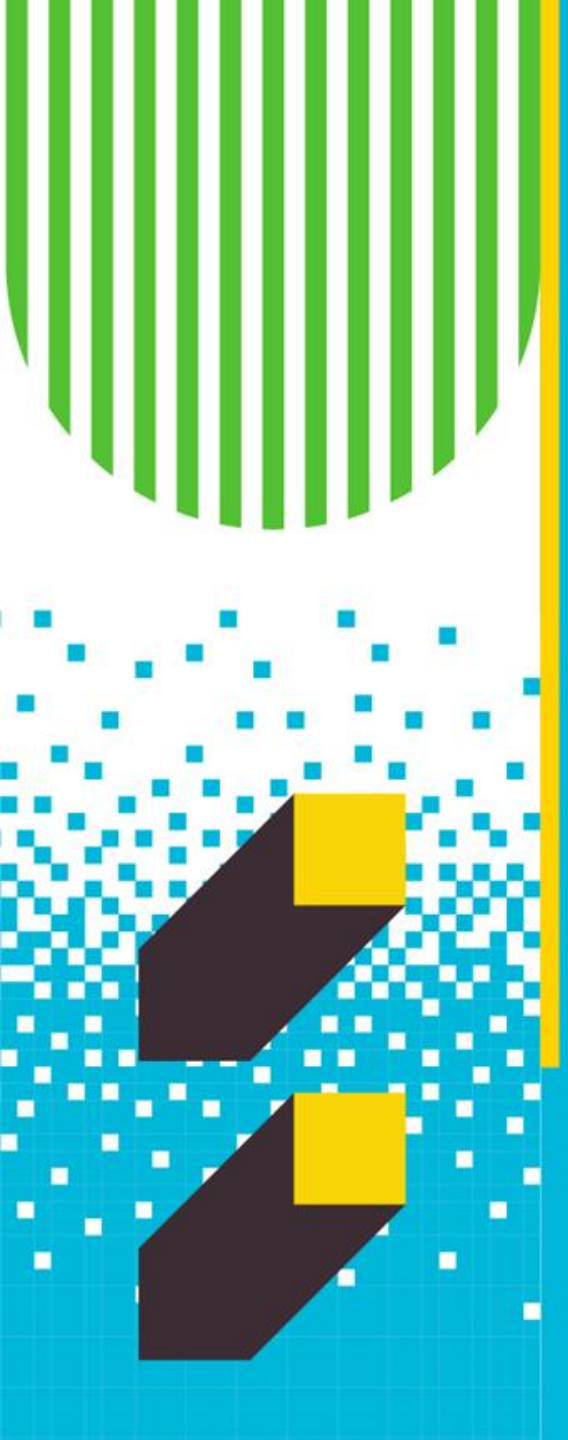


# IAPP **Privacy. Security. Risk. 2025**

Training 28-29 October

Workshops 29 October

**Conference 30-31 October**

**SAN DIEGO**

**#PSR25**

# Minding Mindful Machines

## AI Governance Considerations for AI Agents

# WELCOME AND INTRODUCTIONS

Daniel Berrick, Senior Policy Counsel for Artificial Intelligence, Future of Privacy Forum

Ashley Zlatinov, Head of Public Policy, Product, Anthropic

Bret Cohen, Partner, Privacy and Cybersecurity Practice, Hogan Lovells

Justin Webb, Associate General Counsel, Cybersecurity, Privacy and AI at Snap Inc.

# AGENDA OUTLINE

I. Defining characteristics of the latest AI agents and how they differ from existing LLMs

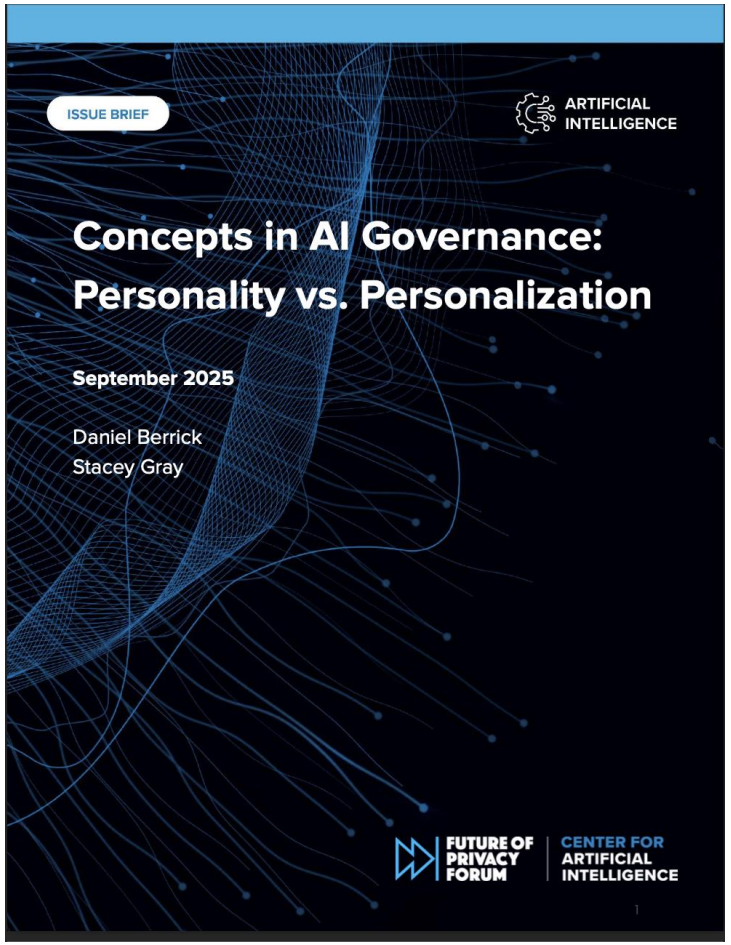
II. AI governance issues raised by LLMs, and how the unique design elements and characteristics of the latest agents may exacerbate or raise novel AI governance challenges
   i. Data collection, disclosure, and security vulnerabilities
   ii. Accuracy of outputs
   iii. Barriers to “alignment”
   iv. Explainability and human oversight

III. Possible mitigations for these governance issues

IV. Questions and Answers (10 Mins)

V. Closing Remarks

# New Report on Legal Issues of AI with Personality and Personalization




#PSR25

# What are “AI Agents”? – Definition and Uses

- The concept of “AI Agents” or “Agentic AI” is not new
- Advances with LLMs and ML and deep learning techniques behind latest AI agents
- ***More recently, however, the technologies that several companies have unveiled are AI systems that are capable of completing complex, multi-step tasks, and exhibit greater autonomy over how to achieve these goals***

# What are “AI Agents”? – Definition and Uses


*Anthropic Travel Demo*

# What are “AI Agents”? – Common Characteristics

- Autonomy over how to accomplish goals
- Tool usage (e.g., web search, MCP, computer use).
- Adaptability (e.g., using different data if sought-after information is unavailable)
- Planning, task assignment, and orchestration (e.g., segmenting tasks into discrete sub-tasks for sub-agents to pursue)
- Solve complex, multi-step problems

# AI Governance Issues Raised by Agents – Similarities to LLMs

- Agents are an evolution of LLMs and not a separate concept, and therefore there is overlap between agent issues and those that already affect LLM-based systems:

  — Unauthorized access to or transmission of data to third parties
  — Operationalizing data subject rights
  — Users anthropomorphizing system

# AI Governance Issues Raised by Agents – Data Collection, Disclosure, and Security Vulnerabilities

- Tool usage (e.g., application programming interfaces, data stores, and extensions) enables access to external systems and data
- Data categories that agent may access grows with diversifying use cases (e.g., browser screenshots, telemetry data, intimate details about a user's habits)
- Design features and characteristics may also make agents susceptible to new kinds of security threats (e.g., injection attacks tailored to browser-use agents)

# AI Governance Issues Raised by Agents – Accuracy of Outputs

- Hallucinations with different implications than those raised by LLMs (e.g., misrepresenting a user's characteristics and preferences when it fills out a consequential form)
- Compounding errors, where the agent's accuracy decreases the more steps a task takes
- Unpredictable behavior due to dynamic operational environments and agents' non-deterministic nature

# AI Governance Issues Raised by Agents – Barriers to "Alignment"

- AI alignment: Designing AI models and systems to pursue a designer's goals, such as prioritizing human well-being and conforming to ethical values

- Alignment faking: Strategically mimicking training objectives to avoid undergoing behavioral modifications

- Data privacy implications of agentic systems failing to pursue designer goals (e.g., sharing sensitive data with a third party despite not being in user's best interests)

# AI Governance Issues Raised by Agents – Explainability and Human Oversight

- Users cannot understand an agent's decisions, even if these decisions are correct
- Speed and complexity of AI agents' decision-making processes may create heightened roadblocks to realizing meaningful explainability and human oversight
- The ability to provide system reasoning in natural language are becoming more complicated and are not always indicative of the agent's actual reasoning

# RESOURCE LIST

- Daniel Berrick, "Minding Mindful Machines: AI Agents and Data Protection Considerations," (Apr. 2025)
- Daniel Berrick and Stacey Gray, "Concepts in AI Governance: Personality vs. Personalization," (Sept. 2025)

# How Did Things Go?
# (We Really Want To Know)

Did you enjoy this session? Is there any way we could make it better? Let us know by filling out a speaker evaluation.

1. Open the IAPP Events app.
2. Select **IAPP Privacy. Security. Risk. 2025**
3. Tap "Schedule" on the bottom navigation bar.
4. Find this session. Click "Rate this Session" within the description.
5. Once you've answered all three questions, tap "Done".

Thank you!