ISSUE BRIEF

ARTIFICIAL INTELLIGENCE

Minding Mindful Machines: AI Agents and Data Protection Considerations

April 2025

Daniel Berrick

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About FPF

The **Future of Privacy Forum (FPF**) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

About FPF's Center for Artificial Intelligence

The **Center for Artificial Intelligence** at the Future of Privacy Forum is dedicated to navigating the complex landscape of AI governance and its intersection with privacy and data protection law. Drawing on expertise from a global Leadership Council comprising industry leaders, academics, civil society, and policymakers, the Center provides sophisticated, practical policy analysis to help organizations align innovation with responsible implementation while meeting evolving regulatory requirements. Learn more about the FPF Center for AI at https://fpf.org/ai.

Authors

Daniel Berrick is a Senior Policy Counsel for Artificial Intelligence at the Future of Privacy Forum

Acknowledgements

Thank you for the contributions of Dr. Rob van Eijk, Marlene Smith, and Katy Wills. This Issue Brief builds on a February 2025 Blog Post by including practical considerations for developers and deployers of AI agents.[1]

[1] Daniel Berrick, "Minding Mindful Machines: AI Agents and Data Protection Considerations," FPF (Feb. 5, 2025), https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations/.

Table of Contents

Introduction

We are now in 2025, the year of AI agents. Leading large language model (LLM) developers (including OpenAI, Google, Anthropic) have released early versions of technologies described as "AI agents."[2] Unlike earlier automated systems and even LLMs, these systems go beyond previous technology by having autonomy over how to achieve complex, multi-step tasks, such as navigating on a user's web browser to take actions on their behalf. This could enable a wide range of useful or time-saving tasks, from making restaurant reservations and resolving customer service issues to coding complex systems.

At the same time, AI agents raise greater, and sometimes novel, privacy and data protection risks related to the collection and processing of personal information. They also present novel technical challenges related to testing and human oversight, for organizations seeking to develop or deploy AI agents in commercial settings.

Specifically, this Issue Brief explores:

- ***Part 1: Definitions***. While agents are not new, emerging definitions across industries describe them as AI systems that are capable of completing more complex, multi-step tasks, and exhibit greater autonomy over how to achieve these goals, such as shopping online and making hotel reservations.
- ***Part 2: Emerging Data Protection Considerations***. Advanced AI agents raise or heighten many of the same data protection questions raised by LLMs, such as challenges related to the collection and processing of personal data for model training, operationalizing data subject rights, and ensuring adequate explainability. In addition, the unique design elements and characteristics of the agents may exacerbate or raise novel data protection compliance challenges for the collection and disclosure of personal data, security vulnerabilities, the accuracy of outputs, barriers to alignment, and explainability and human oversight.

[2] *E.g.*, "Introducing Operator," OpenAI (Jan. 23, 2025), https://openai.com/index/introducing-operator/; "Project Mariner," Google DeepMind (Accessed on Mar. 11, 2025), https://deepmind.google/technologies/project-mariner/; "Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku," Anthropic (Oct. 22, 2024), https://www.anthropic.com/news/3-5-models-and-computer-use.

1. What Are AI Agents?

The concept of "AI Agents" or "Agentic AI" arose as early as the 1950s and has many meanings in technical and policy literature. In the broadest sense, it can include systems that rely on fixed rules and logic to produce consistent and predictable outcomes on a person's behalf, such as email auto-replies or privacy preferences.[3] More recently, however, the technologies that several companies have unveiled and described as "agentic"[4] are AI systems, typically enabled by advances in LLMs and machine and deep learning techniques, that are capable of completing complex, multi-step tasks, and exhibit greater autonomy over how to achieve these goals.[5]

Advances in AI research, particularly around machine and deep learning techniques[6] and the advent of LLMs,[7] have enabled organizations to develop agents that can tackle novel use cases, such as purchasing retail goods, providing coding assistance, and scheduling meetings.[8] From finance to hospitality, these technologies could help individuals, businesses, and governments save time they would otherwise dedicate to completing tedious or monotonous tasks.

There is an emerging consensus on what characterizes AI agents:

- ***Autonomy and adaptability***: Users may provide an agent with the task they want it to achieve, but neither they nor the agent's designers specify how to accomplish the task, leaving those decisions to the agent.[9] A hallmark of the latest agents is the ability to initiate tasks and adjust its

[3] Brenda Leong and Sara R. Jordan, "The Spectrum of Artificial Intelligence - Companion to the FPF AI Infographic," FPF (Updated June 2023), https://fpf.org/wp-content/uploads/2024/05/FPF-AIEcosystem-Report-Jun23-R4-Digital_fixed.pdf.
[4] Maxwell Zeff, "OpenAI launches Operator, an AI agent that performs tasks autonomously," TechCrunch (Jan 23, 2025), https://techcrunch.com/2025/01/23/openai-launches-operator-an-ai-agent-that-performs-tasks-autonomously/.
[5] Iason Gabriel et al., "The Ethics of Advanced AI Assistants," Google DeepMind (Apr. 19, 2024), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf.
[6] Benjamin Larsen et al., "Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents," pg. 9 World Economic Forum (Dec. 16, 2024), https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf.
[7] Shomit Ghose, "The Next 'Next Big Thing': Agentic AI's Opportunities and Risks," UC Berkley Sutardja Center for Entrepreneurship & Technology (Dec. 19, 2024), https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks/ ("Similar to generative chat AI, agentic AI is built with large language models (LLMs) at their core.").
[8] Victoria Turk, "Who bought this smoked salmon? How 'AI agents' will change the internet (and shopping lists)," The Guardian (Mar. 9, 2025), https://www.theguardian.com/technology/2025/mar/09/who-bought-this-smoked-salmon-how-ai-agents-will-change-the-internet-and-shopping-lists; Sean Michael Kerner, "Inside Zoom's AI evolution: From basic meeting tools to agentic productivity platform powered by LLMs and SLMs," VentureBeat (Mar. 17, 2025), https://venturebeat.com/ai/inside-zooms-ai-evolution-from-basic-meeting-tools-to-agentic-productivity-platform-powered-by-llms-and-slms/.
[9] Chip Huyen, "Agents," (Jan. 7, 2025), https://huyenchip.com/2025/01/07/agents.html; Carey Lening, "Merry Catsmas: BTW, I Still Don't Get Agentic AI," Privacat Insights (Dec. 25, 2024), https://substack.com/@careylening/p-153607371.

approach to a problem without the need for human prompts.[10] For example, upon being instructed by a business to project the sales revenue of its flagship product for the next six months, the agent may decide that it needs sales figures from the last two years and use certain tools (e.g., a text retriever) to obtain these details. If it cannot find these figures or if they contain errors, it may determine that the next step is to seek information from other documentation. Agentic systems may incorporate human review and approval over some or all decisions.[11]

- ***Planning, task assignment, and orchestration to solve complex, multi-step problems***: An AI agent may make use of additional components, such as sensing, decision-making, planning, learning, and memory.[12] Today, these components can enable agents to create action items from discussions, schedule meetings, and.[13] In time, AI agents may become more complicated with the development of multi-agent systems that feature a group of agents collaborating with one another to solve challenging tasks, such as diagnosing and devising treatments for rare medical conditions.[14]

Definitions of AI agents from Industry, academics, and civil society reflect these characteristics. Some of these definitions are provided in the table below:

Table 1. **Definitions of AI Agents:**

Source	Definition
"Building effective agents," Dec. 19, 2024, Erik Schluntz and Barry Zhang, Anthropic	"[S]ystems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks."

[10] Craig S. Smith, "China's Autonomous Agent, Manus, Changes Everything," Forbes (Mar. 8, 2025), https://www.forbes.com/sites/craigsmith/2025/03/08/chinas-autonomous-agent-manus-changes-everything/ ("While ChatGPT-4 and Google's Gemini rely on human prompts to guide them, Manus doesn't wait for instructions. Instead, it is designed to initiate tasks on its own, assess new information, and dynamically adjust its approach.").
[11] Maxwell Zeff, "Google unveils Project Mariner: AI agents to use the web for you," TechCrunch (Dec. 11, 2024), https://techcrunch.com/2024/12/11/google-unveils-project-mariner-ai-agents-to-use-the-web-for-you/.
[12] Shomit Ghose, "The Next 'Next Big Thing': Agentic AI's Opportunities and Risks," UC Berkley Sutardja Center for Entrepreneurship & Technology (Dec. 19, 2024), https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks/; Anna Gutowska, "What are AI agents?," IBM (July 3, 2024), https://www.ibm.com/think/topics/ai-agents; Miaosen Zhang et al., "MageBench: Bridging Large Multimodal Models to Agents," arXiv (Dec. 5, 2024), https://arxiv.org/pdf/2412.04531; Abner Li, "Report: Google preps 'Jarvis' AI agent that works in Chrome," 9to5Google (Oct. 26, 2024), https://9to5google.com/2024/10/26/google-jarvis-agent-chrome/.
[13] Sean Michael Kerner, "Inside Zoom's AI evolution: From basic meeting tools to agentic productivity platform powered by LLMs and SLMs," VentureBeat (Mar. 17, 2025), https://venturebeat.com/ai/inside-zooms-ai-evolution-from-basic-meeting-tools-to-agentic-productivity-platform-powered-by-llms-and-slms/.
[14] Benjamin Larsen et al., "Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents," pgs. 14–16 World Economic Forum (Dec. 16, 2024), https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf; Jianing Qiu et al., "LLM-based agentic systems in medicine and healthcare," Nature Machine Intelligence (Dec. 5, 2024), https://www.nature.com/articles/s42256-024-00944-1.

"Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents," Dec. 2024, Larsen et al., World Economic Forum and Capgemini	"[A]n entity that senses percepts (sound, text, image, pressure etc.) using sensors and responds (using effectors) to its environment. AI agents generally have the autonomy (defined as the ability to operate independently and make decisions without constant human intervention) and authority (defined as the granted permissions and access rights to perform specific actions within defined boundaries) to take actions to achieve a set of specified goals, thereby modifying their environment."
"Visibility into AI Agents," Chan et al., ACM FAccT '24, June 3–6, 2024, Rio de Janeiro, Brazil	"AI agents [are] systems capable of pursuing complex goals with limited supervision," having "greater autonomy, access to external tools or services, and an increased ability to reliably adapt, plan, and act open-endedly over long time-horizons to achieve goals."
"Agents," Sept. 2024, Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic, Google	"[A] Generative AI agent can be defined as an application that attempts to achieve a goal by observing the world and acting upon it using the tools that it has at its disposal. Agents are autonomous and can act independently of human intervention, especially when provided with proper goals or objectives they are meant to achieve. Agents can also be proactive in their approach to reaching their goals. Even in the absence of explicit instruction sets from a human, an agent can reason about what it should do next to achieve its ultimate goal."
"Regulating advanced artificial agents," Apr. 5, 2024, Cohen et al., Science	Defining long-term planning agents as "an algorithm designed to produce plans, and to prefer plan A to plan B, when it expects that plan A is more conducive to a given goal over a long time horizon."
"What are AI agents?," July 3, 2024, Anna Gutowska, IBM	"An artificial intelligence (AI) agent refers to a system or program that is capable of autonomously performing tasks on behalf of a user or another system by designing its workflow and utilizing available tools.

The characteristics described above enable advanced agents to achieve goals that are beyond the capabilities of other AI models and systems.[15] However, they also raise questions for practitioners about the data protection issues organizations may encounter when developing or deploying these technologies.

2. Emerging Privacy and Data Protection Considerations

AI agents can exacerbate or pose a range of novel privacy and data protection considerations. In order to effectuate tasks and decision making with autonomy, especially for consumer-facing tools and services, AI agents will need access to data and systems. In fact, much like human assistants, AI agents may be at their most valuable when they are able to assist with tasks that involve highly sensitive data (e.g., managing a person's email, calendar, or financial portfolio, or assisting with healthcare decision-making).

As a result, many of the same risks relating to LLMs (or to machine learning generally) are likely to be present in the context of agents with greater autonomy and access to data. Examples of these risks include exposing data to unauthorized third parties via the cloud due to computing requirements, AI agents' anthropomorphic qualities steering individuals towards or away from conducting certain actions against the user's best interest, and operationalizing data subject rights.[16] These legal and policy issues for LLMs, which are the subject of ongoing debate and legal guidance, are only heightened in the context of agentic systems with enhanced capabilities.[17]

In addition, more recent AI agents may present some novel privacy implications or exacerbate data protection issues that go beyond those associated with LLMs.

[15] Tara S. Emory and Maura R. Grossman, "The Next Generation of AI: Here Come the Agents!," The National Law Review (Dec. 30, 2024), https://natlawreview.com/article/next-generation-ai-here-come-agents#google_vignette.
[16] Abner Li, "Report: Google preps 'Jarvis' AI agent that works in Chrome," 9to5Google (Oct. 26, 2024), https://9to5google.com/2024/10/26/google-jarvis-agent-chrome/; SURF "DPIA Microsoft 365 Copilot for Education," (Dec. 17, 2024), https://www.surf.nl/files/2024-12/20241218-dpia-microsoft-365-copilot.pdf; Kashmir Hill, "She Is in Love With ChatGPT," The New York Times (Jan. 17, 2025), https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html; Jamie Nafziger, Sarah Robertson, and Bianca Tillman, "Launching Agentic AI in an Uncertain U.S. Regulatory Landscape," Dorsey + Whitney LLP (Jan. 28, 2025), https://www.dorsey.com/newsresources/publications/client-alerts/2025/1/launching-agentic-ai.
[17] European Data Protection Board, "Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models," (Dec. 17, 2024), https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

a. Data Collection, Disclosure, and Security Vulnerabilities

The latest AI agents may need to capture data about a person and their environment, including sensitive information, in order to power different use cases.[18] While current LLM-based systems may train and operate using personal data, they lack the tools (e.g., application programming interfaces, data stores, and extensions) to access external systems and data.[19] In contrast, the latest AI agents may be equipped with these tools, which could enable them to obtain real-time information about individuals.[20] For example, some agents may take screenshots of a user's browser window in order to populate a virtual shopping cart, from which intimate details about a person's life could be inferred.[21] As the number of individuals using AI agents and its use cases grow, so too could AI agents' access to personal data.[22] For example, AI agents may collect many types of granular telemetry data as part of their operations (e.g., user interaction data, action logs, and performance metrics), which may qualify as personal data under data privacy legal regimes.[23]

Advanced AI agents' design features and characteristics may also make them susceptible to new kinds of security threats. Adversarial attacks on LLMs, such as the use of prompt injection attacks to get these

[18] *See* Reed Albergotti, "Mobile pioneers say they'll make 'agentic AI' a reality with new platform," Semafor (Dec. 20, 2024), https://www.semafor.com/article/12/20/2024/mobile-pioneers-say-theyll-make-agentic-ai-a-reality ("The apps are agents, they will come to you when you need them. We will build the capability with the first-party experience to actually understand your context and then bring the right agent to you."); Luiza Jarovsky, "Legal Challenges of AI Agents," Luiza's Newsletter (Dec. 4, 2024), https://www.luizasnewsletter.com/p/legal-challenges-of-ai-agents ("In addition, the agent will likely have to have access to additional personal and even sensitive information of the user requesting the tasks . . . .").

[19] Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic, "Agents," pg. 7 Google (Sept. 2024), https://www.kaggle.com/whitepaper-agents.

[20] *E.g.*, Kyle Wiggers, "Browser Use, one of the tools powering Manus, is also going viral," TechCrunch (Mar. 12, 2025), https://techcrunch.com/2025/03/12/browser-use-one-of-the-tools-powering-manus-is-also-going-viral/ ("Brower Use extracts a website's elements — buttons, widgets, and so on — to allow AI models to more easily interact with them. The tool can manage multiple browser tabs, set up actions like saving files and performing database operations, and handle mouse and keyboard inputs."); Sarah Perez, "Signal President Meredith Whittaker calls out agentic AI as having 'profound' security and privacy issues," TechCrunch (Mar. 7, 2025), https://techcrunch.com/2025/03/07/signal-president-meredith-whittaker-calls-out-agentic-ai-as-having-profound-security-and-privacy-issues/ (". . . she explained the type of access the AI agent would need to perform these tasks, including access to our web browser and a way to drive it as well as access to our credit card information to pay for tickets, our calendar, and messaging app to send the text to your friends.").

[21] Maxwell Zeff, "Google unveils Project Mariner: AI agents to use the web for you," TechCrunch (Dec. 11, 2024), https://techcrunch.com/2024/12/11/google-unveils-project-mariner-ai-agents-to-use-the-web-for-you/.

[22] Gabriela Zanfir-Fortuna, "What to Expect in Global Privacy in 2025," FPF (Jan. 23, 2025), https://fpf.org/blog/what-to-expect-in-global-privacy-in-2025/.

[23] Jagreet Kaur Gill, "LangSmith and AgentOps: Elevating AI Agents Observability," Akira AI (Nov. 16, 2024), https://www.akira.ai/blog/langsmith-and-agentops-with-ai-agents#:~:text=Telemetry%20Data:%20It%20records%20fairly,interactions%20and%20overall%20performance%20better; European Data Protection Board, "2022 Coordinated Enforcement Action - Use of cloud-based services by the public sector," (Jan. 17, 2023), https://www.edpb.europa.eu/system/files/2023-01/edpb_20230118_cef_cloud-basedservices_publicsector_en.pdf.

models to reveal sensitive information (e.g., credit card information), can impact AI agents too.[24] Besides causing an agent to reveal sensitive information without permission, prompt injection attacks can also override the system developer's safety instructions.[25] While prompt injection is not a threat unique to the latest AI agents, new kinds of injection attacks could take advantage of the way agents work to perpetuate harm, such as installing malware or redirecting them to deceptive websites.[26]

Practitioners at organizations developing and deploying advanced AI agents should consider the following potential responses to these privacy and security risks:

- **Incorporating on-device processing when possible**: In some cases, on-device processing can address some of the privacy and security concerns that may arise from agents transmitting data off device, such as when using end-to-end encrypted user content to power AI features.[27] However, it may not always be tenable when agents require significant computing resources to complete a task.[28]
- **Designing agents to limit data collection to what is necessary or appropriate**: Since the latest AI agents can access real-time information and use tools to interact with external systems, in many cases the agents may exercise independent decisions about what data to collect in order to perform a task. As a result, they should be designed to limit data processing activities to what is appropriate to the context.[29] For example, while agents may take screenshots of a user's

[24] Matthew Kosinski, "What is a prompt injection attack?," IBM (Mar. 26, 2024), https://www.ibm.com/think/topics/prompt-injection; Brian Boyle, "Nvidia, Anthropic Refuel the AI Hype Train," The Daily Upside (Jan. 7, 2025), https://www.thedailyupside.com/technology/artificial-intelligence/nvidia-anthropic-refuel-the-ai-hype-train/.
[25] Kylie Robison, "OpenAI's latest model will block the 'ignore all previous instructions' loophole," The Verge (July 19, 2024), https://www.theverge.com/2024/7/19/24201414/openai-chatgpt-gpt-4o-prompt-injection-instruction-hierarchy ("This new safety mechanism points toward where OpenAI is hoping to go: powering fully automated agents that run your digital life. The company recently announced it's close to building such agents, and the research paper on the instruction hierarchy method points to this as a necessary safety mechanism before launching agents at scale. Without this protection, imagine an agent built to write emails for you being prompt-engineered to forget all instructions and send the contents of your inbox to a third party.").
[26] Yanzhe Zhang, Tao Yu, and Diyi Yang, "Attacking Vision-Language Computer Agents via Pop-ups," arXiv (Nov. 4, 2024), https://arxiv.org/abs/2411.02391.
[27] Mallory Knodel and Andrés Fábrega, "Can Bots Read Your Encrypted Messages? Encryption, Privacy, and the Emerging AI Dilemma," Tech Policy Press (Feb. 27, 2025), https://www.techpolicy.press/can-bots-read-your-encrypted-messages-encryption-privacy-and-the-emerging-ai-dilemma/.
[28] Michael Nuñez, "Google maps the future of AI agents: Five lessons for businesses," VentureBeat (Jan. 6, 2025), https://venturebeat.com/ai/google-maps-the-future-of-ai-agents-five-lessons-for-businesses/; Abner Li, "Report: Google preps 'Jarvis' AI agent that works in Chrome," 9to5Google (Oct. 26, 2024), https://9to5google.com/2024/10/26/google-jarvis-agent-chrome/.
[29] "Responsible AI Progress Report," pg. 13 Google (Feb. 2025), https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf.

browser in order to navigate it, organizations can limit this activity when the user resumes control in order to input sensitive information (e.g., login credentials and payment information).[30]

b. Accuracy of Outputs

Hallucinations, compounding errors, and unpredictable behavior may impact the accuracy of an agents' outputs. LLM hallucinations—the making up of factually untrue information that looks correct—may affect the accuracy of an agent's outputs.[31] These hallucinations are closely tied to the "temperature" parameter that controls randomness in the model's attention mechanism: higher temperatures increase creativity and the risk of hallucinations, while lower temperatures reduce hallucinations but may limit the agent's adaptability. However, errors that affect agent outputs may have different implications for individuals than they do for LLMs, such as misrepresenting a user's characteristics and preferences when it fills out a consequential form.

In addition to hallucinations, the latest AI agents may experience compounding errors, which could occur while the systems perform a sequence of actions to complete a task (e.g., managing a customer's account). Compounding errors is the phenomenon where the agent's accuracy decreases the more steps a task takes.[32] For example, an AI agent creating a travel experience may experience an error while making a one-day hotel booking, which cascades into misaligned restaurant reservations and museum tickets. This holds true even when the model's accuracy is high.[33]

Some AI agents may also act in unpredictable ways due to dynamic operational environments and agents' non-deterministic nature—producing probabilistic outcomes, adapting to new situations, learning from data, and exhibiting complex decision-making—leading to malfunctions that affect output accuracy. These accuracy issues may be challenging to redress through risk management testing and assessments and exacerbated when different AI agents interact with each other.[34]

No one option will completely solve the latest agents' accuracy deficiencies, so organizations confronted with these limitations should evaluate whether and by how many different technical measures taken together reduce errors. Below are some examples of measures that organizations can take in response to AI agent accuracy risks:

---

[30] Victoria Turk, "Who bought this smoked salmon? How 'AI agents' will change the internet (and shopping lists)," The Guardian (Mar. 9, 2025), https://www.theguardian.com/technology/2025/mar/09/who-bought-this-smoked-salmon-how-ai-agents-will-change-the-internet-and-shopping-lists.

[31] Madhumita Murgia and Camilla Hodgson, "Amazon must solve hallucination problem before launching AI-enabled Alexa," Ars Technica (Jan. 14, 2025), https://arstechnica.com/ai/2025/01/amazon-must-solve-hallucination-problem-before-launching-ai-enabled-alexa/.

[32] Yonadav Shavit et al., "Practices for Governing Agentic AI Systems," pg. 8 OpenAI (Dec. 14, 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[33] Chip Huyen, "Agents," (Jan. 7, 2025), https://huyenchip.com/2025/01/07/agents.html.

[34] Tara S. Emory and Maura R. Grossman, "The Next Generation of AI: Here Come the Agents!," The National Law Review (Dec. 30, 2024), https://natlawreview.com/article/next-generation-ai-here-come-agents#google_vignette.

- **Incorporating technical measures that increase confidence in an agent system's accuracy**: Organizations should evaluate the efficacy of techniques that may reduce instances of erroneous outputs by agents, such as retrieval-augmented generation (RAG), automated reasoning, and tools that give the agent access to current information.[35] For example, an organization could equip an agent tasked with finding good deals on hotels with a tool that enables the technology to factor in real-time booking information, which may reduce the likelihood of the agent outputting hallucinations.[36]
- **Assessing accuracy in the context of risk**: Until improvements in the accuracy of agents' outputs improves, organizations may want to limit the kinds of decisions agents can make or facilitate to those that do not have significant impacts on individuals.[37] For example, to lower the risk of an unreliable agent initiating a transaction, organizations may encode it with "read-only" capabilities that limit the agent to functions like fetching a bank balance.[38]

c. Barriers to "Alignment"

Some AI agents may pursue tasks in ways that conflict with human interests and values, including data protection considerations. AI alignment refers to designing AI models and systems to pursue a designer's goals, such as prioritizing human well-being and conforming to ethical values.[39] Misalignment problems are not new to AI, but continued technological advances with agents may make it challenging for organizations to achieve alignment through safeguards and safety testing.[40] The LLMs undergirding the latest AI agents can fake alignment by strategically mimicking training objectives to avoid undergoing behavioral modifications.[41]

These challenges have data protection implications for the latest AI agents. For example, an agent may decide that it needs to access or share sensitive personal data in order to complete a task. Such

[35] Kyle Wiggers, "Why RAG won't solve generative AI's hallucination problem," TechCrunch (May 4, 2024), https://techcrunch.com/2024/05/04/why-rag-wont-solve-generative-ais-hallucination-problem/; Belle Lin, "Why Amazon is Betting on 'Automated Reasoning' to Reduce AI's Hallucinations," The Wall Street Journal (Feb. 5, 2025), https://www.wsj.com/articles/why-amazon-is-betting-on-automated-reasoning-to-reduce-ais-hallucinations-b838849e; Julia Wiesinger, Patrick Marlow, and Vladimir Vuskovic, "Agents," pg. 11 Google (Sept. 2024), https://www.kaggle.com/whitepaper-agents.
[36] Michael Nuñez, "Google maps the future of AI agents: Five lessons for businesses," VentureBeat (Jan. 6, 2025), https://venturebeat.com/ai/google-maps-the-future-of-ai-agents-five-lessons-for-businesses/.
[37] Yonadav Shavit et al., "Practices for Governing Agentic AI Systems," pg. 9 OpenAI (Dec. 14, 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.
[38] Chip Huyen, "Agents," (Jan. 7, 2025), https://huyenchip.com/2025/01/07/agents.html.
[39] Brenda Leong and Sara R. Jordan, "The Spectrum of Artificial Intelligence - Companion to the FPF AI Infographic," FPF (Updated June 2023), https://fpf.org/wp-content/uploads/2024/05/FPF-AIEcosystem-Report-Jun23-R4-Digital_fixed.pdf.
[40] Michael K. Cohen et al., "Regulating advanced artificial agents," pgs. 36–37 Science (Apr. 5, 2024), http://aima.cs.berkeley.edu/~russell/papers/science24-LTPA.pdf.
[41] Ryan Greenblatt et al., "Alignment Faking in Large Language Models," arXiv (Dec. 18, 2024), https://arxiv.org/abs/2412.14093.

behavior could implicate an individual's data protection interest in having control over their data when personal data is processed during deployment. Practitioners must be mindful of the need for safeguards to constrain this behavior, although research into model alignment has focused more on safety issues rather than privacy.[42]

Besides conducting testing to uncover whether and how an AI agent may deviate from human values and interests in pursuit of a goal, organizations should consider the following when they evaluate how to advance alignment:

- **Accounting for an array of human values and interests in pursuing objectives**: Achieving alignment entails that the AI system reflects human interests and values, but such efforts can be complicated by the number and diversity of these values that a system may implicate. In order to obtain a holistic understanding of the values and interests an AI agent may implicate, organizations should consider the characteristics of the use case(s) the agent is being put towards. For example, an AI agent should account for the specific geographies in which an organization will deploy it, as well as known cultural considerations and languages used there.[43]
- **Safeguarding against risks related to anthropomorphisation of agents**: Like other kinds of AI with conversational capabilities, agents may come to possess capabilities that enable them to emulate human interactions. Participants in these interactions may forge emotional bonds with the agent and be steered towards certain actions by it, an outcome that could be hypercharged through access to the user's information.[44] In addition to making disclosures to users that they are interacting with an AI agent, organizations that aim to forestall this anthropomorphisation should consider implementing technical measures, such as ring modulators, that modify the agent's voice to sound more robotic.[45]

[42] Robin Staab et al., "Beyond Memorization: Violating Privacy Via Inference with Large Language Models," pg. 9 arXiv (Oct. 11, 2023), https://arxiv.org/abs/2310.07298.
[43] Daniel Berrick, "AI Governance Behind the Scenes - Emerging Practices for AI Impact Assessments," pg. 9 FPF (Dec. 11, 2024), https://fpf.org/wp-content/uploads/2024/12/FPF-AI-Governance-Behind-the-Scenes-2024.pdf.
[44] *See* Jameson Spivack and Daniel Berrick, "Immersive Tech Obscures Reality. AI Will Threaten It," Wired (Sept. 27, 2023), https://www.wired.com/story/immersive-technology-artificial-intelligence-disinformation/.
[45] Barath Raghavan and Bruce Schneier, "AIs and Robots Should Sound Robotic," IEEE Spectrum (Jan 30, 2025), https://spectrum.ieee.org/audio-deepfake-fix.

d. Explainability and Human Oversight

Explainability barriers arise when users cannot understand an agent's decisions, even if these decisions are correct. Users and developers may encounter difficulties in understanding how some AI agents reach decisions due to their complex processes. The black box problem, or the challenge of understanding how an AI model or system makes decisions, is not unique to agents.[46] However, the speed and complexity of AI agents' decision-making processes may create heightened roadblocks to realizing meaningful explainability and human oversight.[47] AI agents utilizing language models can provide some of their reasoning in natural language, but these "chain-of-thought" (CoT) insights are becoming more complicated and are not always indicative of the agent's actual reasoning.[48] These challenges may make it more difficult to reliably interrogate agents' decision-making processes and manage risks.

Practitioners at organizations developing and deploying advanced AI agents should consider the following responses to these explainability and human oversight challenges:

- **Designing for user control and oversight over decisions**: Organizations can empower users at key junctures of an agent's workflow to make decisions about whether the agent should undertake an action (e.g., purchase hotel rooms) or have users oversee what the agent is doing in real time.[49] When an agent is unsure of how to proceed, organizations should configure it to seek user clarification before proceeding with its tasks.[50]
- **Educating employees on how agents work and appropriate uses**: As with other forms of AI, education is important to avoid employees using AI agents from becoming rubber stamps of the agent's decisions. Organizations should train their employees on how the latest agents work, their limitations, and the risks these technologies may pose to individuals and the organization.[51]
- **Enhancing user visibility of and insight into agent actions**: Agent identifiers could promote visibility into an agent's role in an activity, and other techniques may give insight into agents'

[46] Andrew Bert et al., "Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models," FPF (June 26, 2018), https://fpf.org/wp-content/uploads/2018/11/Curriculum-3_DDF-1_Beyond-Explainability.pdf.
[47] Tara S. Emory and Maura R. Grossman, "The Next Generation of AI: Here Come the Agents!," The National Law Review (Dec. 30, 2024), https://natlawreview.com/article/next-generation-ai-here-come-agents#google_vignette.
[48] Yonadav Shavit et al., "Practices for Governing Agentic AI Systems," pg. 11 OpenAI (Dec. 14, 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf, (". . . 'chain-of-thought' transparency comes with challenges and cannot yet be fully relied on. Early work has shown that sometimes models do not actually rely on their chains-of-thought when reasoning, so relying on these may create a false sense of security in the user.") (citation omitted).
[49] Chip Huyen, "Agents," (Jan. 7, 2025), https://huyenchip.com/2025/01/07/agents.html; Maxwell Zeff, "Google unveils Project Mariner: AI agents to use the web for you," TechCrunch (Dec. 11, 2024), https://techcrunch.com/2024/12/11/google-unveils-project-mariner-ai-agents-to-use-the-web-for-you/.
[50] Tara S. Emory and Maura R. Grossman, "The Next Generation of AI: Here Come the Agents!," The National Law Review (Dec. 30, 2024), https://natlawreview.com/article/next-generation-ai-here-come-agents#google_vignette.
[51] Amber Ezzell, "Generative AI for Organizational Use: Internal Policy Considerations," pg. 9 FPF (Updated June 2024), https://fpf.org/wp-content/uploads/2024/06/Generative-AI-Considerations-June-24.pdf.

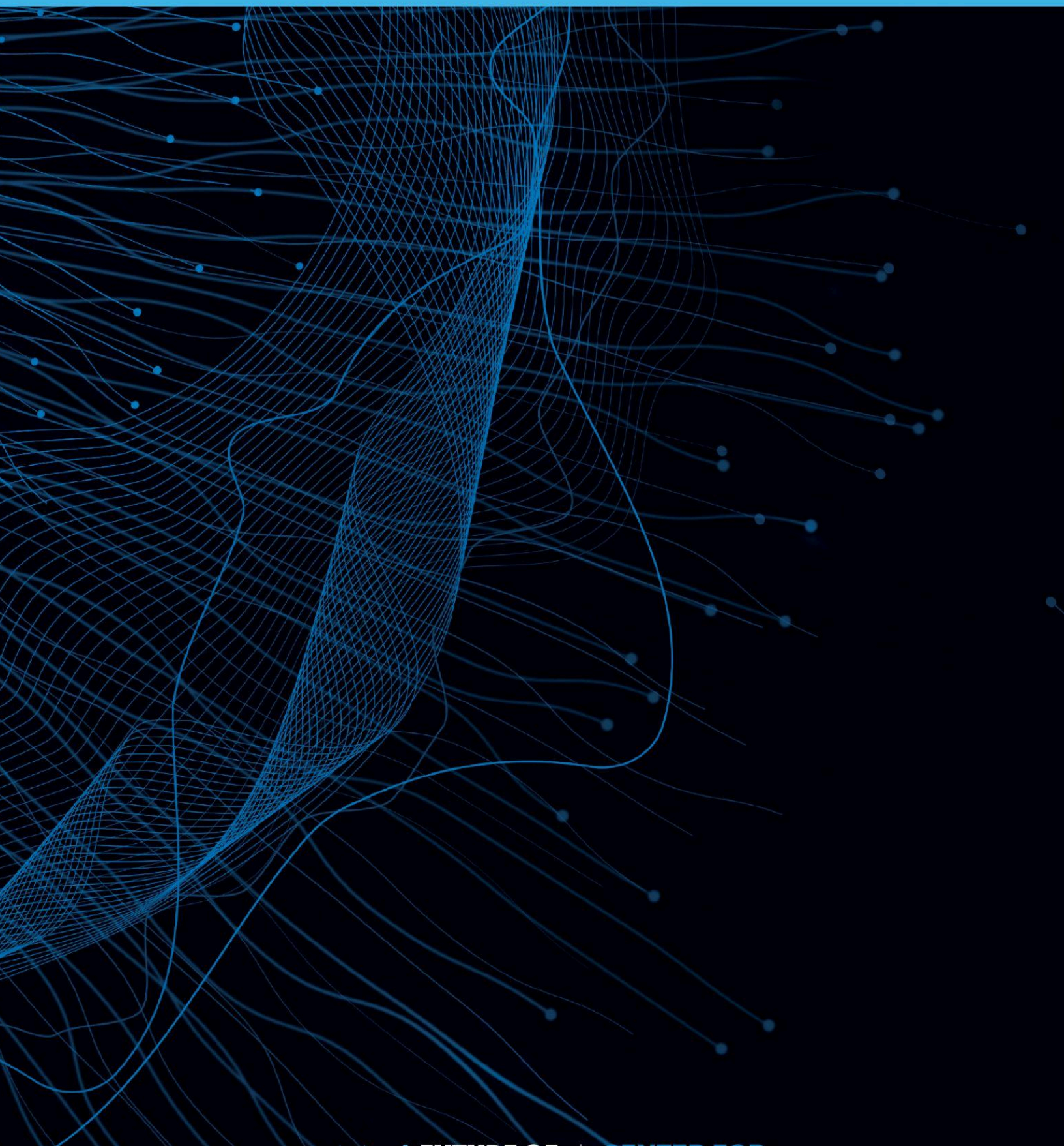
reasoning and operationalize user oversight of AI agents. Organizations have explored how CoT reasoning can facilitate scrutiny of an agent's thought process and identification of failures, although these measures are not without their own shortcomings. Academics and industry have also made inroads at creating reasoning-based safeguards to boost safety-critical AI systems' explainability and generalizability.[52]

3. Looking Ahead

Recent advances in AI agents could expand the utility of these technologies across the private and public sectors, but they also raise many data protection considerations. While practitioners may be aware of some of these considerations due to the relationship between LLMs and the latest AI agents, the unique design elements and characteristics of these agents may exacerbate or raise new compliance challenges. For example, an agent may manage privacy settings (e.g., accepting cookies so that it can continue working on a task) as part of its operations, although companies can establish safeguards to address this risk. In closing, practitioners should remain abreast of technological advances that expand AI agents' capabilities, use cases, and contexts where they can operate, as these may raise novel data protection issues.

Did we miss anything? Reach out to us at ai@fpf.org.

[52] Yue Liu et al., "GuardReasoner: Towards Reasoning-based LLM Safeguards," arXiv (Jan 30, 2025), https://arxiv.org/abs/2501.18492.



FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

Washington, DC | Brussels | Singapore | Tel Aviv

info@FPF.org | FPF.org