The Privacy Guardian Agent: Towards Trustworthy AI Privacy Agents

VINCENT FREIBERGER, Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Germany

The current "notice and consent" paradigm is broken: consent dialogues are often manipulative, and users cannot realistically read or understand every privacy policy. While recent LLM-based tools empower users seeking active control, many with limited time or motivation prefer full automation. However, fully autonomous solutions risk hallucinations and opaque decisions, undermining trust. I propose a middle ground — a Privacy Guardian Agent that automates routine consent choices using user profiles and contextual awareness while recognizing uncertainty. It escalates unclear or high-risk cases to the user, maintaining a human-in-the-loop only when necessary. To ensure agency and transparency, the agent's reasoning on its autonomous decisions is reviewable, allowing for user recourse. For problematic cases, even with minimal consent, it alerts the user and suggests switching to an alternative site. This approach aims to reduce consent fatigue while preserving trust and meaningful user autonomy.

Additional Key Words and Phrases: Agents, Notice and Consent, LLMs, Trust

1 Motivation

The current approach of "notice and consent" on the web is dysfunctional by design. It places an impossible cognitive burden on users who cannot read, let alone comprehend, every privacy policy they encounter [12, 15]. Further, consent dialogues tend to employ dark patterns that manipulate users into habituated acceptance rather than informed choice [14]. While recent Large Language Model (LLM) tools promise to restore informational sovereignty by summarizing and explaining policies [3, 9, 16], they primarily serve motivated users seeking active control. For the majority, particularly "Unconcerned" or "Lazy Experts" privacy profiles [4], the friction of manual review remains a barrier, often exacerbated by time constraints and familiarity bias [9, 15]. However, fully automated agentic solutions risk eroding trust through hallucinations or misaligned, opaque decision-making [7]. As a result, we face a dilemma: manual control is impractical for most, but full automation removes the user agency necessary for meaningful consent.

2 A Privacy Guardian Agent

I argue for a middle ground – a hybrid system that combines user privacy profiles [4, 11] with contextual analysis of data flows to automate routine consent decisions while escalating uncertain or high-risk cases to the user. Going beyond choosing consent options, the agent should help users decide whether to use a service at all when even "essential-only" consent implies unacceptable risks to users, like sensitive inferences or suspicious third-party sharing. The agent alerts the user and, inspired by PrivacyCheck [13], proposes alternative services that better align with their privacy profile.

This approach minimizes consent fatigue while preserving user agency. Users can always review the agent's rationale for any decision and override it when needed. By delegating routine decisions to the agent, users avoid habituation and retain cognitive resources and control for uncertain or high-stakes decisions, aligning with Zhang et al.'s (2026) [17] design directions for context-aware just-in-time interventions.

User Profiling: New users disengage when faced with an extensive number of setup questions. Instead, utilizing privacy personas [4, 11] along the dimensions of motivation and privacy knowledge can help calibrate the agent to users' needs. For example, an "Unconcerned" user might want to auto-accept analytics cookies on news sites, while a "Fundamentalist" profile would auto-reject third-party tracking. The tool could utilize a user self-assessment on

Author's Contact Information: Vincent Freiberger, Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Leipzig, Germany, freiberger@cs.uni-leipzig.de.

motivation and privacy knowledge following Marky et al. (2024) [11] for initial privacy profiling and potentially refine and adjust based on locally stored continued usage data.

Contextual Integrity analysis: The agent should understand the context of a data request (like sensitivity of data, estimated urgency of users tasks or evaluation of stated data collection purposes) and leverage the obtained user privacy profile. Guardian Agents could utilize LLMs to parse privacy policies into Contextual Integrity norms [2]. So purposes for collected data types for a given service could be evaluated considering how privacy-sensitive users actually are. In case CI norms cannot be reliably extracted, the decision should be escalated to the user.

Reliability Calibration: To reduce the risk of AI hallucination or factual errors, I suggest reliability calibration. The system should provide privacy policy evidence to support its interpretation, making its reasoning transparent and auditable. Further, the agent should communicate the reliability of its risk assessment utilizing measures like model uncertainty, consistency checks or following more elaborate approaches like Tanneru et al (2024) [10]. It should flag inconsistencies, missing information, or ambiguous language that prevent reliable Contextual Integrity analysis. When the agent cannot confidently determine whether a data flow violates the user's profile, it escalates to the user. The escalation message explains what is uncertain and why it matters for the user's privacy preferences. In cases where the agent identifies unacceptable risk, it also communicates its confidence level in that assessment, allowing users to calibrate their trust appropriately.

3 Discussion

A key question is how far consent automation can legitimately extend under GDPR, which requires consent to be "freely given" and "specific." I argue the agent should be framed as a decision-support tool rather than a legal proxy: the user configures the profile, reviews high-stakes decisions, and can audit past choices. Future work should examine under which conditions such delegation enhances rather than undermines autonomy in practice, and how reliability calibration influences trust. Moreover, the question of accountability should be explored: Who bears responsibility if the agent makes mistakes or misinterprets a policy? Beyond hallucinations, the agent could introduce new manipulation risks if it is captured by vendor incentives or adversarial policies requiring guardrails like a transparent rationale and robustness measures.

Ironically, a Privacy Guardian Agent that protects user privacy requires collecting and processing sensitive behavioral data, which creates a new privacy risk. User privacy profiles combined with browsing context and interaction patterns could reveal details about users' values, concerns, and online habits. Consequently, profile data must be stored locally and processed either locally or with privacy-preserving techniques like privacy sanitization in place. The agent should retain only aggregated preference patterns, not granular histories. Without such safeguards, Guardian Agents risk becoming data protection risks themselves, undermining the informational sovereignty they aim to preserve.

Embedding agents in data cooperatives could improve collective protections but raises questions about whose risk tolerance dominates when preferences conflict, and whether shared norms reduce individual agency or enhance it through collective bargaining power.

4 About the Author

I am a PhD student focused on empowering users regarding their privacy online by building LLM-based privacy policy assistants. I have published a Late Breaking Work at CHI 2025 [8] and have an accepted full paper at CHI 2026 [9], which present an interactive LLM-based privacy policy assistant aiming to raise user awareness and understanding. My earlier work mapped the concept of fairness to privacy policies [6], discussed challenges with LLM-based privacy policy

assessments from a technical, legal and ethical perspective [1], and ran early experiments on LLM-based privacy policy assessments [5]. As I found varying usage patterns and user needs, aligning with related work on privacy profiles [4, 11], my current goal is to personalize privacy assistants to maximize their impact. Last CHI, I attended the Human-centered XAI workshop, proposing ideas on how assessments could be made more transparent and trustworthy [7]. In this regard, I am also particularly interested in calibrated trust and maintaining user agency, which is well aligned with the workshop's CfP.

References

[1] Irem Aydin, Hermann Diebel-Fischer, Vincent Freiberger, Julia Möller-Klapperich, Erik Buchmann, Michael Färber, Anne Lauber-Rönsberg, and Birte Platow. 2024. Assessing Privacy Policies with AI: Ethical, Legal, and Technical Challenges. arXiv:2410.08381 [cs.CY] https://arxiv.org/abs/2410.08381

[2] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. 2006. Privacy and contextual integrity: Framework and applications. In *2006 IEEE symposium on security and privacy (S&P'06)*. IEEE, 15–pp.

[3] Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Jia-Jun Li, and Yaxing Yao. 2025. CLEAR: Towards Contextual LLM-Empowered Privacy Policy Analysis and Risk Generation for Large Language Model Applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 277–297. doi:10.1145/3708359.3712156

[4] Janna Lynn Dupree, Richard Devries, Daniel M. Berry, and Edward Lank. 2016. Privacy Personas: Clustering Users via Attitudes and Behaviors toward Security Practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5228–5239. doi:10.1145/2858036.2858214

[5] Vincent Freiberger and Erik Buchmann. 2024. Fair balancing? Evaluating LLM-based privacy policy ethics assessments. In *Proceedings of the Third European Workshop on Algorithmic Fairness (EWAF'24)*. CEUR Workshop Proceedings, Aachen, Germany, 1–19.

[6] Vincent Freiberger and Erik Buchmann. 2024. Legally Binding but Unfair? Towards Assessing Fairness of Privacy Policies. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics* (Porto, Portugal) *(IWSPA '24)*. Association for Computing Machinery, New York, NY, USA, 15–22. doi:10.1145/3643651.3659890

[7] Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2025. Explainable AI in Usable Privacy and Security: Challenges and Opportunities. In *Proceedings of the 2025 Workshop on Human-Centered Explainable AI @CHI*. Zenodo, Genève, Switzerland, 56–64.

[8] Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2025. "You Don't Need a University Degree to Comprehend Data Protection This Way": LLM-Powered Interactive Privacy Policy Assessment. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 36, 12 pages. doi:10.1145/3706599.3719816

[9] Vincent Freiberger, Arthur Fleig, and Erik Buchmann. 2026. Helping Johnny Make Sense of Privacy Policies with LLMs.

[10] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying Uncertainty in Natural Language Explanations of Large Language Models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 238)*, Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (Eds.). PMLR, 1072–1080. https://proceedings.mlr.press/v238/harsha-tanneru24a.html

[11] Karola Marky, Alina Stöver, Sarah Prange, Kira Bleck, Paul Gerber, Verena Zimmermann, Florian Müller, Florian Alt, and Max Mühlhäuser. 2024. Decide Yourself or Delegate - User Preferences Regarding the Autonomy of Personal Privacy Assistants in Private IoT-Equipped Environments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 851, 20 pages. doi:10.1145/3613904.3642591

[12] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *Isjlp* 4 (2008), 543.

[13] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K Suzanne Barber. 2020. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In *Proceedings of the 29th ACM international conference on information & knowledge management*. Association for Computing Machinery, New York, NY, USA, 3441–3444.

[14] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376321

[15] Varun Shiri, Maggie Xiong, Jinghui Cheng, and Jin L.C. Guo. 2024. Motivating Users to Attend to Privacy: A Theory-Driven Design Study. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 258–275. doi:10.1145/3643834.3661544

[16] Bolun Sun, Yifan Zhou, and Haiyun Jiang. 2025. Empowering Users in Digital Privacy Management through Interactive LLM-Based Agents. In *The Thirteenth International Conference on Learning Representations*. International Conference on Learning Representations, Appleton, WI, USA, 1–21.

[17] Shuning Zhang, Eve He, Sixing Tao, Yuting Yang, Ying Ma, Ailei Wang, Xin Yi, and Hewu Li. 2026. A Scoping Review and Guidelines on Privacy Policy's Visualization from an HCI Perspective. arXiv preprint arXiv:2601.17368.