From Chat to Agent: Why Consent Breaks When AI Acts on Your Behalf

MEHMET HAKLIDIR

TUBITAK BILGEM Artificial Intelligence Institute, Gebze, Kocaeli, Turkiye • mehmet.haklidir@tubitak.gov.tr

The rapid evolution from conversational chatbots to autonomous AI agents has outpaced consent mechanisms designed to protect users. While millions share sensitive information with LLM-based systems, evidence shows a widening gap between users' mental models of data handling and actual platform practices. This disconnect deepens as AI browsers navigate the web, handle credentials, and execute transactions. This paper argues that consent is structurally broken across the LLM ecosystem and identifies three design tensions: personalization versus data minimization, agent autonomy versus user control, and transparency versus usability. Building on these tensions, this paper proposes LLM-specific consent mechanisms: contextual in-flow transparency, task-scoped permissions, and AI-mediated privacy guardians, and highlight their consequences for security, comprehension, and power asymmetries in agentic systems.

CCS CONCEPTS • Human-centered computing → HCI theory, concepts and models • Security and privacy → Usability in security and privacy

Additional Keywords and Phrases: consent, privacy, large language models, agentic AI, trustworthy AI

1 INTRODUCTION

When users confide in ChatGPT about anxiety or ask Gemini for help with a sensitive medical question, many assume these conversations vanish once the tab closes. Often, they do not. In practice, we increasingly see people treat LLMs like a private scratchpad. They are then surprised when history, retention, or training settings behave more like a platform than a diary. Analyses of major LLM providers' privacy policies suggest user conversations may be retained by default and may be used to improve models, often via opt-out controls [1]. Yet users who share health information (47%), financial details (35%), and legal concerns (13%) with chatbots remain largely unaware of these practices [2].

This gap between user expectations and platform practices suggests that consent mechanisms designed for the static web cannot keep pace with conversational AI. The situation becomes more precarious as chatbots evolve into AI-powered browsers that can both act and reach into sensitive accounts. In other words, they combine high autonomy with high access. This is the riskiest quadrant in emerging "autonomy–access" threat framing for agentic browsers [3]. OpenAI notes that prompt injection (malicious instructions that hijack agent behavior) may be "unlikely to ever be fully solved" [4].

From what we observe in deployments and everyday use, consent is already structurally strained across the LLM ecosystem, and the gap widens as systems become more agentic. We organize this paper around three design tensions that keep recurring in practice, then propose three provocations intended to spark disagreement: contextual in-flow transparency, task-scoped permission models, and AI-mediated "privacy guardians."

2 THE CONSENT CRISIS ACROSS THE LLM ECOSYSTEM

As these systems move from "talking" to "doing," we see meaningful consent erode in predictable ways. This paper traces this across three contexts: chatbots, AI browsers, and agentic systems.

Chatbots: The Mental Model Gap. Many users treat chatbot exchanges as ephemeral, closer to talking with a forgetful confidant than posting on a platform [5]. That mental model is frequently wrong in practice. Conversational intimacy can create a "disclosure trap," where interface cues encourage oversharing [6]. A December 2024 data breach at chatbot provider WotNot exposed 346,000 private conversations, including identity documents and medical records [7]. Meanwhile, consent is buried in sprawling policy documents and dense settings, fragmenting accountability across providers [1].

AI Browsers: When Consent Meets Autonomy. When ChatGPT Atlas navigates your inbox, it accesses data you never explicitly shared, taking actions with real-world consequences. An agent asked to "summarize my unread emails" ingests messages from senders who never consented to AI processing, a *third-party consent externality* with no clear resolution. Prompt injection vulnerabilities documented across major AI browsers [3, 8] mean users may authorize Task A while a compromised context nudges the agent into Task B. In effect, consent can become non-binding: what a user authorizes is not always what the system actually executes.

Agentic AI: The Consent Ceiling. As AI systems coordinate across services (email, calendar, file storage) they infer information existing in none of these sources individually. Persistent memory means agents' knowledge compounds in ways no initial consent could anticipate. Requiring consent for every action produces debilitating fatigue; blanket permissions ask users to authorize an unknowable future. Neither honors the principle that consent should be informed, specific, and revocable [9].

3 THREE DESIGN TENSIONS

These failures are not accidents. They reflect structural tensions that demand real design trade-offs.

Personalization vs. Data Minimization. Users want AI systems that know them: their preferences, projects, communication patterns. Yet personalization stands in tension with privacy principles that systems should collect only what is necessary. Many current implementations lean toward maximal collection, with opt-out mechanisms that users often fail to notice or understand. A more honest approach would make trade-offs explicit, allowing session-level personalization without long-term retention. *Design implication:* session-based personalization primitives with legible retention toggles.

Agent Autonomy vs. User Control. Users delegate to agents precisely because they want tasks completed without supervision. But every decision delegated removes user oversight. Human-in-the-loop confirmations for "sensitive" actions merely relocate the problem: who defines sensitivity? Any threshold will be over-inclusive (generating fatigue) or under-inclusive (allowing consequential actions without oversight). *Design implication:* graduated autonomy levels with reversibility guarantees.

Transparency vs. Usability. Informed consent requires transparency, but transparency mechanisms that interrupt conversation flow degrade the experience making conversational AI compelling. Research on cookie banners demonstrates a familiar failure mode: users either abandon the interaction or click through without reading [9]. LLM interfaces risk recreating the same pattern, only in conversational form. The challenge is developing transparency that is contextual, intelligible, and minimally intrusive. *Design implication:* ambient, progressive disclosure that learns user preferences over time.

4 DESIGN PROVOCATIONS

This paper offers three provocations to challenge current assumptions and stimulate discussion.

Consequences & trade-offs. Moving beyond one-time clicks shifts consent toward continuous, revocable negotiation. This can reduce the blast radius of prompt injection and cross-service inference (e.g., limiting what an agent can do after a single compromised webpage), but it may also create new fatigue patterns, new intermediaries, and new concentrations of power.

Contextual In-Flow Transparency (Q1/Q3). What if consent interfaces were woven into conversation, surfacing data practices at moments of relevance? This could appear as subtle indicators when exchanges are stored, used for model improvement, or shared with tools, or as prompts when sensitive information is detected: "This seems personal. This conversation is saved in your history and may be used to improve the service. Tap to adjust." The risk is interruption fatigue, especially if "sensitive" prompts fire too often. But the alternative (opaque data flows governed by unread policies) has already failed at scale.

Task-Scoped, Time-Limited Permissions (Q2/Q3). Consider an alternative: agent permissions as bounded contracts, specific to a task, limited in duration, and automatically revoked. Before execution, the agent would present: "To complete this task, I will access [calendar, email, payment]. This permission expires in [30 minutes] or upon completion." Automatic expiry limits what later prompt injection can exploit. The trade-off is friction, and poorly designed prompts could devolve into "banner fatigue." But users may accept the cost when boundaries are concrete, time-limited, and easy to revoke.

AI-Mediated Privacy Guardian (Q2/Q3). A more radical proposal: users could deploy their own AI agent protecting privacy, monitoring, negotiating, and intervening when other systems seek data access. The guardian could translate requests into offers aligned with user preferences: "The travel agent wants your full email history. I suggest offering only flight confirmations from the past month. Approve?" The recursion problem remains: who guards the guardian? Guardians should therefore operate under maximal transparency: auditable logs, local-first processing, and verifiable code or third-party audits. Otherwise, we risk simply relocating the consent problem to a new, powerful intermediary.

5 DISCUSSION: OPEN QUESTIONS FOR THE WORKSHOP

The tensions and provocations above remain genuinely open. They point toward questions that no single discipline can answer alone.

Defining thresholds. If not every agent action can require explicit consent, which actions should? Is sensitivity determined by data type (health, finance), action consequence (payment, deletion), or reversibility? Should thresholds be universal, culturally variable, or individually calibrated? And if individually calibrated, how do users specify preferences without becoming privacy experts?

Standardization possibilities. The current landscape fragments consent across platforms: each LLM provider implements its own policies and opt-out mechanisms. Would cross-platform standards (analogous to Creative Commons for content licensing) be feasible for LLM interactions, especially now that agents routinely span multiple tools and identities within a single task? What would a "consent portability" framework look like, allowing users to carry preferences across AI services?

Regulatory adequacy. Existing frameworks (GDPR, CCPA, EU AI Act) were not designed with agentic LLM systems in mind. Do their core principles (lawful basis, purpose limitation, data minimization) extend adequately to autonomous agents? Or do we need LLM-specific provisions addressing training data consent, agent delegation boundaries, and cross-service inference?

Trust architecture. Our privacy guardian provocation raises trust delegation questions. What institutional or technical arrangements make guardians trustworthy? Open-source code? Independent audits? Structural separation from commercial incentives? How do we avoid relocating the trust problem one level up?

Measuring success. If we redesign consent for LLM systems, how will we know we have succeeded: higher user comprehension, fewer privacy harms, or smaller security blast radii? What would "meaningful consent" look like as a measurable outcome?

Alternative futures. In a data-minimal future where personalization is local and session-scoped, what consent primitives remain necessary? In a memory-rich future where agents coordinate across services by default, what new collective and third-party consent mechanisms become unavoidable?

We do not expect the workshop to resolve these questions definitively. But bringing together researchers and practitioners across disciplines can surface assumptions and expose blind spots. Consent in the age of agentic AI demands cross-disciplinary friction, not just imagination.

6 AUTHOR BIOGRAPHY

Mehmet HAKLIDIR is the Director of the Artificial Intelligence Institute at TÜBİTAK BİLGEM in Türkiye. He leads national-scale AI initiatives across six competence centers, coordinating more than one hundred public, private, and academic partners to operationalise Türkiye's National AI Strategy. He has over 20 years of experience in artificial intelligence, modelling and simulation, robotics, and large-scale digital transformation projects. His research and leadership focus on trustworthy AI, AI governance and policy, safety-critical and human-interactive AI systems, and the deployment of large-scale AI and LLM infrastructures for public good. He has contributed to national and international AI strategies and accountability frameworks, including Türkiye's Trustworthy-AI Accountability Framework and "Reliable AI" certification, and serves as a member or delegate in bodies such as OECD ONE AI, the Global Partnership on AI (GPAI), UNESCO AI Ethics initiatives, and NATO's data and AI governance fora.

Areas of Expertise and Interest: Trustworthy AI; AI governance and policy; LLM and foundation-model ecosystems; Safety-critical and human-interactive AI (including social robots and autonomous systems); Privacy-preserving and accountable AI deployment; AI for national-scale digital transformation.

REFERENCES

[1] Jennifer King et al. 2025. User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies. Stanford HAI.

[2] Sarah Tran et al. 2025. Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective. AAAI/ACM AIES.

[3] Wiz Research. 2025. Agentic Browser Security: 2025 Year-End Review.

[4] OpenAI. 2025. Continuously Hardening ChatGPT Atlas Against Prompt Injection Attacks. OpenAI Blog.

[5] Jabari Kwesi et al. 2025. Exploring User Security and Privacy Attitudes and Concerns Toward the Use of General-Purpose LLM Chatbots for Mental Health. USENIX Security.

[6] Lu Fan et al. 2024. Self-disclosure in Human-Chatbot Interaction. CHI '24.

[7] Pieter Arntz. 2024. AI Chatbot Provider Exposes 346,000 Customer Files. Malwarebytes.

[8] Brave Security Research. 2025. Indirect Prompt Injection in Perplexity Comet.

[9] Alessandro Acquisti et al. 2015. Privacy and Human Behavior in the Age of Information. Science 347, 6221.