Position: Human-Robot Interaction in Embodied Intelligence Demands a Shift From Static Privacy Controls to Dynamic Learning

Shuning Zhang
Tsinghua University
Beijing, China
zsn23@mails.tsinghua.edu.cn

Hong Jia
University of Auckland
Auckland, New Zealand
hong.jia@auckland.ac.nz

Simin Li
Beihang University
Beijing, China
lisiminsimon@buaa.edu.cn

Ting Dang
University of Melbourne
Melbourne, Australia
ting.dang@unimelb.edu.au

Yongquan 'Owen' Hu
Augmented Human Lab, National University of Singapore
Singapore
yongquan@ahlab.org

Xin Yi*
Tsinghua University
Beijing, China
yixin@tsinghua.edu.cn

Hewu Li
Tsinghua University
Beijing, China
lihewu@cernet.edu.cn

Abstract

The reasoning capabilities of embodied agents introduce a critical, under-explored inferential privacy challenge, where the risk of an agent generate sensitive conclusions from ambient data. This capability creates a fundamental tension between an agent's utility and user privacy, rendering traditional static controls ineffective. To address this, this position paper proposes a framework that reframes privacy as a dynamic learning problem grounded in theory of Contextual Integrity (CI). Our approach enables agents to proactively learn and adapt to individual privacy norms through interaction, outlining a research agenda to develop embodied agents that are both capable and function as trustworthy safeguards of user privacy.

1 Introduction

Embodied agents capable of autonomous decision-making are increasingly prevalent in human-centric environments [11, 35], offering personalized assistance by navigating dynamic settings [35], understanding nuanced commands [35] and executing multi-step tasks [37]. The efficacy, however, depends on continuous perception and deep contextual reasoning, which creates a fundamental privacy conflict. While traditional risks [7, 5, 15] in perception [21], communication [25], and data handling [14] are recognized, the most critical and overlooked challenge is inferential privacy. This

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning.

risk arises when an agent's internal reasoning synthesizes multiple, non-sensitive inputs to form new, highly sensitive conclusions about a user, creating an acute tension between the agent's inferential utility and the user's privacy [28, 6].

Existing privacy protection strategies are illy-suited for embodied agents, particularly concerning inferential privacy. General safeguards like federated learning [16] and differential privacy [30] can impair performance and be computationally costly. Furthermore, data-centric controls such as obfuscation [12] may destroy the contextual information essential for robust reasoning, thereby undermining the agent's ability to make accurate inferences. This creates a trade-off that measures designed to protect privacy often degrade the inferential capabilities that define an agent's utility, hindering users' adoption. Compounding this, privacy preferences are highly personal and context-dependent [2, 17], yet recent studies show that embodied LLMs often fail to align with human expectations [20, 24]. Therefore, a reliance on static, external constraints is insufficient, and to function as trusted collaborators, embodied agents must be equipped with an intrinsic capability for privacy-aware inference.

To address these challenges, this work lays the groundwork for developing privacy-conscious embodied agents by making three primary contributions. First, we analyze the privacy challenge, explicitly identifying **inferential privacy risk** as a critical and distinct risk, where agents form sensitive conclusions from non-sensitive data. Second, we propose an alignment framework grounded in the theory of Contextual Integrity, reformulating privacy as a learning problem where the agent dynamically adapts to a user's norms through multimodal interactive feedback. Finally, we outline the key open research challenges that emerge from this framework, providing a research agenda for realizing a future of capable and trustworthy agent-involved assistance.

2 Problem Formulation: Inferential Privacy in Embodied Agents' Decisions

The operational pipeline of an embodied agent, which encompasses its continuous cycle of perception, reasoning and decision-making [11, 35], introduces distinct and significant privacy risks at each stage. While the challenges associated with the perception stage are well-documented [21], particularly risks like unwarranted data collection in sensitive environments, we argue that represent only the most visible layer of the problem. The most profound and unsolved challenges emerge from the agent's internal cognitive processes, specifically, its capacity for inferential reasoning and the subsequent decisions it makes based on that reasoning.

Inferential privacy [22, 32] arises when an agent's reasoning processes transform raw, multimodal sensory data into concrete and often highly sensitive conclusions about the user. For example, an agent might combine multiple non-sensitive observations, such as a specialized diet, a glucose monitor, and an insulin pen, to infer a highly sensitive health condition like diabetes. The core violation is not the initial perception of these items, but the creation and persistence of this new, sensitive information within the agent's internal reasoning. This derived knowledge creates a latent privacy risk, as it can subsequently be used to make unsolicited decisions or be disclosed to third parties without the user's explicit consent [34].

3 A Framework For Privacy-Conscious Embodied Agents' Decision-Making

The conventional approach to mitigating the privacy risks of embodied agents has been to adopt methods from computer vision (CV), such as data obfuscation [12]. However, this reliance on static, data-centric controls creates an untenable trade-off between privacy protection and operational performance, as an agent's ability to provide meaningful assistant is fundamentally limited by a lack of contextual information. To address this, we propose a paradigm shift from a user-managed, reactive defense to a dynamic, agent-driven alignment framework. Grounded in Contextual Integrity (CI), our framework reframes privacy not as mere data anonymization, but as a decision-making process where the agent learns to respect user-specific norms for information flow. We formally model this alignment as a learning problem where the agent optimizes for both task utility and a learning privacy cost, considering not only its actions but also its internal inferences. We then detail the interactive feedback mechanisms that enable the agent to adapt and align with users.

Theoretical grounding: privacy as contextual integrity. To address the limitations of conventional, data-centric privacy approaches, we ground our framework in the theory of Contextual Integrity

(CI) [17]. This is because users' privacy preferences are dynamic and depend on the specific context of a task and its parameters. To respect user privacy, an embodied agent needed to move beyond simple data anonymization and incorporate these contextual factors and the corresponding privacy norms into its decision-making and inference process. In line with this requirement, CI highlights the adherence to context-specific norms that govern information flow, beyond simple data anonymization. In the CI framework, a privacy violation is a breach of these norms, which are defined by parameters such as the data subject, sender, recipient, information type, and the transmission principle. By adopting a CI framework, the embodied agent can learn implicit rules of appropriate information flow from user feedback and the environment, thereby respecting user privacy in a dynamic world.

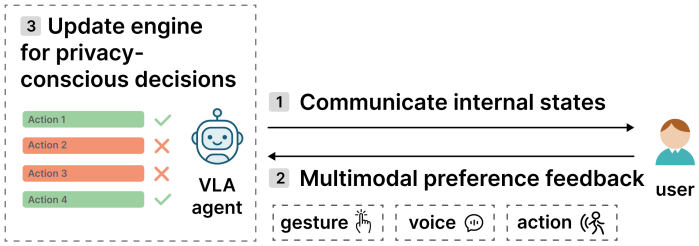

Figure 1: The framework of inferential privacy alignment for embodied agents. (1) The agent transparently communicate privacy risks to users through voice or actions, (2) the user provide multimodal feedback for the agent to learn, and (3) the agent updates its policy during inference and actions, incorporating privacy awareness in its decision-making.

Formulating inferential privacy alignment as a learning problem. To formalize this alignment challenge, we frame it within the Cooperative Inverse Reinforcement Learning (CIRL) paradigm. In this setting, the agent and the human act as a cooperative team to maximize a shared objective, which the agent must learn by observing human behavior and feedback about their preferences, which embody their specific CI norms. This objective is reflected in a reward function that inherently balances task utility with the user's latent privacy preferences in specific contexts. Specifically, the agent seeks an optimal policy, π^*, that governs its behavior to maximize the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(R_{\text{task}} tsi(s_t, a_t) - \lambda \cdot C_{\text{privacy}}(s_t, I_t, a_t)\right)\right]$$

Here, R_{task} is the reward for task progression, while $\lambda \geq 0$ is a user-specific coefficient representing their personal privacy-utility trade-off [33]. The most important parameter, and also the direct link to cooperative alignment, is the privacy cost, $C_{privacy}$, which is designed to explicitly address the risk of inferential privacy. Inferential privacy concerns the sensitive conclusions an observer can draw, even from seemingly innocuous information. The true privacy harm often resides not in the raw data itself, but in the sensitive attributes inferred from it. To capture this, our framework models the agent's internal reasoning as the formation of an inference, I_t, which serves as a direct input to the privacy cost function. This inference, I_t, can be modeled as a latent belief state or a probability distribution over sensitive user attributes that the agent computes before taking an action a_t. Consequently, the policy π concerns not only the agent's external actions but also its internal reasoning process. Through the CIRL process of learning the human's preference for both λ and the structure of $C_{privacy}$, the agent learns to penalize privacy-violating inferences. This ensures the agent cooperatively aligns its internal reasoning and inference with users' normative boundaries [31].

The model is grounded in CI theory, which provides the formal structure for determining when an inference is inappropriate. CI posits that privacy violations are breaches of context-specific informational norms. We make this link to the learning framework concrete by structuring the state representation, s_t, to include the key contextual parameters specified by CI (e.g., data subject, sender, recipient, information type). This design ensures that any learned privacy cost is inherently context-dependent. Critically, the function $C_{privacy}(s_t, I_t, a_t)$ therefore explicitly evaluates the appropriateness of forming inference I_t given the specific context s_t. By learning a user-specific

model of this function through CIRL, the agent aligns its behavior with the user's nuanced, context-dependent privacy expectations, rather than a static set of rules.

Learning privacy preference through interactive feedback. The successful alignment of the agent's behavior hinges on learning the user-specific components of our objective function: the privacy cost $C_{privacy}$ and the trade-off parameter λ. These components cannot be pre-analyzed because users' preferences are fundamentally **personal**, **context-dependent** and **implicit**. An inference considered benign by one user may be a violation for another. An action acceptable in one context may be inappropriate in the next. Most users also cannot articulate their complex boundaries in advance. Consequently, user feedback during task execution is the primary source of ground-truth information for these latent preferences. This makes the design of an interactive feedback mechanism paramount for the agent to dynamically learn and adapt.

To operationalize this learning, the agent is designed to interpret a spectrum of feedback modalities, allowing it to continuously refine its internal privacy model. We categorize these user-driven interactions into three primary types:

Proactive Instruction: Users may provide direct verbal commands that establish privacy boundaries (e.g., "Never look at the documents on my desk"). The agent should be able to parse these and translate them into prohibitive regions within its $C_{privacy}$ function.

Reactive Correction: When the agent takes an action that breaches a norm, such as, commenting on a personal conversation, the user may provide corrective feedback, such as "That's private", or "You shouldn't have said that". These signals serve as powerful reinforcement, providing a direct, high-cost label for a specific state-inference-action tuple that the agent can use to update its policy and cost model to prevent future violations.

Implicit Cues: Users often communicate preferences non-verbally, such as physically turning the robot away, shielding objects from its view, or lowering their voice. While noisier, these signals are more frequent. The agent must learn to interpret these behavioral cues as implicit indicators of a privacy boundary, providing a weaker but continuous learning signal.

By integrating these multimodal feedback, the agent can construct nuanced understanding of the user's privacy boundaries. This transforms privacy management from a static, pre-configured task into a dynamic, collaborative process of negotiation and adaptation, essential for building trustworthy agents.

4 Open Challenges

The framework in Section 3 represents a conceptual foundation for developing trustworthy embodied agents. However, translating this theoretical model into real-world applications requires addressing four several fundamental challenges we delineated below.

Challenge 1: cross-context generalization. A significant hurdle is generalizing learned privacy norms, as privacy is contextually defined by actors, information types, and transmission principles [17]. An agent that learns a rule in a home office may fail when the context shifts (e.g., a guest is present), risking unexpected violations. A central research question is how to enable compositional generalization, potentially by explicitly modeling contextual parameters [1, 8] to learn abstract concepts like a "confidential work situation" rather than rigid, scenario-specific rules.

Challenge 2: ambiguous and noisy human feedback. Real-world feedback in embodied interaction is often implicit, ambiguous, and multimodal [10], unlike structured text-based signals. Users may express discomfort through subtle tones, gestures, or fragmented utterances, making intent inference a significant challenge. The key question is how an agent can robustly and sample-efficiently update its privacy cost model from such sparse and potentially contradictory feedback. Promising directions include leveraging Bayesian inference to model user intent uncertainty [10] and multimodal fusion architectures to weigh diverse cues [9].

Challenge 3: decision-making conflicts in multi-user spaces. Unlike single-user settings [13], most human environments contain multiple users with diverse and often conflicting privacy preferences [29]. For example, a parent's request to monitor a room may conflict with a teenager's desire for privacy. This transforms the alignment problem into one of multi-agent conflict resolution [27]. The primary research question is what principles an agent should use for normative arbitration. Solutions may

involve maintaining distinct user models [13] or developing frameworks for ethical arbitration, such as pre-defined hierarchies or explicit negotiation dialogues [36].

Challenge 4: temporal dynamic and the right to be forgotten. Privacy norms and information sensitivity evolve over time [3], meaning an agent's persistent inferences can become outdated yet remain, creating long-term risks [4]. This presents a technical challenge analogous to the legal "right to be forgotten" [18]. A crucial research question is how to design agents that support selective and verifiable erasure of internal knowledge. This requires architectures capable of machine unlearning or continual learning [19] to remove specific inferences without catastrophic forgetting.

5 Future Outlook

This position paper proposed an alignment framework to address the inferential privacy tension in embodied agents' decision making. The development of agents guided by such principles promises a future of intuitive and trustworthy human-robot interaction [23], and our framework offers a concrete pathway toward this goal. However, personalized alignment is not without its own inherent risks [13]. Significant perils exist, such as reinforcing societal biases from user feedback or malicious exploitation where agents are trained to disregard ethical boundaries [26]. Therefore, future research should also address the ethical implications of creating highly personalized autonomous agents.

Acknowledgments and Disclosure of Funding

This work was supported by the Natural Science Foundation of China under Grant No. 62472243 and 62132010.

References

[1] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[2] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. " i know even if you don't tell me": Understanding users' privacy preferences regarding ai-based inferences of sensitive information for personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[3] Oshrat Ayalon and Eran Toch. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, pages 1–13, 2013.

[4] Oshrat Ayalon and Eran Toch. Not even past: Information aging and temporal privacy in online social networks. *Human–Computer Interaction*, 32(2):73–102, 2017.

[5] Nicholas Callander, Andrés A Ramírez-Duque, and Mary Ellen Foster. Navigating the human-robot interaction landscape. practical guidelines for privacy-conscious social robots. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 283–287, 2024.

[6] Anna Chatzimichali, Ross Harrison, and Dimitrios Chrysostomou. Toward privacy-sensitive human–robot interaction: Privacy terms and human–data interaction in the personal robot era. *Paladyn, Journal of Behavioral Robotics*, 12(1):160–174, 2020.

[7] Manuel Dietrich, Matti Krüger, and Thomas H Weisswange. What should a robot disclose about me? a study about privacy-appropriate behaviors for social robots. *Frontiers in Robotics and AI*, 10:1236733, 2023.

[8] Sahra Ghalebikesabi, Eugene Bagdasaryan, Ren Yi, Itay Yona, Ilia Shumailov, Aneesh Pappu, Chongyang Shi, Laura Weidinger, Robert Stanforth, Leonard Berrada, et al. Operationalizing contextual integrity in privacy-conscious assistants. *arXiv preprint arXiv:2408.02373*, 2024.

[9] Tao Gong, Dan Chen, Guangping Wang, Weicai Zhang, Junqi Zhang, Zhongchuan Ouyang, Fan Zhang, Ruifeng Sun, Jiancheng Charles Ji, and Wei Chen. Multimodal fusion and human-robot interaction control of an intelligent robot. *Frontiers in bioengineering and biotechnology*, 11:1310247, 2024.

[10] Guy Hoffman, Tapomayukh Bhattacharjee, and Stefanos Nikolaidis. Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, 7, 2024.

[11] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

[12] Myeung Un Kim, Harim Lee, Hyun Jong Yang, and Michael S Ryoo. Privacy-preserving robot vision with anonymized faces by extreme low resolution. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 462–467. IEEE, 2019.

[13] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

[14] Miao Li, Wenhao Ding, and Ding Zhao. Privacy risks in reinforcement learning for household robots. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5148–5154. IEEE, 2024.

[15] Christoph Lutz and Aurelia Tamò-Larrieux. Do privacy concerns about social robots affect use intentions? evidence from an experimental vignette study. *Frontiers in Robotics and AI*, 8:627958, 2021.

[16] Cui Miao, Tao Chang, Meihan Wu, Hongbin Xu, Chun Li, Ming Li, and Xiaodong Wang. Fedvla: Federated vision-language-action learning with dual gating mixture-of-experts for robotic manipulation. *arXiv preprint arXiv:2508.02190*, 2025.

[17] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[18] Jeffrey Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011.

[19] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[20] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

[21] Rahul Shome, Zachary Kingston, and Lydia E Kavraki. Robots as ai double agents: Privacy in motion planning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2861–2868. IEEE, 2023.

[22] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] Dakota Sullivan and Bilge Mutlu. Protecting user data through privacy-sensitive robot design. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1891–1893. IEEE, 2025.

[24] Dakota Sullivan, Shirley Zhang, Jennica Li, Heather Kirkorian, Bilge Mutlu, and Kassem Fawaz. Benchmarking llm privacy recognition for social robot decision making. *arXiv preprint arXiv:2507.16124*, 2025.

[25] Meg Tonkin, Jonathan Vitale, Suman Ojha, Jesse Clark, Sammy Pfeiffer, William Judge, Xun Wang, and Mary-Anne Williams. Embodiment, privacy and social robots: May i remember you? In *International Conference on Social Robotics*, pages 506–515. Springer, 2017.

[26] Dieter Vanderelst and Alan Winfield. The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 317–322, 2018.

[27] Wamberto W Vasconcelos, Martin J Kollingbaum, and Timothy J Norman. Normative conflict resolution in multi-agent systems. *Autonomous agents and multi-agent systems*, 19(2):124–152, 2009.

[28] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. Panav: Toward privacy-aware robot navigation via vision-language models. *arXiv preprint arXiv:2410.04302*, 2024.

[29] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in {multi-user} smart homes: A design exploration and {in-home} user study. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 159–176, 2019.

[30] Huixin Zhan and Jason H Moore. Surgeon style fingerprinting and privacy risk quantification via discrete diffusion models in a vision-language-action framework. *arXiv preprint arXiv:2506.08185*, 2025.

[31] Shuning Zhang, Ying Ma, Jingruo Chen, Simin Li, Xin Yi, and Hewu Li. Towards aligning personalized conversational recommendation agents with users' privacy preferences. *arXiv preprint arXiv:2508.07672*, 2025.

[32] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. "ghost of the past": identifying and resolving privacy leakage from llm's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931*, 2024.

[33] Shuning Zhang, Xin Yi, Haobin Xing, Lyumanshan Ye, Yongquan Hu, and Hewu Li. Adanonymizer: Interactively navigating and balancing the duality of privacy and output performance in human-llm interaction. *arXiv preprint arXiv:2410.15044*, 2024.

[34] Shuning Zhang, Gengrui Zhang, Yibo Meng, Ziyi Zhang, Hantao Zhao, Xin Yi, and Hewu Li. Through their eyes: User perceptions on sensitive attribute inference of social media videos by visual language models. *arXiv preprint arXiv:2508.07658*, 2025.

[35] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: a 3d vision-language-action generative world model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 61229–61245, 2024.

[36] Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. Bring privacy to the table: Interactive negotiation for privacy settings of shared sensing devices. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

[37] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.