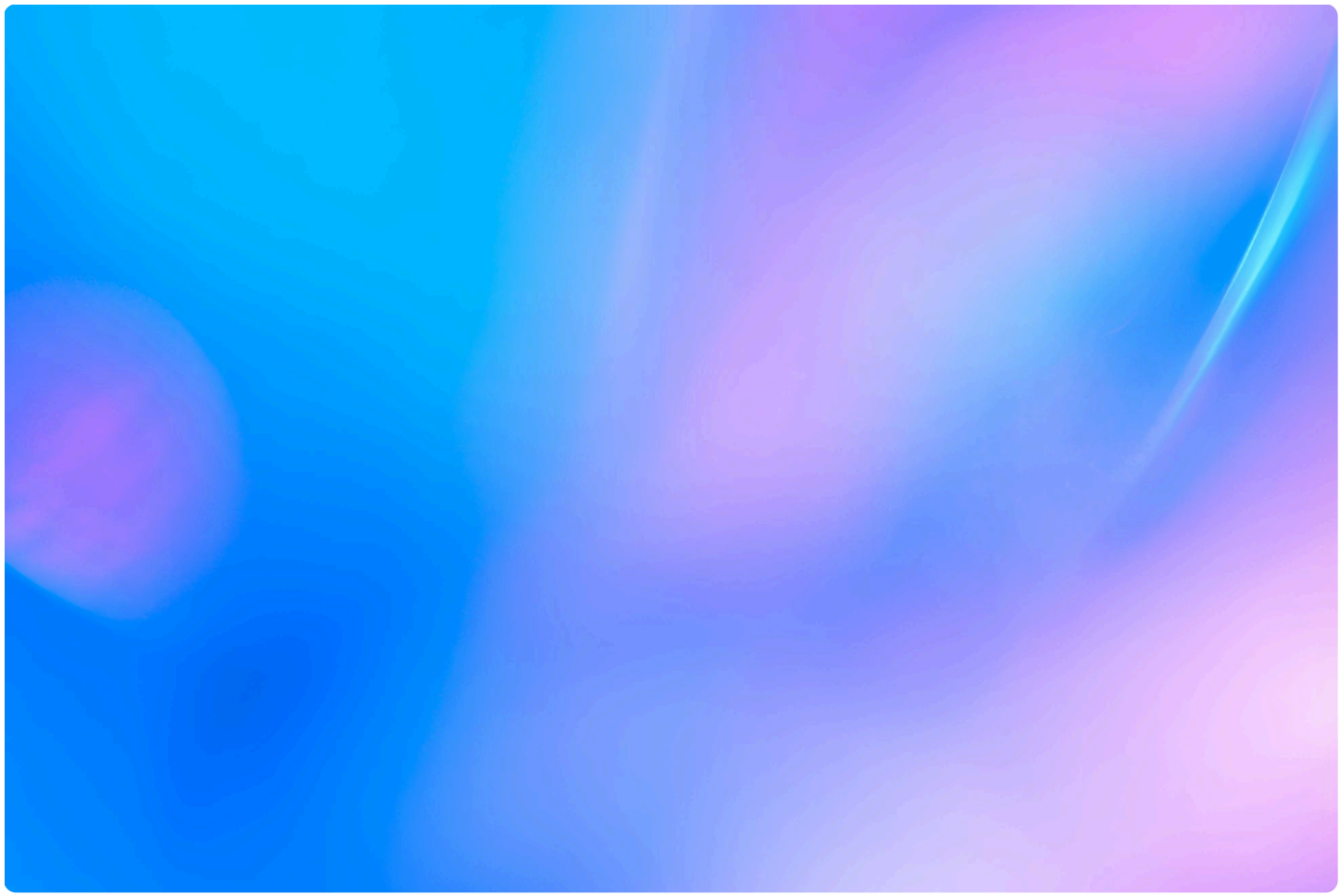
OpenAI

OpenAI privacy hackathon report



January 2026

Contents

Forward

The OpenAI Privacy Hackathon was conceived with a simple but ambitious aim: to move the conversation on privacy in AI from principle to practice.

Across Europe, privacy discourse is rightly rigorous, risk-focused, and grounded in law. Yet as AI systems become more capable, autonomous, and embedded in everyday life, there is a growing need to complement that legal foundation with practical, technical mechanisms that make privacy work in real systems, at real moments of interaction. This hackathon was designed to explore exactly that gap.

By bringing together engineers, founders, developers, privacy technologists, policymakers, academics, regulators and practitioners in a hands-on, time-bounded environment, the event tested whether the collaborative energy of a hackathon - long associated with product innovation - could be applied to one of the most complex challenges in AI: operationalizing privacy by design at scale.

Forward

What emerged over the course of the day exceeded expectations. Participants did not focus on abstract compliance or theoretical safeguards. Instead, they produced concrete proposals that address privacy at critical points in the AI lifecycle: at the user interface, within data architectures, and increasingly, at the level of autonomous, agent-mediated interactions. Many of the ideas captured in this report reflect emerging realities – agentic systems, multimodal data, inference risk - that regulators, developers, and organizations are only beginning to grapple with.

This report does not present endorsed solutions, nor does it claim to resolve the hard questions raised by advanced AI systems. Rather, it documents a moment of collective exploration: an experiment in how policy, engineering, and design communities can work together to imagine privacy-enhancing futures that are both technically plausible and normatively grounded. This report does not present endorsed solutions, nor does it claim to resolve the hard questions raised by advanced AI systems. Rather, it documents a moment of collective exploration: an experiment in how policy, engineering, and design communities can work together to imagine privacy-enhancing futures that are both technically plausible and normatively grounded.

We hope this publication serves three purposes. First, as a record of the ideas generated through the hackathon. Second, as a signal that privacy innovation deserves the same creative and technical investment as other dimensions of AI development. And third, as an invitation - to regulators, technologists, and policymakers alike - to continue building shared spaces where privacy is treated not as a constraint on innovation, but as a capability to be engineered deliberately and responsibly.

This hackathon was intended as a starting point of a journey. The conversations, prototypes, and questions it generated are ones we expect – and hope – to revisit, expand and keep the drumbeat of privacy for the intelligence age.

Emma Redmond
Associate General Counsel, Head of OpenAI Ireland

Rafaela Nicolazzi
Head of EMEA Data, Privacy & Consumers Protection, Global Affairs

Idriss Kechida
Global Data Protection Officer

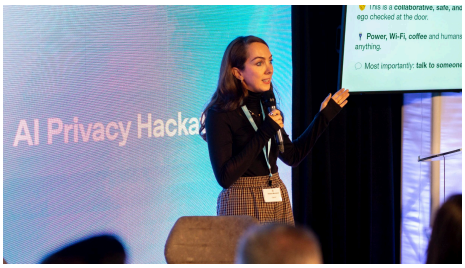
Disclaimer

This report brings together the outputs of the OpenAI Privacy Hackathon, a privacy–focused hackathon held in Dublin on 4 December 2025 and organized by OpenAI.

It reflects the discussions and recommendations developed by the 75 participants across three thematic tracks during the event and does not necessarily represent the views of the organizers or of all participants.

The content is provided for informational purposes only and does not constitute an official position or endorsement by OpenAI or any other organization. Within each of the three tracks, the policy proposals are presented in line with the content of the final group presentations of the respective teams during the hackathon. No substantive changes have been made in compiling this report; only minor edits for clarity, formatting, and presentation have been applied.

Introduction



The OpenAI Privacy Hackathon was conceived as a first-of-its-kind, AI-focused policy and technology hackathon dedicated exclusively to advancing privacy innovation globally.

Held in Dublin on 4 December 2025 and organized by OpenAI, the event brought together 63 participants from across Europe, including AI engineers, privacy technologists, policymakers, academics, and legal and technical experts.

By applying the fast-paced, solution-oriented methodology of hackathons to privacy challenges, the event aimed to bridge the persistent gap between policy frameworks and technical implementation. Rather than focusing solely on regulatory compliance, the hackathon was designed to inspire hands-on problem-solving, elevate awareness of Privacy-Enhancing Technologies (PETs), and explore how AI can be used to strengthen privacy in practice. A further objective was to seed a lasting community of European practice connecting developers, privacy professionals, and policy stakeholders.

This approach responds to a critical gap in current privacy discourse, which remains largely theoretical or legal in nature, and although appropriately centered on risk, has not yet matched that focus with a comparable effort to harness AI to meaningfully strengthen privacy. As the first AI-specific hackathon in Europe focused solely on enhancing, rather than merely complying with, privacy standards, the event provided a unique venue for cross-disciplinary collaboration to co-create practical prototypes and policy-relevant frameworks that advance ethical innovation.

Participants worked across three thematic tracks, each addressing a core dimension of privacy-by-design in AI systems:

Track 01	Privacy by design
Track 02	Data minimization and synthetic data tools
Track 03	Agentic AI & privacy-preferences interoperability

Each track worked intensively over the course of the hackathon to identify challenges and to develop actionable proposals, combining technical expertise with policy insight. This report compiles the outputs of that collective effort, presenting the ideas developed across the three tracks and reflecting the collaborative spirit that defined the OpenAI Privacy Hackathon. While diverse in scope, the proposals share a common ambition: to move privacy from abstract principle to practical, individual-centered implementation in the age of AI.

Overview of policy proposals

01	**AIngel Extension**	Browser-native AI privacy intelligence layer for real-time form analysis and disclosure awareness.
02	**Ghost**	On-device AI privacy agent that warns users about inferences and risks before posting or submitting content.
03	**Data subject rights (DSR) resolution for multimodal systems**	LLM + graph database approach to identifying, mapping, and resolving personal data across complex, multimodal systems.
04	**Precision tool permissions**	JWT- and OAuth-based permission propagation to ensure AI agents only retrieve data users are authorized to access.
05	**Blackboard / MindGuard**	On-device privacy control layer that lets users define granular, purpose-bound permissions for AI agents, with interception, blocking, or synthetic responses.
06	**TrustAI**	Agent-level enforcement of privacy preferences so that consent is respected during agent-mediated interactions.
07	**Agentic enforcement of privacy preferences via CMPs**	Making consent signals enforceable throughout the agent journey by embedding them directly into agent behavior rather than relying on voluntary website compliance.

Breakout 1: Privacy by design

Facilitator
Rodolpho Eckhardt | Privacy Engineering Manager, OpenAI

Context

As artificial intelligence systems become increasingly embedded in products, services, and decision-making processes, the need to integrate robust privacy protections directly into system design and architecture has become more pressing than ever. Privacy by Design (PbD) is a foundational principle that calls for privacy to be proactively embedded throughout the entire lifecycle of AI systems, from ideation and requirements gathering through development, deployment, and ongoing operation, rather than treated as an afterthought or a compliance exercise.

This breakout session explored how Privacy by Design principles can be meaningfully operationalized in the context of AI, with a particular focus on systems handling complex, multimodal data such as text, images, audio, and video. Participants examined how architectural choices, product decisions, design patterns, and governance mechanisms can systematically reduce privacy risks while preserving functionality, innovation, and user trust.

While tools, processes, and governance models that embed privacy directly into system design are widely recognized as central to scalable and trustworthy AI development, there is no single blueprint for what effective Privacy by Design looks like in practice. Participants were therefore invited to explore a broad range of approaches, including privacy-first system architectures, consent-aware data flows, privacy-preserving defaults, automated impact assessments, and design patterns that enforce purpose limitation and user control.

Breakout 1: Privacy by design

The session focused on identifying promising design strategies, surfacing open implementation challenges, and developing hackathon concepts that advance practical, system-level applications of Privacy by Design in AI systems.

Proposal 1: AIngel extension

Overview

Online forms are one of the largest and least monitored sources of personal data exposure on the web. Individuals routinely disclose sensitive information without real-time visibility into what they are sharing, whether the requested fields are necessary, who will receive the data, or how the disclosure relates to information they have shared previously. Existing safeguards do not address this moment of risk. Security tools are designed to protect systems, and privacy laws regulate overall organizational behavior, but neither protects the individual at the point where personal data is created and submitted. This creates a critical gap in user awareness and control precisely when privacy risks materialize.

Key insights

The primary barrier identified is the absence of real-time privacy intelligence at the moment of data entry. Users lack meaningful signals that help them understand the sensitivity of individual fields, the proportionality of data requests, or the broader context in which their information will be processed. There is also no persistent memory of past disclosures, leaving individuals unable to assess cumulative privacy exposure across sites. As a result, data entry remains a blind and fragmented activity, even though it represents a major vector for personal data leakage. The opportunity lies in embedding privacy awareness directly into the interface where data is entered, transforming abstract privacy risks into immediate, intelligible signals.

Solution

AIngel Extension is a browser-native AI system that provides real-time privacy risk assessments for online forms before data is submitted. It operates directly within the browser to analyze form fields dynamically, identifying personally identifiable, sensitive, behavioral, and financial data through AI-driven semantic analysis. Privacy risk is assessed by examining data minimization, the coherence between purpose and requested fields, domain trust, jurisdictional context, and the presence of scripts or third-party elements. These assessments are translated into clear, visual explanations that enable users to make informed decisions at the moment of disclosure. The system is implemented through a layered architecture. A local extension layer scans the page structure and classifies fields in real time, performing lightweight on-device pre-processing. An AI processing layer conducts deeper semantic analysis, identifies domains and organizations, models purpose-to-field plausibility, and generates a dynamic privacy risk score using a standardized index. A user intelligence layer presents this information through dashboards and visual narratives, including a timeline of past disclosures and cross-site risk comparisons. Privacy safeguards are built into the design through local-first extraction, privacy-preserving logs, and a zero-knowledge user vault that securely records what data was shared, where, and when.

Prototype

The prototype envisions AIngel Extension as an unobtrusive, browser-integrated interface that overlays privacy intelligence directly onto online forms. Fields are visually highlighted in real time as users interact with them, with risk indicators and simplified explanations appearing contextually. A companion dashboard presents aggregated insights through imagery, infographics, or auto-generated visual narratives, allowing users to review their disclosure history and compare privacy risks across websites. The visual design is intended to make privacy visible and actionable without interrupting normal user workflows, reinforcing informed decision-making at the point of data entry.

Proposal 2: Ghost

Overview

People, particularly younger users, routinely share text, posts, comments, and form entries online without realizing what can be inferred about them from that content. While the shared information may appear harmless, it can reveal sensitive attributes such as political views, routines, locations, family details, or other aspects of identity through inference rather than explicit disclosure. Existing privacy tools fail to address this risk because they require significant effort from users, relying on separate apps, lengthy policies, or buried settings pages that are rarely consulted. As a result, long-term risks related to reputation, personal safety, and profiling remain invisible at the moment of sharing, when intervention would be most effective.

Key insights

A central insight is that friction significantly reduces adoption: users are unlikely to engage with privacy tools that require leaving their current flow or consulting a separate interface. While fear-based warnings can be effective, they only work when paired with user autonomy and simple, in-context choices. The most effective intervention point is immediately before content is posted or submitted, rather than after the fact or within static settings pages. Another key insight is that inferred information can be more sensitive than raw data itself, as patterns in language and behavior can reveal deeply personal attributes. Finally, users exhibit different sharing behaviors, with some tending to overshare and others being more cautious, which means that privacy interventions need to adapt dynamically rather than apply uniform warnings.

Solution

The proposed solution is an on-device AI privacy agent that operates as a browser extension supported by a backend service. The agent monitors text in real time as users are about to submit forms, sign up for services, or post comments and content. This text is analyzed using an LLM-based classifier, running locally on the device or through a privacy-preserving API, to assess potential privacy risks before the content is shared.

The system performs inference detection and risk scoring by identifying explicit identifiers such as names, phone numbers, and email addresses, as well as sensitive categories, locations, routines, references to children, and other high-risk signals. It combines this content analysis with site context to generate a simple, intelligible risk level, such as low, medium, or high.

At the point of action, the agent presents a just-in-time coaching interface that appears as a small banner or bar within the user's flow. This interface explains what can be inferred from the content and highlights risky fragments directly in the text. It then offers agentic, one-tap options that preserve user autonomy, allowing the user to post anyway, edit a suggested safer version, or choose not to post.

In parallel, the agent maintains a personal privacy profile stored entirely on the device. This profile learns from the user's typical behavior, distinguishing between more cautious users and habitual oversharers, and adjusts the strength and frequency of nudges accordingly. It also maintains a private, encrypted on-device log of interactions to support this adaptive behavior without external data exposure.

Breakout 2: Data minimization & synthetic data tools

Facilitator
Charles de Bourcy | Member of Technical Staff, OpenAI

Context

As artificial intelligence systems increasingly rely on rich, real-world data to deliver high-value capabilities, protecting individuals' privacy is becoming more complex, particularly where data spans unstructured text, images, audio, and video. Data minimization is therefore a crucial area of practice in the development and operation of AI systems. Organizations must consider not only how to limit the collection and retention of data, but also how to transform or abstract it so that it remains useful while reducing privacy risk.

This breakout session examined these challenges in the context of multimodal data, including approaches to removing personally identifiable information, working with synthetic datasets, and defining practical retention and deletion strategies. Tools, processes, and governance models that operationalize PII removal, synthetic data practices, and retention minimization are likely to be central to trustworthy AI development and evaluation, yet there is no single template for how they should be implemented.

Participants were invited to explore a wide range of ideas, such as redaction or transformation pipelines, synthetic data generation methods, and workflow or policy mechanisms for enforcing data

minimization across systems and teams. The session also considered how AI itself can support these efforts by enabling automation, analysis, and control at a scale that would be difficult to achieve through manual processes alone. The focus was on surfacing promising design directions, open research questions, and hackathon concepts that advance privacy-preserving AI and responsible data minimization across modalities.

Proposal 3: Data subject rights

Problem statement

Complying with data subject rights is increasingly difficult in modern data environments where personal data is distributed across large, complex systems. This challenge is amplified in multimodal settings, where personal data exists across text, images, audio, and video. The core problem is accurately identifying which personal data belongs to which individual. Without reliable subject–data linkage, organizations struggle to respond effectively to access, deletion, or correction requests, particularly at scale.

Key insights

Participants identified that data subject rights workflows are hindered by the size and complexity of organizational data scopes. Multimodal data introduces additional challenges, as some data types are harder to process and parse than others. Disambiguating personal data between multiple individuals is a recurring difficulty, especially where references are indirect or contextual. Issues of scale, granularity, and relevance further complicate compliance, and existing detection and identification methods are often insufficient to address these challenges in a reliable or scalable way.

Solution

The proposed solution combines fine-tuned large language models with a graph database to improve data-subject identification and mapping. In this approach, data is ingested into a structured graph database that

represents relationships between individuals and data elements. A first fine-tuned LLM extracts and classifies personal data, while a second LLM traverses and reasons over the graph structure to resolve identities and link data points to the correct data subjects. The system outputs a clear, queryable mapping between individuals and their associated personal data, enabling more effective data subject rights fulfillment.

Graph databases were selected because they support faster querying over complex relationships and integrate well with LLM-based reasoning, making them suitable for large-scale, multimodal data environments.

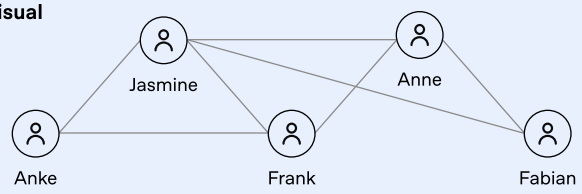
Prototype

The proposal was illustrated through both textual and visual representations of the graph-based architecture. These visuals showed how data flows from ingestion into a structured graph, how LLMs interact with the graph to extract and resolve personal data, and how the final output provides a consolidated view linking data subjects to their personal data. The visual concept emphasized clarity, traceability, and the ability to reason over complex relationships to support data subject rights processes.

Textual

Name	Works with
Anke	Frank, Jasmine
Anne	Anke, Fabian, Jasmine
Fabian	Anne, Jasmine
Frank	Anke, Anne, Jasmine
Jasmine	Fabian, Frank

Visual



Proposal 4: Precision tool permissions

Problem statement

Teams often rely on SQL experts for even basic data queries, creating bottlenecks, slowing decision-making, and increasing the risk of exposing sensitive data through over-broad queries. Organizations need a safe, proportional, and privacy-aware method to transform natural-language questions into queries that only retrieve data users are authorized to access.

Key insights

Current workflows introduce unnecessary friction and risk. Manual querying by specialists centralizes knowledge but limits speed and scalability. Over-broad or improperly scoped queries can expose sensitive information unnecessarily. There is a clear opportunity to leverage AI to automate query generation while embedding strict access controls, ensuring that users can “ask freely” but only retrieve data they are permitted to see.

Solution

The proposed solution involves propagating the user’s JSON Web Token (JWT) into the system prompt so that GPT agents generate queries on behalf of the user, constrained by their specific permissions. Data access is centrally governed through OAuth, ensuring that GPT agents only retrieve data the user is authorized to access. This allows natural-language questions to be translated into accurate queries that respect organizational permission boundaries, reducing bottlenecks and exposure risk.

Prototype

The team’s presentation included code examples from a notebook environment demonstrating JWT propagation, system-instruction handling to constrain data access, and sample outputs showing GPT agent calls enforcing user-specific permissions. The architecture enables GPT agents to safely interact with sensitive data while maintaining compliance with permission policies. The prototype emphasizes clear enforcement of access controls and the ability to handle natural-language queries at scale.

Breakout 3: Agentic AI & privacy-preferences interoperability

Facilitator
Justin B. Weiss | Privacy Consultant

Context

As artificial intelligence (AI) evolves from reactive, assistive technologies to more autonomous, agentic systems, the potential to transform user experiences across diverse platforms, service providers, and digital interfaces is significant. However, this transition introduces a fundamental challenge: enabling the portability of user privacy preferences as agentic AI systems act on users' behalf and interact with third parties that may have differing privacy policies and practices. Agentic AI platforms can perform a wide spectrum of actions, such as executing purchases, confirming shipping and personal details, or directly connecting with other users, platforms, and services, which raises critical privacy and data governance concerns that must be addressed as a condition of trustworthy use.

Tools that facilitate the portability of user privacy preferences are likely to play a pivotal role in advancing agentic AI services and their interoperability in the consumer context. Nevertheless, challenges remain regarding the availability, compatibility, and enforceability of these tools,

and their adoption at scale. Previous initiatives, including the Platform for Privacy Preferences (P3P) and Do Not Track (DNT), offer valuable lessons; although these efforts had the potential to improve user experience and safety, their limitations should inform areas for improvement as we strive toward effective privacy preference interoperability for agentic AI.

Proposal 5: Blackboard

Problem statement

AI agents often require deep access to data or services to be useful, but users do not know exactly how their data is being used. Current permission models can be crude, typically seeking and granting broad access rights such as “access your files,” rather than task-, purpose-, and time-bound permissions. As a result, data frequently leaves the user’s device, creating privacy, compliance, and trust risks. Users lack mechanisms to specify: “This agent can do these things, with this data, for this purpose, and it stays on my device.”

Key insights

Participants identified several barriers and gaps in current privacy controls. Permissions are often binary, forcing users to either over-trust agents or avoid them altogether. Existing privacy controls (cookies, mobile app settings) fail to provide meaningful granular consent or transparency regarding data use. There is no standard, machine-readable vocabulary for agent behavior covering data, purpose, retention, identity, and value, limiting interoperability. Cloud-by-default agents produce opaque data flows with weak auditability, and agents may make decisions that conflict with user intent. Questions relating to data sovereignty, such as between the EU and the US, create additional compliance uncertainty.

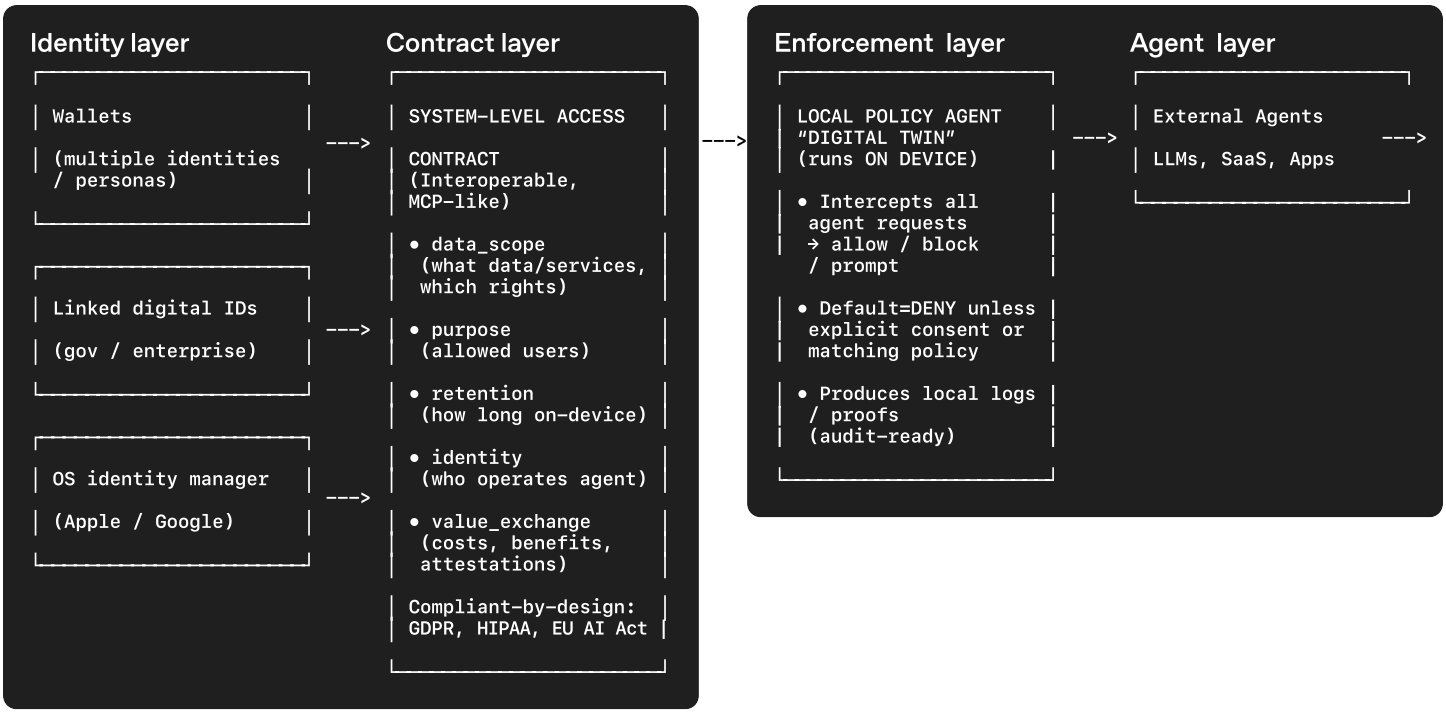
These challenges present an opportunity for an on-device control layer that is trusted, auditable, and enterprise-ready, allowing users to enforce identity-based, purpose-limited, and time-bound permissions.

Solution

MindGuard is a privacy agent that runs on the user's device, giving users control over AI agent interactions according to a structured privacy vocabulary. Users define preferences once, including which data or services an agent can access, for what purposes, retention policies, identity of the agent, and any value exchange or attestations. The agent enforces these preferences across applications, intercepting sensitive requests and either blocking them or returning synthetic neutral data while notifying the user. Input privacy ensures sensitive data never leaves the device; output privacy ensures that applications only receive controlled outputs; flow governance enforces the rules across apps, and verification dashboards show users what was blocked or allowed.

Prototype

The prototype was demonstrated as a mobile app. Users first set privacy preferences, specifying data sources, access modes, and retention rules. When a prospective agent attempts to analyse sensitive information (e.g., emotional state), MindGuard intercepts the request and either blocks it or returns synthetic data. A dashboard provides feedback, showing, for example, “This week, 47 requests blocked.” The internal schema supports structured enforcement of preferences, covering data scope, purpose, retention, identity, and value exchange, and enables fine-grained, machine-readable control over agent behaviour. Visuals included phone mockups for setup screens, interception alerts, and dashboards, illustrating real-time enforcement and transparent reporting.



Proposal 6: TrustAI — enforcing privacy preferences in agentic AI

Problem statement

As AI agents increasingly act on behalf of users, existing privacy and consent mechanisms are not designed for agent-mediated interactions. Consent today is largely managed through website-specific consent banners and settings, which are fragmented, inconsistently enforced, and difficult for users to understand or maintain over time. When agents interact with services autonomously, users lose visibility and control over how their preferences are applied, creating gaps in compliance, trust, and accountability.

Key insights

Participants identified that current privacy tools rely heavily on voluntary compliance and place the burden of configuration on users. Consent is often binary, context-blind, and disconnected across services. As a result, users must repeatedly express preferences without any guarantee that they are respected, particularly when interactions are mediated by agents rather than direct user action. The group noted that emerging wallet-based approaches to identity and to identity and consent in the EU illustrate a shift toward portable, user-controlled trust signals. However, these mechanisms alone do not solve the enforcement problem unless agents are technically constrained to act in line with user preferences.

Solution

TrustAI proposes an agent-level enforcement layer that ensures user privacy preferences are consistently applied across interactions. Instead of relying solely on websites to honour consent, the agent itself becomes responsible for enforcing preferences when acting on the user's behalf. Privacy preferences are expressed in a structured, machine-readable format and carried by the agent during interactions with external services. This allows agents to make context-aware decisions about data

sharing, usage, and disclosure. The approach is compatible with existing consent frameworks and identity systems, including wallet-based models such as the EU Digital Identity Wallet, while remaining independent of any single implementation. By shifting enforcement to the agent layer, TrustAI reduces reliance on voluntary compliance and improves transparency, accountability, and user trust in agentic systems.

Prototype

The prototype demonstrates a front-end implementation showing how agent-enforced privacy preferences can be applied during interactions with websites and services. It illustrates how preferences persist across sessions and are respected even when actions are performed autonomously by an AI agent. The visual concept focuses on clarity of preference expression, visibility of enforcement, and alignment with existing consent and identity ecosystems.

Proposal 7: Enforcing privacy preferences across agent-mediated interactions

Problem statement

User privacy preferences are not consistently respected across the web, particularly when interactions are mediated by AI agents acting on a user's behalf. While Consent Management Platforms (CMPs) are intended to communicate and enforce user choices, compliance is uneven and often relies on voluntary adherence by websites. As agentic AI becomes more prevalent, this gap risks undermining user trust and weakening the effectiveness of existing consent frameworks.

Key insights

Participants identified enforcement, not preference expression, as the primary failure point of current consent mechanisms. CMPs exist, but there are limited incentives or technical safeguards ensuring that websites actually comply with the preferences they receive. When agents

perform actions autonomously, users lose visibility into how their choices are applied, increasing the risk that preferences are ignored or misinterpreted. Without stronger enforcement mechanisms, agentic AI could amplify existing weaknesses in the consent ecosystem rather than resolve them.

Solution

This proposal introduces an agent-level approach to enforcing privacy preferences consistently, regardless of whether an interaction is initiated directly by the user or by an AI agent. Instead of relying solely on website-side compliance, the agent itself carries and enforces the user's privacy preferences during interactions. The solution ensures that consent signals communicated via CMPs are respected by making them enforceable at the point of action. By embedding preference awareness into the agent's decision-making process, the system reduces reliance on voluntary compliance and creates stronger incentives for alignment with user intent. The approach is compatible with existing consent infrastructures and identity-based mechanisms, without being tied to a specific implementation.

Prototype

The prototype demonstrates a front-end implementation showing how an agent can enforce user privacy preferences during interactions with websites. It illustrates how preferences persist across sessions and are applied consistently whether actions are taken manually or autonomously. The visual concept focuses on making enforcement visible and understandable, reinforcing user trust in agent-mediated workflows.