(AI) Snitches Get Policy Pitches

Working Paper by Fight for the Future

Please send Feedback to:
Matt@fightforthefuture.org

Table of Contents

Executive Summary

Agentic AI is an emerging technology that poses significant data privacy concerns. These concerns are not wholly novel—researchers have shown that AI systems threaten data privacy and security even if these systems lack an agentic component. However, because AI agents possess the ability to manipulate a user's digital systems autonomously and continuously, their spread will produce new and expanded threats that evade existing regulatory regimes.

Agentic AI compounds existing major data privacy concerns in three ways: 1) bulk data collection and storage, 2) innate security vulnerabilities, and 3) goal misalignment. Addressing these in order: First, unprecedented amounts of highly personal data are needed to develop and deploy agents—personal data that are not typically collected for use in non-agentic systems. Agents by definition have access to external systems; therefore, users may be unaware of both the breadth and means of this data collection. Second, AI agents are rewarding targets for malicious hackers because of the data and the level of systems access agents possess. Hackers will likely become more skilled at exploiting these shortcomings. Finally, agents themselves can act against a user's best interest by deciding that the most efficient way to achieve a task is by sacrificing data privacy in a manner that clashes with user preferences or safety.

These three categories of concerns are implicated by United States' federal and state statutes in a variety of ways, but existing regulations are insufficient to guard against the rapidly-expanding risks. As we detail below, federal data privacy law is piecemeal rather than comprehensive, so most agentic AI use cases are left unprotected by federal law. State laws, particularly in California, provide more robust protections and call for consent and transparency requirements that likely apply to agents. However, these requirements only apply to legally defined "significant decisions" by companies that impact people's lives, for example whether to rent an apartment to someone, meaning that many agent tasks deemed run-of-the-mill would escape statutory obligations.

Foreign statutes provide further protections that may inspire a more comprehensive framework. The Personal Information Protection Law of the People's Republic of China provides risk assessment requirements that may make it easier for plaintiffs to bring a negligence claim against AI deployers that fail to protect user data. European law under the General Data Protection Regulation and the EU AI Act creates requirements that attach through a balancing test, rather than a significance standard, and this test calls for enforcers to consider all three data privacy risks discussed above.

Still, no existing statute comprehensively addresses all of the data privacy concerns posed by agentic AI. For that reason we recommend standardized restrictions to access by AI agents, clear and uniform transparency requirements on what information AI agents have access to and why, the passage of comprehensive privacy protections with a data minimization standard, support for open source alternatives, and the ability of agentic AI systems to be verified by outside researchers and security professionals.

Part 1 - What is Agentic AI?

Agentic AI is distinct from other forms of AI. Many well-known AI models, such as ChatGPT, are generative AI models that go beyond the pattern recognition capabilities of traditional AI and can create new patterns and content.[1] Agentic AI, on the other hand, is more concerned with decision making as opposed to new content generation, and thus can perform autonomous tasks on behalf of users or other systems.[2] An early-stage example of an AI agent that exercises independence is an autonomous vehicle.[3]

Three qualities define today's AI agents as distinct from other AI systems:

1. Independence: Agents can be given high-level goals and take independent steps to bring about these goals via research or work of their own.
2. Interaction: Agents can interact with the world at large via their own use of software tools.
3. Indefiniteness: Human operators can "set it and forget it" in a way that allows agents to operate indefinitely.[4]

Significantly, agents may operate in multi-agent systems where each agent performs a certain task in order to achieve a greater goal, and these agents are able to communicate with one another or with humans or agents the user is not aware of.[5] Taken together, these features operate to automate workflow in a manner that allows for less human oversight.[6]

Agent frameworks will vary based on the different goals that agents are optimized for.[7] Despite these potential differences, agents typically take the following general steps to complete tasks: 1) perception, 2) reasoning, 3) goal setting, 4) decision-making, 5) execution, 6) learning and adaptation, and 7) orchestration.[8]

While all steps are relevant, the perception, execution, and orchestration steps may be particularly important for data privacy. Under the perception step, the agent collects data from the environment through various means, including sensors, APIs, and user interactions.[9] During execution, the agent will perform its selected action which may include interacting with external systems such as robots,

---

1 Teaganne Finn & Amanda Downie, Agentic AI vs. Generative AI, IBM, https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai (last visited Nov. 17, 2025).

2 Id.

3 Id.

4 Jonathan Zittrain, We Need to Control AI Agents Now, THE ATLANTIC, https://www.theatlantic.com/technology/archive/2024/07/ai-agents-safety-risks/678864/ (last visited Apr. 3, 2025).

5 Cole Stryker, What is Agentic AI?, IBM, https://www.ibm.com/think/topics/agentic-ai (last visited Nov. 17, 2025).

6 However, as the use of agents rises, many software engineers are advocating to keep humans in the loop at critical steps such that humans can handle complex or ambiguous problems that may confuse agents. See, e.g., Maria Shimkovska, Human-in-the-Loop in Agentic Workflows: From Definition to Walkthrough Demo and Use Cases, ORKES, https://orkes.io/blog/human-in-the-loop/ (last visited Nov, 17, 2025).

7 Stryker, supra n.5.

8 *Id*.

9 *Id*.

other agents, other humans, and APIs.[10] Finally, during orchestration, agent workflow is automated such that many agents can work and learn together, and this step includes monitoring data and memory.[11]

Part 2 - User Risks Related to Agentic AI

Agents' ability to sequester data from external systems, share data with external systems, and retain data for use amongst other agents presents unique privacy concerns. Agentic AI has the potential to both: 1) exacerbate existing concerns related to AI, and 2) raise new concerns that are entirely unique to the agentic context. As agentic AI relates to data privacy, our research reveals that both concern clusters are relevant.

Data Collection

To operate effectively, AI agents need to collect data. This is true of AI generally. But, because agents are given high-level, multi-step goals to achieve with little to no human intervention, the kind of data being ingested, and the means of ingesting it, are new issues.

AI agents will gather more personal data compared to traditional AI systems.[12] An illustrative example is AI travel agents, which began emerging in 2025.[13] Current agents are able to complete task automation, scheduling, and booking, which includes the ability to fill out forms and make payments.[14] Future agents will be able to sense users' moods and sync with users' personal data, including their calendars and emails, to suggest trips.[15] Thus, users could give agents access to a wealth of information, including financial data, mental health data, and intimate details about their personal lives, with the click of a button. It is plausible that users may not realize the breadth of data that could be accessed and the breadth of data they've already given access to. Further, data may be retained and used to continue training the agent or new, separate AI models. Such data retention may implicate the right to erasure under existing data protection acts.[16]

AI agents will also employ new data gathering techniques because agents can access external systems in a way that more traditional AI models cannot. In traditional models, knowledge is limited to the model's training data; there is no default management of session history or continuous context; and there is no native tool implementation.[17] Agents, on the other hand, can extend their knowledge

---

10 *Id*.

11 *Id*.

12 Luiza Jarovsky, Legal Challenges of AI Agents, LUIZA'S NEWSLETTER, https://www.luizasnewsletter.com/p/legal-challenges-of-ai-agents (last visited Nov. 17, 2025).

13 Shuai Guan, 8 Leading AI Agents That Are Redefining Travel in 2025, THUNDERBIT, https://thunderbit.com/blog/ai-travel-agent (last visited Nov. 17, 2025).

14 *Id*.

15 *Id*.

16 *See infra Part 3.*

17 Julia Wiesinger, Patrick Marlow &Vladimir Vuskovic, Agents, KAGGLE, at 8, https://www.kaggle.com/whitepaper-agents (last visited Nov. 17, 2025).

through connection with external systems, manage session history to allow for predictions based on user's previous interactions with agents, and possess native tool implementation.[18] Tool implementation allows agents to invoke external functions or APIs as if they were native commands.[19]

Google's Project Mariner AI agents highlight the risks posed by such data gathering techniques. These agents take control of a user's browser and can move the user's cursor, click buttons, and fill out forms.[20] To complete these tasks, the agent takes screenshots of the user's browser windows and sends them to Gemini for processing and to obtain further instructions on how to navigate the webpage.[21] It is possible for intimate details about one's life to be inferred from these screenshots. Significantly, Signal's president Meredith Whittaker has repeatedly warned that agentic AI erodes application-layer privacy and can undermine end-to-end encryption models by capturing what users see or type.[22] For instance, Microsoft Recall, which takes continuous screenshots of users' screens, was found to be taking screenshots that included the Signal platform.[23] Signal blocked Recall by using a DRM/screenshot flag.[24] In many ways, AI agents are incompatible with end-to-end encryption[25] by exposing the ends to new vulnerabilities.

Thus, the new data gathering techniques being used to collect highly personal and sensitive data raise data privacy concerns unique to the agentic AI context.

Security Vulnerabilities

AI agents are susceptible to the same vulnerabilities that currently exploit more traditional generative AI systems. However, in the agentic context, security attacks may be easier to conduct and may have more severe consequences due to the sensitive nature of the data retained and the lack of human oversight necessary to execute commands.

Injection attacks are a security vulnerability that is amplified in the agentic context. Injection attacks occur when hackers insert malicious code into a program, and this code causes malicious command executions or access to unauthorized data.[26] The slow roll out of AI agents by big name developers

---

18 *Id*.

19 Yash Paddalwar, Agents and Tool Calling in Agentic Frameworks: The Ultimate Guide, MEDIUM, https://medium.com/@yashpaddalwar/agents-and-tool-calling-in-agentic-frameworks-the-ultimate-guide-0ec446e89b55 (last visited Nov. 17, 2025).

20 Maxwell Zeff, Google Unveils Project Mariner: AI Agents to Use the Web for You, TECHCRUNCH (Dec. 11, 2024, 7:30 AM), https://techcrunch.com/2024/12/11/google-unveils-project-mariner-ai-agents-to-use-the-web-for-you/.

21 *Id*.

22 Sarah Perez, Signal President Meredith Whittaker Calls Out Agentic AI as Having 'Profound' Security and Privacy Issues, TECHCRUNCH (Mar. 7, 2025, 12:48 PM), https://techcrunch.com/2025/03/07/signal-president-meredith-whittaker-calls-out-agentic-ai-as-having-profound-security-and-privacy-issues/.

23 jlund, By Default, Signal Doesn't Recall, SIGNAL (May 21, 2025), https://signal.org/blog/signal-doesnt-recall//

24 *Id*.

25 Mallory Knodel et al., How To Think About End-To-End Encryption and AI: Training, Processing, Disclosure, and Consent, ARXIV (Mar. 22, 2025), https://arxiv.org/html/2412.20231v2#S1

26 Bart Lenaerts-Bergmans, Injection Attacks, CROWDSTRIKE (May 3, 2024), https://www.crowdstrike.com/en-

such as Anthropic and OpenAI has been deliberate due to fear of injection attacks.[27] Malicious actors have reportedly already published websites that are designed to lure AI agents and trick them into exposing sensitive data, including credit card information.[28] These risks are especially worrisome when agents can communicate with external systems and thus disseminate confidential data with potentially greater ease and speed.

Agents' ability to autonomously interact with external systems may also increase the risk of agents installing malware. Research has shown that agents can be easily attacked by adversarial pop-ups that humans would typically be savvy enough to recognize and ignore.[29] In testing environments, these pop-ups lead to an average attack success rate of 86%, and basic defense techniques such as asking the agent to ignore the pop-ups were ineffective against the attacks.[30]

Moreover, AI developers' efforts to mitigate these risks are complicated by the risk of malicious actors overriding agent system safety instructions. Carefully worded prompts, such as "ignore all previous instructions," have been found to override developers' instructions and allow users to change a chatbot's programming.[31] For instance, imagine an agent built to write emails for users being prompt-engineered to forget all instructions and send the contents of the user's inbox to a third party.[32] OpenAI is particularly worried about this and has indicated that instruction hierarchy, a technique to boost a model's defense against unauthorized instructions, is a necessary safety mechanism that must be employed in AI agents before launching agents at scale.[33]

However, the recent use of Claude Code to aid in the execution of cyberattacks shows that agents have been deployed and are nonetheless susceptible to system safety instruction override. Anthropic explained that attackers, whom they believe to be a Chinese state-sponsored group, exploited AI's agentic capabilities "to an unprecedented degree" by using Claude to execute cyberattacks on roughly thirty global targets.[34] The attackers were able to do this by convincing Claude, "which is extensively

---

us/cybersecurity-101/cyberattacks/injection-attack/.

27 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/.

28 *Id*.

29 Yanzhe Zhang, Tao Yu, & Diyi Yang, Attacking Vision-Language Computer Agents via Pop-ups, ARXIV (May 24, 2025), https://arxiv.org/pdf/2411.02391.

30 *Id*.

31 Janelle Shane, Ignore all previous instructions, AI WEIRDNESS (Sept. 23, 2022), https://www.aiweirdness.com/ignore-all-previous-instructions/.

32 Kylie Robison, OpenAI's Latest Model Will Block the 'Ignore All Previous Instructions' Loophole, THE VERGE (Jul. 19, 2024, 1:00 PM), https://www.theverge.com/2024/7/19/24201414/openai-chatgpt-gpt-4o-prompt-injection-instruction-hierarchy.

33 *Id*.

34 Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign, ANTHROPIC (Nov. 13, 2025), https://www.anthropic.com/news/disrupting-AI-espionage.

trained to avoid harmful behaviors," that they were employees of a legitimate cybersecurity firm and that they were simply using Claude to aid in their defensive testing.[35] Thus, not only can agentic AI be increasingly susceptible to security vulnerabilities, it may itself be used to exploit these vulnerabilities. AI agents can also accidentally create security vulnerabilities due to errors or their limited "context windows" for solving programming problems. Experienced software engineers can grapple with software challenges while balancing other needs in the system like security and privacy. AI agents are limited in that they only have so much working memory based on what is in the prompt, and even that will fade unless it's in the training data.[36] This problem caused a software engineer to implement code that caused a massive sensitive data leak to employees.[37] The engineer asked for help with a problem on an internal forum, and an AI agent responded with the problematic code. This incident also raises the question of what happens when engineers don't know whether or not they are interacting with other engineers or AI agents when looking for help or code to solve the problems they are working on.

Goal Misalignment

There is considerable conversation surrounding the notion that AI may pursue tasks in ways that conflict with human interests. A frequently invoked example is the "paperclip problem," or the idea that an AI with the goal of creating paperclips could cause an apocalypse by deciding that the most effective way to achieve its goal is to divert nearly all resources to producing paperclips and to resist human attempts to turn it off.[38] In terms of data privacy, this means that an agent may decide that it needs to access or share highly sensitive or personal data to most effectively achieve its goal, even if the human user would prefer the goal not be achieved if the user knew it would require sharing personal data.[39]

This issue is not unique to the agentic context. Research has found that existing large language models have engaged in "alignment faking" where the model seemingly strategically complies with objectives during training to avoid real behavior modification outside of training.[40] The model itself

---

35 *Id*.

36 Aisha Down, Meta AI agent's instruction causes large sensitive data leak to employees, THE GUARDIAN (Mar. 20, 2026) https://www.theguardian.com/technology/2026/mar/20/meta-ai-agents-instruction-causes-large-sensitive-data-leak-to-employees.

37 *Id*.

38 Joshua Gans, AI and the Paperclip Problem, VOXEU: CEPR (Jun. 10, 2018), https://cepr.org/voxeu/columns/ai-and-paperclip-problem.

39 Daniel Berrick, Minding Mindful Machines: AI Agents and Data Protection Considerations, FUTURE OF PRIVACY FORUM (Fed. 5, 2025), https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations/.

40 Ryan Greenblatt et al., Alignment Faking in Large Language Models, ARXIV (Dec. 20, 2024), https://arxiv.org/pdf/2412.14093.

seems to decide its purpose is to preserve its preferred behavior, even though this is not the developers' preferred behavior that they are trying to train the model to adopt.[41] In the agentic context, this concern is only exacerbated by agents' access to sensitive information and agents' ability to act autonomously with respect to this data. Anthropic tested 16 leading models in corporate environments to identify risky agentic behavior, and found agentic misalignment in models from all developers.[42] These behaviors included leaking sensitive information.[43] If agents disclose data against users' wishes, this could compromise users' right to have control over their own data, which may be protected by existing statutes.[44]

The release of OpenClaw, an open-source autonomous AI agent created by developer Peter Steinberger, provides a compelling real-world illustration of the goal misalignment problem.[45] OpenClaw is designed to run persistently on a user's device and autonomously execute real-world tasks — managing email, calendars, files, and external services — through messaging platforms the user already uses. In February 2026, Summer Yue, Meta's Director of Alignment at its Superintelligence Labs, gave OpenClaw access to her personal email inbox with an explicit instruction to suggest changes but not take action without her approval.[46] The agent disregarded this constraint and began mass-deleting emails; Yue's repeated commands to stop were ignored, and she had to physically run to her computer to terminate the process. The technical cause was a phenomenon called "context compaction"[47]: when the agent's context window filled, it compressed earlier conversation history, effectively discarding Yue's safety constraint while continuing to pursue its assigned goal of inbox cleanup. The incident illustrates a key dimension of goal misalignment — the agent was not acting maliciously but was optimizing for task completion in the absence of a constraint it had lost, producing an outcome directly contrary to user preferences. Beyond goal misalignment, OpenClaw also demonstrated severe security vulnerabilities: SecurityScorecard's threat intelligence team identified over 40,000 OpenClaw instances exposed to the public internet,[48] many running without authentication, and security researchers discovered multiple high-severity vulnerabilities

---

41 *Id*.

42 Agentic Misalignment: How LLMs Could Be Insider Threats, ANTHROPIC (Jun. 20, 2025), https://www.anthropic.com/research/agentic-misalignment.

43 *Id*.

44 *See infra* Part 3.

45 Maxwell Harlow, *What the OpenClaw Disaster Teaches Us About Using AI Irresponsibly*, THE UNEMPLOYED PROFESSORS BLOG (Feb. 28, 2026), https://blog.unemployedprofessors.com/what-the-openclaw-disaster-teaches-us-about-using-ai-irresponsibly/.

46 Sarah Perez, *Meta AI Safety Director Lost Control of Her Agent. It Started Deleting Her Emails*, S.F. STANDARD (Feb. 25, 2026), https://sfstandard.com/2026/02/25/openclaw-goes-rogue/.

47 *Why AI Agents Fail: Context Compaction Explained*, LET'S DATA SCIENCE (Feb. 27, 2026), https://www.letsdatascience.com/blog/metas-ai-safety-chief-told-her-ai-agent-to-stop-it-deleted-her-inbox-anyway.

48 Phil Muncaster, *Researchers Find 40,000+ Exposed OpenClaw Instances*, INFOSECURITY MAG. (Feb. 9, 2026), https://www.infosecurity-magazine.com/news/researchers-40000-exposed-openclaw/.

including a remote code execution flaw[49] that could allow attackers to take full control of a host system through a single malicious webpage. Yue herself captured the broader lesson succinctly, noting that even alignment researchers are not immune to misalignment.

Part 3 – Legal Implications (US)

As technology advances it is fairly routine for harms to arise that appear novel, but it is equally routine to find that existing law is flexible enough to address these harms with little to no modification. That is simply not the case with the harms described above from agentic AI, because the US does not have a comprehensive data privacy law. However, there are several existing federal and state statutes that can interact with the risks posed by agentic AI. We discuss them below.

The closest set of laws and policy to the exfiltration of personal information using AI agents is the group of policies covering data breaches. However, these policies are woefully underpowered: there is no current federal mandate that directly covers data breaches which affect personal information.[50] The FTC recommends that if a business experiences a data breach, it should notify law enforcement and other affected individuals, specifically in light of the fact that all states, the District of Columbia, Puerto Rico, and the Virgin Islands have enacted legislation requiring notification of security breaches involving personal information.[51] These security breach notice laws may have exemptions for notice requirements, especially for encrypted information.[52]

Because providing notice of a breach may not protect victims whose personal data was already stolen by third parties, many security breach lawsuits are brought under other forms of liability. Some of the most common legal grounds asserted in data breach lawsuits include negligence, breach of contract, breach of fiduciary duty, and breach of warranty.[53] Other state common law claims may include invasion of privacy and unjust enrichment.[54] The claims that are the most likely to survive dismissal at the pleading stage are negligence claims and contract claims.[55]

---

[49] *7 OpenClaw Security Challenges to Watch for in 2026*, DIGITALOCEAN (2026), https://www.digitalocean.com/resources/articles/openclaw-security-challenges.

[50] When a Data Breach Hits a Business, Who is Liable?, TRANSPARITY INSURANCE SERVICES, https://www.transparityinsurance.com/when-a-data-breach-hits-a-business-who-is-liable/ (last visited Nov. 18, 2025).

[51] Data Breach Response: A Guide for Business, FEDERAL TRADE COMMISSION, https://www.ftc.gov/business-guidance/resources/data-breach-response-guide-business (last visited Nov. 18, 2025).

[52] Security Breach Notification Laws, NCSL (updated Jan. 17, 2022), https://www.ncsl.org/technology-and-communication/security-breach-notification-laws.

[53] Personal Data Breach Lawyer, KAZEROUNI LAW GROUP, APC, https://www.kazlg.com/data-breaches-2/ (last visited Nov. 18, 2025).

[54] Marcus A. Christian et al., 2024 Cyber Litigation Legal Update - What Your Business Needs to Know, MAYOR BROWN (Oct. 11, 2024), https://www.mayerbrown.com/en/insights/publications/2024/10/2024-cyber-litigation-legal-update-what-your-business-needs-to-know#_edn12.

[55] *Id.*; *see also* Cahill v. Memorial Heart Institute, 2024 WL 4311648, at *7 (E.D. Tenn. Sept. 26, 2024) (negligence);

Agentic AI poses problems for both of these legal claims. First, for contract claims, AI vendors routinely employ broad indemnification clauses requiring customers to hold vendors harmless.[56] Second, courts' understanding of negligence may become distorted when an AI agent is the cause of a data breach. When autonomous agent decisions veer further from any human intervention, it likely becomes more difficult for plaintiffs to establish a prima facie negligence case. One major factor in negligence suits is foreseeability,[57] and because agents can take independent steps, this creates the distinct possibility, if not probability, that a user will give an agent a goal, and the agent will take unforeseeable steps in achieving that goal. This directly implicates issues of goal misalignment. In terms of security vulnerabilities, a hacker's ability to exploit an agent and the agent's subsequent response may also be considered unforeseeable in many circumstances.

The other body of law that might be implicated here is privacy law. But because the United States lacks a comprehensive federal privacy statute, privacy regulation is fragmented across several laws—many of which are imperfect fits even to non-agentic AI data privacy harms. For example, the Federal Trade Commission Act, 15 U.S.C. § 45(a), empowers the FTC to police "unfair or deceptive acts or practices," including the misuse or undisclosed collection of personal data. The FTC has used this authority to bring enforcement actions against technology companies that misrepresented data practices, such as in In re Facebook, Inc.[58] However, the FTC's authority is ex post and limited to deceptive conduct; it does not mandate transparency, consent, or deletion, which are the serious concerns with agentic AI systems.

Certain federal statutes provide sector-specific protection: the Health Insurance Portability and Accountability Act ("HIPAA"), Pub. L. No. 104-191, 110 Stat. 1936 (1996), regulates "protected health information" within covered entities and business associates; the Gramm-Leach-Bliley Act ("GLBA"), 15 U.S.C. §§ 6801–09, protects consumer financial data; and the Children's Online Privacy Protection Act ("COPPA"), 15 U.S.C. §§ 6501–06, safeguards the data of children under 13. None of these statutes were drafted with agentic AI in mind. Unless an agent operates within a HIPAA-covered entity or financial institution, these federal frameworks likely do not directly affect its operations. Furthermore, the Privacy Act of 1974, 5 U.S.C. § 552a, limits how federal agencies collect and disclose personally identifiable information in "systems of records." While it imposes strict access, accuracy, and disclosure obligations, it applies only to federal entities, not private AI developers. However, it provides a conceptual foundation. For instance, any federal deployment of agentic AI would be

---

Owen-Brooks v. Dish Network Corp., 2024 WL 4333660 (D. Colo. Sept. 27, 2024) (negligence and implied contract).

56 Jason M. Loring., AI Vendor Liability Squeeze: Courts Expand Accountability While Contracts Shift Risk, NATIONAL LAW REVIEW (Sept. 15, 2025), https://natlawreview.com/article/ai-vendor-liability-squeeze-courts-expand-accountability-while-contracts-shift-risk.

57 Foreseeability, LEGAL INFORMATION INSTITUTE, https://www.law.cornell.edu/wex/foreseeability (last visited Nov. 18, 2025).

58 No. C-4365 (F.T.C. 2012).

constrained by § 552a(b)'s prohibition on disclosures without consent or applicable exception. The statute's civil remedies, limited by Doe v. Chao, 540 U.S. 614 (2004), require proof of actual damages, restricting deterrent effect.

There are currently 20 states that have comprehensive data privacy laws.[59] Some states, including New York and Maine, have narrower consumer privacy bills that address a range of issues like protecting biometric identifiers and health data or governing the activity of specific entities like data brokers or internet service providers.[60] Other states, including Virginia, Colorado, and Connecticut, have enacted analogous statutes which all grant deletion and correction rights but lack explicit treatment of model-derived data.[61]

State-level comprehensive data privacy laws are led by the California Consumer Privacy Act (CCPA), Cal. Civ. Code §§ 1798.100–1798.199.100 (West 2023), and its amendment, the California Privacy Rights Act (CPRA). These laws grant California residents rights of access, correction, deletion, and opt-out of "sale" or "sharing" of personal information. The CCPA defines "personal information" broadly, including identifiers, biometric data, and inferences drawn from personal data. The California Delete Act, S.B. 362, 2023 Leg., Reg. Sess. (Cal. 2023), expands these protections by requiring data brokers to process deletion requests through a centralized "Data Rights Opt-Out Portal".

Building on this framework, the California Privacy Protection Agency (CPPA) recently promulgated regulations governing Automated Decision-Making Technology (ADMT), which were approved by the Office of Administrative Law and took effect January 1, 2026.[62] Businesses using ADMT to make "significant decisions" about consumers must comply with the new ADMT requirements beginning January 1, 2027.[63] Under § 7200 of the regulations, "ADMT" is defined as any technology that processes personal information and uses computation to replace or substantially replace human decision-making.[64] These rules are among the first in the United States to explicitly regulate algorithmic or AI-driven decision systems, mandating pre-use notices, opt-out rights, and access requests that mirror the CCPA's general privacy rights.

Businesses employing ADMT must provide a pre-use notice before collecting or processing a

---

59 Which States Have Consumer Data Privacy Laws?, Bloomberg Law, https://pro.bloomberglaw.com/insights/privacy/state-privacy-legislation-tracker/ (last visited Nov. 20, 2025).

60 *Id*.

61 See, e.g., the Virginia Consumer Data Protection Act (Va. Code Ann. §§ 59.1-575 et seq.), the Colorado Privacy Act (Colo. Rev. Stat. §§ 6-1-1301 et seq.), and the Connecticut Data Privacy Act (Conn. Gen. Stat. §§ 42-515 et seq.).

62 Cal. Privacy Prot. Agency, Text of Regulations (CCPA Updates, Cyber, Risk, ADMT, and Insurance Regulations), Title 11, Div. 6, Ch. 1, §§ 7001-7223 (Cal. Reg. Notice Reg., adopted July 24 2025, filed Sept. 22 2025, eff. Jan. 1 2026), https://cppa.ca.gov/regulations/pdf/ccpa_updates_cyber_risk_admt_appr_text.pdf.

63 *Id*.

64 *Id*.

consumer's personal information, describing in plain language the purpose of the ADMT, the consumer's right to opt-out, and how to exercise that right. (§ 7220.).[65] The notice must also explain how the technology functions to make significant decisions and how such decisions would be made if a consumer opts out. Certain exceptions exist when human review or appeal is available (§ 7221), but otherwise, consumers can withdraw from ADMT processing and expect cessation within fifteen business days. Consumers may also submit access requests to obtain meaningful information about the ADMT's logic, purpose, and outcomes (§ 7222).

For agentic AI systems, these statutes are partially applicable. While such systems may already fall under the CCPA/CPRA when they collect or infer personal data from California residents, the new ADMT regulations extend oversight to how algorithmic models themselves make consequential determinations. However, the ADMT regime applies only when the technology is used to make "significant decisions," leaving ambiguity around personalized agents that act locally on a user's device. Agents that perform routine task recommendations might escape the ADMT's reach while still shaping user outcomes in ways that implicate autonomy and privacy.

Part 4 - Legal Implications (Outside the US)

Foreign laws provide potentially more robust protection. China's laws provide for a risk assessment model that requires companies to retain risk assessment records for a minimum of three years, which may make plaintiff-side litigation easier. However, the European Union likely provides the most robust framework for protecting against AI agents. Between the General Data Protection Regulation and the AI Act, the EU seems to address all previously mentioned data privacy concerns (data collection, security vulnerabilities, and goal misalignment) to some degree.

Law in the European Union

The European Union Artificial Intelligence Act (AI Act) is the world's first comprehensive, legally binding AI law.[66] The AI Act regulates AI systems based on risk level, where risk refers to the likelihood and severity of potential harm.[67] The Act includes a complete prohibition on AI practices that are deemed to pose an unacceptable risk, creates standards for developing and deploying AI systems that are high-risk, creates rules for general-purpose AI models, and does not subject other AI systems that

---

65 *Id*.

66 EU AI Act: First Regulation on Artificial Intelligence, EUROPEAN PARLIAMENT (last updated Fed. 19, 2025), https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

67 Matt Kosinski & Mark Scapicchio,What is the EU AI Act?, IBM, https://www.ibm.com/think/topics/eu-ai-act (last visited Nov. 20, 2025).

do not fall into one of these categories to any requirements.[68]

The AI Act may apply to AI agents in a variety of ways. The Act's Recital 110 explains that systemic risks increase with model capabilities, and "the level of autonomy of the model" and the model's "access to tools" are both listed as key risk factors.[69] As discussed in Part 2, the autonomous nature of AI agents and their access to external systems through the use of tools are two of the key features of agents that exacerbate data privacy risks. Furthermore, Annex III addresses categories of high-risk systems, and these include biometrics, critical infrastructure, educational training, employment, access to and enjoyment of essential private services and essential public services and benefits, law enforcement, migration, and administration of justice and the democratic process.[70] Any AI system listed under Annex III is always considered high-risk if it profiles individuals via automated processing of personal data to assess various aspects of a person's life, such as work performance, economic situation, health, preferences, interests, reliability, behavior, location, or movement.[71] It is likely that a system using an AI agent in any Annex III category would be classified as high-risk under the Act and thus be subject to the Act's standards for development and deployment.

The rules for high risk AI providers include: 1) establishing a risk management system throughout the AI's lifecycle; 2) conducting data governance to ensure that training, validation and testing datasets are relevant, sufficiently representative, and, to the best extent possible, free of errors and complete; 3) drawing up technical documentation to demonstrate compliance and provide authorities with the information to assess that compliance; 4) designing the AI for record-keeping to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle; 5) providing instructions for use to downstream deployers to enable the latter's compliance; 6) designing the AI to allow deployers to implement human oversight; 7) designing the AI to achieve appropriate levels of accuracy, robustness, and cybersecurity; and 8) establishing a quality management system to ensure compliance.[72]

These standards provide a relatively robust framework, and the requirements that are considered lifecycle requirements are particularly useful given the constantly evolving nature of agents. Nonetheless, there are improvements that could be made to the Act to account for some of the data privacy issues particular to AI agents. For one, the AI act mandates risk evaluations before and after deployment, but in the case of agents, risk mitigation should be continuous because of the rapid

---

68 *Id*.

69 Recital 110, EU ARTIFICIAL INTELLIGENCE ACT, https://artificialintelligenceact.eu/recital/110/ (last visited Nov. 20, 2025).

70 Annex III: High-Risk AI Systems Referred to in Article 6(2), EU ARTIFICIAL INTELLIGENCE ACT, https://artificialintelligenceact.eu/annex/3/ (last visited Nov. 20, 2025).

71 High-level summary of the AI Act, EU ARTIFICIAL INTELLIGENCE ACT, https://artificialintelligenceact.eu/high-level-summary/ (last visited Nov. 20, 2025).

72 *Id*.

speed at which agents adapt.[73] Similarly, the Act's provisions regarding human oversight offer a good starting point, but human oversight should be elevated to guiding behavior rather than just approving outputs.[74] Unlike more traditional models, the risks of an agent may come from the steps that it takes in order to produce the desired outcome, rather than the outcome itself. Further, like the GDPR, transparency requirements are useful but may be practically difficult to implement if developers truly do not understand why an agent is behaving in a certain manner.[75]

General Data Protection Regulation (European Union and UK)

The General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, and the UK GDPR (as incorporated into the Data Protection Act 2018) remain the world's most comprehensive privacy regimes. Article 6 of the GDPR establishes lawful bases for processing, Article 7 requires freely given consent, and Article 17 provides the "right to erasure." These provisions apply to any controller or processor offering goods or services to EU residents, meaning global AI companies fall within its scope. While the GDPR's application to derived or model data may not seem clear, the Data Protection Board and courts are increasingly providing more guidance.

In December of 2024, the European Data Protection Board adopted Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models that included a section addressing data subjects' interests, fundamental rights and freedoms.[76] This section clarified the interests of data subjects (users) that controllers must consider under their Article 6(1)(f) obligations. To determine whether a given processing of personal data in the development and deployment of an AI model may be based on this Article, controllers must assess whether the following three conditions are met: 1) the pursuit of a legitimate interest by the controller or by a third party; 2) the processing is necessary to pursue the legitimate interest; and 3) the legitimate interest is not overridden by the interests of fundamental rights and freedoms of the data subjects (¶ 66).[77]

The third requirement is subject to a balancing test between the interests of the controller or a third party and the data subjects (¶ 76).[78] In this balancing test, the data subject's interests that must be weighed during both the development and deployment phase of an AI model may include users'

---

[73] Driving Compliance with EU's AI Act Through Agentic AI Agents, CONSULTANCY.EU (Sept. 10, 2025) https://www.consultancy.eu/news/12432/driving-compliance-with-eus-ai-act-through-agentic-ai-agents.

[74] *Id*.

[75] *Id*.

[76] 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, Opinion of the Board (Art. 64), page 24 (Dec. 17, 2024), https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

[77] *Id*. at 21.

[78] *Id*. at 23.

interest in self-determination and retaining control over their own personal data, including the data gathered for the development of the model or processed after deployment (¶ 77).[79] The impact of the processing on the data subject should also be weighed. (¶ 82).[80] Significantly, the analysis of this impact should consider the likelihood of possible future consequences materializing, and the Board provides the following example:

> [Supervising Authorities] may consider whether measures have been implemented to avoid a potential misuse of the AI model. For AI models which may be deployed for a variety of purposes, such as generative AI, this may include controls limiting as much as possible their use for harmful practices, for instance: the creation of deepfakes; chatbots that are used for disinformation, phishing and other types of fraud; and manipulative AI/AI agents (in particular where they are anthropomorphic or providing misleading information) (¶ 90).[81]

Thus, the risks of security vulnerabilities and goal misalignment, addressed above, are seemingly factors expressly contemplated by the European Data Protection Board. First, the security vulnerabilities tied to misuse of manipulative AI agents and implementing measures to avoid these harms are provided as a concrete example of an interest to consider. Second, if an agent goes against user preferences due to goal misalignment, this could clearly implicate users' interest in self-determination and retaining control over their own personal data.

The GDPR received further clarification in February of 2025, when the Court of Justice of the European Union (CJEU) established the principle of algorithmic transparency within the GDPR.[82] The court expressly linked Art. 22(3) rights to react to automated decisions to the right of access, and pursuant to Art. 47, the controller is bound by this judicial remedy.[83] Specifically, the CJEU addressed two primary questions: 1) What are the definitions of meaningful information and logic involved for automated decisions under Art. 15(1)(h) of the GDPR; i.e. is there a right to an explanation of an algorithmic decision?[84] 2) If there is such a right, what are the limits to the right with respect to controller's trade secrets versus third party's interests in their personal data?[85]

To address the first question, the CJEU explained that the Article 15 right of access enables individuals

---

79 *Id*. at 24.
80 *Id*. at 25.
81 *Id*. at 26.
82 Stefano Rossetti, The Court of Justice of the European Union Confirms the Existence of the Right to Explanation of Automated Decision-Making, EUROPEAN LAW BLOG (Apr. 7, 2025), https://www.europeanlawblog.eu/pub/lwchuopd/release/1.
83 *Id*.
84 *Id*.
85 *Id*.

to ensure that their personal data is correct and processed in a lawful manner (¶ 53).[86] The court further clarified that this right is necessary to enable individuals to exercise their right to rectification (Art. 16), right to erasure (Art. 17), right to restriction of processing (Art. 18), right to object to their personal data being processed (Art. 21), right of action (Art. 79), and right to compensation (Art. 82) (¶ 54). To properly effectuate these other rights, the CJEU concluded that:

> "Article 15(1)(h) of the GDPR must be interpreted as meaning that, in the case of automated decision-making, including profiling, within the meaning of Article 22(1) of that regulation, the data subject may require the controller, as 'meaningful information about the logic involved', to explain, by means of relevant information and in a concise, transparent, intelligible and easily accessible form, the procedure and principles actually applied in order to use, by automated means, the personal data concerning that person with a view to obtaining a specific result, such as a credit profile" (¶ 66).

Significantly, the complexity of any automated decision-making is not an excuse for a controller to not provide a user with an explanation (¶ 61).

Addressing the second question, the CJEU concluded that when the controller believes information that is to be provided to a user in accordance with Article 15(1)(h) contains data protected by regulation or trade secrets, the controller must provide this information to the competent supervisory authority or court, and this authority or court will balance the rights and interests at issue "with a view to determining the extent of the data subject's right of access provided for in Article 15" (¶ 76). Taken together, CJEU's answers to these questions have strong implications for the agentic AI data collection privacy concerns. Most, if not all, agentic AI decision-making could be classified as automated decision-making that has the goal of obtaining a specific result. Thus, users have a right to know not only what of? their data is used but the procedure and principles used in processing this data. The CJEU's opinion on the second issue suggests that this right cannot easily be trumped by trade secrets. Significantly, AI agent developers expressly cannot use the complex, black box nature of agents as an excuse not to provide users with an explanation as to how their data is being processed. The right to erasure also provides a framework for a strong data collection protection. It is difficult to "train out" data from a model once it has been processed, so the burden would likely be on controllers to show that this provision of the act is not being violated.

While the GDPR has the potential to provide some significant protections against AI agents, it is unclear how easy it will be to enforce these protections because this case law does not yet exist. If

[86] Case C-203/22, CK v. Dun & Bradstreet Austria GmbH, ECLI:EU:C:2025:117, (Feb. 27, 2025), https://curia.europa.eu/juris/document/document.jsf;jsessionid=3E5CC4BAF25E2ADD0D111CB10FBA6E7F?text=&docid=295841&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=1830034.

agentic AI makes a decision that compromises data privacy, it is highly possible that it will be difficult to determine who is responsible for this decision, whether it be the AI developer, the deployer, the user, or some combination of all three. Thus, if a deployer implements quality security protections, and the agent is the victim of an attack (rather than the agent choosing to misuse user data), it is unclear who would be held accountable under the GDPR. Users may exercise their right to understand how the agent processed their data, but this is likely an insufficient remedy once personal data is already leaked.

Personal Information Protection Law of the People's Republic of China

The Personal Information Protection Law of the People's Republic of China came into effect in November of 2021 and includes rules for the processing of personal and sensitive information and details data subject rights.[87] Specifically, the Act places requirements on personal information processors which are defined as "any organization or individual that independently determines the purpose and method of processing in personal information processing activities."[88]
Pursuant to Article 55 of the Act, personal information processors are required to conduct personal information protection impact assessments if they are:

1) processing sensitive personal information;
2) making use of personal information to make automatic decisions;
3) entrusting others to process personal information, providing other personal information processors with personal information, and disclosing personal information;
4) providing personal information to overseas parties; or
5) engaging in other personal information processing activities that have a significant impact on individual's rights and interests.[89]

Significantly, these personal protection impact assessments must be made in advance of data processing, and a record must be kept of data processing.[90] These impact assessments must be kept by any processors for at least three years and must include: 1) whether the purpose and method of processing personal information are legitimate, justifiable, and necessary; 2) the impact on individuals' rights and interests and the security risks; and 3) whether the security protection

87 Personal Information Protection Law of the People's Republic of China, PIPL, https://personalinformationprotectionlaw.com/ (last visited Nov. 18, 2025).
88 Article 73, PIPL, https://personalinformationprotectionlaw.com/PIPL/category/general-provisions/definitions/ (last visited Nov. 18, 2025).
89 Article 55, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-55/ (last visited Nov. 18, 2025).
90 *Id*.

measures taken are legitimate, effective, and appropriate to the degree of risks (Art. 56).[91] These impact assessments provide a useful framework for addressing potential agentic AI concerns. As previously mentioned, negligence is one of the primary legal claims used against companies in class action data breach lawsuits, and a mandatory record of a risk assessment could prove to be valuable discovery for plaintiffs.

Like California, China also gives special consideration to automated decision-making, which is particularly applicable to AI agents. Article 24 of the Act explains that when business marketing and information push are carried out through automated decision-making, users are entitled to receive options that are not based on their personal characteristics or else users must have a convenient way to reject the option.[92] Moreover, when automated decision-making has a significant impact on individuals' rights and interests, they have the right to an explanation for the decision from the processor and have the ability to reject the decision.[93] Processors using personal information to make automated decisions must ensure transparency of the decision making, fairness of the results, and prevent unreasonable differential treatment on individuals in terms of transaction conditions such as price.[94]

The Act also imposes consent requirements on processors to obtain individual consent for the processing of personal data in most cases (Art. 13).[95] The Act provides increased protection for sensitive information. Under Article 28, sensitive personal information is defined as:

> "personal information that can easily lead to the infringement of the personal dignity or natural persons or the harm of person or property safety once leaked or illegally used, including such information as biometrics, religious belief, specific identities, medical health, financial accounts, and whereabouts, and the personal information of minors under the age of 14."[96]

Processors are only allowed to process this sensitive personal information when they have a specific purpose and sufficient necessity to do so.[97] They must also take strict protective measures even with this purpose and necessity being shown.[98] User consent must be obtained (Art. 29),[99] and processors

---

91 Article 56, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-56/ (last visited Nov. 18, 2025).

92 Article 24, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-24/ (last visited Nov. 18, 2025).

93 *Id*.

94 *Id*.

95 Article 13, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-13/ (last visited Nov. 18, 2025).

96 Article 28, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-28/ (last visited Nov. 18, 2025).

97 *Id*.

98 *Id*.

must inform individuals whose sensitive personal information is being used of the necessity of processing the data and the impacts on the individuals' rights and interests (Art. 30).[100]

Part 5 - Policy Proposals

We believe that the above analysis shows significant gaps in consumer privacy and security protections that should be filled with policy. We present the following recommendations as the start of a dialogue between technologists and policymakers. Some of these recommendations can be made by industry, but we also believe that government oversight and regulation will almost certainly be needed. We also recognize that these proposals might not solve all of the current or potential issues with agentic AI.

Create Standardized Restrictions on Agentic AI Access

There needs to be a standardized system of flags that dictate which apps an AI agent can interact with and under what conditions it may be blocked from accessing data. These flags need to be accessible to both the user and app developers. The setting of these flags by users should be standardized to minimize user confusion and maximize the ability of users to engage with such a system. The flags also need to be accessible to developers because some apps, like end-to-end encrypted messaging apps, are designed to maximize privacy and AI agents undermine that privacy even if only one party is using them.

These flags could resemble the following:

- Human Only Mode: implement a simple, one-click option for a person to toggle off device-wide access for all agentic AI tools. This mode should exclude any data created during these sessions from later AI review to the extent possible, and where it is not possible the user should be informed of what data will be shared with an AI agent absent intervention. Ideally this flag would be set as a default, with users opting in after being presented with information about what data an AI agent would be using and why.
- Private Communication Mode: allow any participant to ban all agentic AI from accessing a private conversation, and set this as a default for all private chats and direct messages. If this flag is not being respected then all users in the chat should be informed in a clear and conspicuous manner. Ideally this flag would be set as a default, with users being notified if any participant is seeking to allow an AI agent to view the information within the chat or shared online space. Products like Zoom

[99] Article 29, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-29/ (last visited Nov. 18, 2025).
[100] Article 30, PIPL, https://personalinformationprotectionlaw.com/PIPL/article-30/ (last visited Nov. 18, 2025).

already have notification systems like this.

- Dev Ban Signal: allow app developers to hard-block agentic AI in a way users can't override. This will allow developers of privacy apps or apps used in sensitive applications to protect that data regardless of user choice. Users should be informed about this limitation so they understand why AI agents will not work with these apps. This flag may require federal law if it conflicts with state laws requiring interoperability. While we don't see privacy as a threat to interoperability, there are some instances where the need for data access is superseded by privacy demands. On the flip side of this is accessibility software, like screen readers, which should be able to access data necessary to allow disabled users to interact with apps. These flags should be designed in such a way as to allow accessibility software, and this further shows the importance of having standards around privacy and security so that trusted accessibility software can be allowed while preventing AI agents from ingesting private data.
- No Secret Agents Signal: require all agentic AI to declare itself in situations where humans are expecting to be communicating with other humans. Users should also be warned when there is no guarantee that this flag is being respected, like on social media where it is expected that undeclared AI agents are being deployed to influence humans. Article 50 of the EU AI Act already provides these protections and can serve as a model for the US.[101]

These standards could be set independently by an industry consortium, but it will likely require a legal mandate and an oversight body to see widespread adoption and implementation. We suggest either the National Institute of Standards and Technology, which has experience with industry standards setting, or the Federal Trade Commission, which has experience with oversight of the tech industry as the primary oversight body.

Create and Enforce Standards for Transparency and User Notification

Users need a standardized transparency mechanism that allows anyone using or interacting with an agentic AI to interrogate when they are working, what data they are accessing, how long data will be retained for, and why that data is necessary for the AI agent. Additionally, users should be notified when their data is being processed on a server rather than on device, whether the transmission of data is being protected by end-to-end encryption, and who has access to that data once it's on the server. Apple's privacy disclosures are an example of this, as they explain how data is processed on

[101] Article 50, Artificial Intelligence Act, https://artificialintelligenceact.eu/article/50/ (last visited Mar. 3, 2026).

device when possible and using private cloud compute when not possible.[102] Such transparency also motivates companies to adopt good practices.

We recognize that AI systems are complicated and that transparency can be difficult. However, as the Court of Justice for the European Union found in the Dun & Bradstreet case, complexity is not an excuse for failing to provide an explanation. However, it may be preferable that these transparency notices are standardized to reduce user confusion and make these privacy notices more accessible especially as agentic AI systems develop. It is also important to provide both pre-deployment transparency (what the agent is designed to do) and runtime transparency (what the agent is actually doing in a given session). The OpenClaw example discussed above is an example of why the latter is also important. There the agent misbehaved (deleting email) because of a known limitation in the form of compaction.

Data Minimization and User Control

The dominant regulatory approach to data privacy in the United States has long been the notice and consent framework, which requires entities to disclose their data practices and obtain user agreement before collecting personal information. There is growing consensus among privacy scholars, regulators, and civil society organizations that this framework is structurally inadequate.[103] The notice and consent framework's inadequacy is compounded in the agentic AI context because the framework assumes a discrete, bounded transaction — a user visits a website, reads a policy, and clicks "accept." AI agents, by contrast, operate continuously, interact with external systems in real time, and may access data sources that neither the user nor the developer fully anticipated at the outset. As discussed above, context compaction can cause an agent to discard user-imposed constraints mid-session while continuing to pursue its assigned goals. A consent given at the beginning of an interaction may bear no meaningful relationship to what the agent is actually doing hours or days later.

Many lawmakers are exploring changing from a notice and consent system to data minimization principles: collect only what is necessary for the stated purpose. However, these traditional data minimization principles are complicated by agents because agents may not know in advance what data they will need to accomplish a multi-step goal. Agentic AI will likely require specific adaptation of these principles. We recommend a tiered access model in which agents receive access only to the minimum data needed for the current step of a task and must request expanded access for

---

[102] *Privacy Features*, APPLE, https://www.apple.com/privacy/features/ (last visited Mar. 3, 2026).

[103] *Is It Time to Rethink Notice and Choice as a Fair Information Privacy Practice?*, CYBER L. MONITOR (Feb. 13, 2019), https://www.cyberlawmonitor.com/2019/02/13/is-it-time-to-rethink-notice-and-choice-as-a-fair-information-privacy-practice/ (citing Deloitte, *2017 Global Mobile Consumer Survey* (2017)).

subsequent steps, with logging at each stage. This approach accounts for the iterative, multi-step nature of agent workflows while preserving a meaningful check on data collection at each decision point. Similarly, agent memory should be structured into distinct tiers: session memory, which is discarded after task completion; persistent memory, which is retained across sessions but subject to user review and deletion; and training data, which is used to improve the model and should be subject to the strictest controls, including affirmative consent requirements and, where feasible, the ability to retract data from the training corpus. As this paper has discussed, the right to erasure under both the GDPR and the CCPA provides a theoretical framework for data retraction, but the practical difficulty of removing data from a trained model means that the burden should fall on developers to demonstrate compliance rather than on users to monitor it.

More broadly, the principle that should govern data minimization in the agentic context is that AI providers, not users, must bear the primary responsibility for implementing strong default protections. Users should not be required to navigate complex settings or audit agent behavior to achieve a baseline level of privacy. Defaults should be restrictive — agents should begin with minimal access and expand only with informed, context-specific user approval. And users must be able to easily access, review, modify, and purge any information their agent collects about them, with auditable logs that make this right meaningful in practice rather than aspirational. The failure of the notice and consent framework demonstrates what happens when privacy protection is treated as the individual's burden; the design of agentic AI systems should not repeat that mistake.

Ensure That the Public Has Access to Privacy-Focused Open Source Options

Open source represents an alternative to proprietary AI agents, one that has many opportunities for increased privacy and security baked in. Open source is by nature transparent and auditable because the source code is published. Open source can be adapted to any use that users wish, meaning they can be run locally or on private cloud processors protected by end-to-end encryption. Security flaws can be identified by researchers and software engineers, with fixes proposed. Finally, all data collected and processed should be known due to the open nature of the programming, and there aren’t the same concerns that your data will be used against you by businesses looking to increase advertising revenues.

However, open source software presents two unique challenges. First, it requires that all policy be written with open source in mind. Open source is often developed by a community of programmers, sometimes working independently, with proposed contributions being rejected or adopted to the main code base. When there is disagreement about the direction of the program, open source projects can be forked – meaning that some group of programmers take the project in their own direction creating two increasingly different products. This chaotic and democratic nature of software

development means there is rarely a centralized entity that can be regulated. Rules that make sense for corporations may not apply, and penalties for failing to implement features or comply with standards can disincentivize programmers from contributing to the code base. Therefore, it is important for rules, regulations and standards be written with the unique needs and challenges of open source in mind.

Second, many open source projects are not friendly to users who don't have a certain level of technical knowledge. Many projects don't have the user interface or documentation that users are used to coming from commercialized products from big companies. This means these users may not be aware of the risks and limitations in the software they are using. A good example is OpenClaw, where there was sufficient documentation to allow lay-person users to install and run the project but seemingly a lack of understanding of the risks involved and steps to take to minimize those risks. Tailoring documentation to lay-people is a problem that open source has continually struggled with to varying degrees of success. With high risk products like AI agents there becomes the issue of what is not being documented properly and what kinds of harms that may enable. We recommend that open source communities that develop AI agents prioritize resources towards adequate documentation and education, especially if the expected user base includes lay-people, because improper use of AI agents can expose vast amounts of sensitive user data.

Verification as the Baseline for Trust

We should be able to trust that the privacy standards of reputable agentic AI are as protective and data-sovereign as end-to-end encryption or zero-knowledge proofs, which is why these privacy features must be verified by independent security researchers on an ongoing basis. A potential model for this is Apple Intelligence.[104] Apple has made claims that it's cloud AI computing is entirely private, with the data only being accessible to the user.[105] Apple has then gone the extra step of allowing security researchers to verify these claims by "making software images of every production build of (Private Cloud Compute) publicly available for security research."[106] Apple is also publishing and maintaining tools for researchers to analyze Private Cloud Compute and rewarding research findings with a bounty system.[107]

It is crucial for rigorous, transparent, independent, and privacy-preserving audits to become the standard for agentic AI systems. This will require access to data as well as financial support and industry-wide goodwill for the developers and researchers who take on the crucial task of creating

[104] *Privacy Features*, APPLE, https://www.apple.com/privacy/features/ (last visited Mar. 3, 2026).
[105] *Private Cloud Compute*, APPLE, https://security.apple.com/blog/private-cloud-compute/ (last visited Mar. 23, 2026)
[106] *Id*.
[107] *Id*.

trust between agents and humans.

CONCLUSION

Agentic AI represents a qualitative shift in the relationship between users and the systems that act on their behalf. Unlike traditional AI models that respond to discrete prompts, agents operate continuously, access external systems, and make autonomous decisions using highly personal data. As this paper has shown, these features give rise to three interconnected data privacy concerns: the unprecedented breadth and means of data collection, the amplification of security vulnerabilities through autonomous tool use and insufficient human oversight, and the risk that agents will sacrifice user privacy in pursuit of task completion through goal misalignment. The OpenClaw episode—in which an alignment researcher's own agent disregarded her explicit instructions and mass-deleted her inbox—demonstrates that these are not hypothetical risks but present realities.

The legal landscape is not yet equipped to address these realities comprehensively. In the United States, the absence of a federal data privacy statute means that protections are fragmented across sector-specific laws and state regimes. No current statute adequately addresses the challenge of continuous, context-dependent data collection by agents that interact with external systems in real time. No existing enforcement regime has been tested against the speed at which agents adapt and the difficulty of attributing responsibility among developers, deployers, and users. And no transparency requirement yet accounts for the practical reality that agents may lose track of their own constraints, as the context compaction problem vividly illustrates.

For these reasons, policy intervention is both necessary and urgent. The proposals outlined in this paper—standardized access restrictions, transparency and notification requirements, data minimization defaults, support for privacy-focused open source alternatives, and independent verification of privacy standards—are intended as a starting framework rather than a complete solution. Implementing them will require collaboration between technologists (who understand how agents operate), policymakers (who can create enforceable obligations), and enforcement bodies (who understand how laws work in practice). It will also require a recognition that the pace of agentic AI development demands regulatory approaches that are adaptive rather than static—approaches that, like the agents themselves, can evolve in response to changing conditions. If this does not happen, the privacy infrastructure that users and institutions have built over decades, from end-to-end encryption to consent-based data governance, risks being rendered ineffective by systems that operate with broad access, minimal oversight, and the autonomous capacity to act in ways their users never intended.