Defining and Auditing Dependency-Inducing Human-Robot Interaction

Tianqi Zhu
School of Computing, National University of Singapore

***Abstract*—As embodied foundation models enable increasingly capable assistive robots, well-intentioned assistance may gradually erode human agency and foster unhealthy dependence. Existing robot-safety frameworks emphasize physical harm, reliability, and task success but lack principled methods for defining and auditing dependency-inducing interaction. We introduce a cognitive-state model capturing perceived capability, delegation habit, and graded need for assistance, together with a stochastic user-response model. From this model, we define the Dependency Induction Index (DII), a counterfactual metric measuring the expected increase in unnecessary reliance caused by a robot action relative to non-intervention. We then present a post-deployment auditing methodology that infers latent cognitive states from interaction logs through Bayesian filtering and computes aggregate risk metrics, including the Dependency Induction Rate (DIR) and thresholded violation rate. The framework provides an operational basis for assessing long-term agency risk in assistive robots.**

I. Introduction

Assistive robots are increasingly deployed in homes, healthcare, and elder care. Although existing safety frameworks address physical harm, reliability, and task success, they largely overlook assistance that gradually erodes human agency and fosters unhealthy dependence. Higher robot autonomy can improve task performance while reducing users' sense of agency [1], and users may adapt their behaviour to exploit robotic assistance [2]. As embodied foundation models become more capable and ubiquitous, this risk becomes increasingly important.

We address this gap with a formal framework for dependency-inducing interaction. We model user cognition through latent variables for perceived capability, delegation habit, and task-grounded need for assistance, together with a stochastic user-response model. From these dynamics, we derive a counterfactual **Dependency Induction Index (DII)**, measuring the expected increase in unnecessary help reliance caused by a robot action relative to waiting. We further develop a post-deployment auditing pipeline that reconstructs latent states from interaction logs through Bayesian filtering and computes the **Dependency Induction Rate (DIR)** as an interpretable measure of long-term agency risk.

The contributions are:

1) A computational definition of dependency-inducing human-robot interaction based on a latent-state model, stochastic user behavioral model, and counterfactual DII.
2) A post-deployment auditing pipeline that uses belief inference and log replay to compute dependency-risk metrics.
3) A discussion of identifiability, calibration requirements, synthetic validation, and red-teaming protocols for dependency-risk auditing.

II. Related Work

A. Adaptive Assistance and Foundation Models

Shared-autonomy research models assistive interaction as collaboration between adaptive human and robot partners. Representative approaches include policy blending for shared control, mutual adaptation that reasons about latent human adaptability, and trust-aware planning that incorporates latent human trust into robot decision making [3–5]. Recent work further shows that users intentionally adapt their behaviour in response to robotic assistance [2] and examines when robots should involve users in uncertain or risky decisions [6]. Building on robotics foundation models that enables task generalization and natural language interaction [7–10], assistive systems infer user intent from multimodal observations and proactively initiate assistance without explicit commands [11–13]. Bayesian user models have also been proposed to determine when to intervene and with what confidence for personalized proactive assistance [14]. Existing work, however, primarily optimizes assistance quality and task performance rather than evaluating how adaptive assistance influences users' long-term autonomy.

B. Human-Centered Evaluation of Assistive Robots

Research on automation established calibrated trust and appropriate reliance as central principles for human–automation interaction, distinguishing appropriate reliance from misuse, disuse, and automation bias [15–18]. HRI extends these ideas by examining users' sense of agency, perceived control, and long-term independence. Studies show that increasing robot autonomy can reduce users' sense of agency [1], while users prefer assistive systems that preserve meaningful opportunities for control [19]. Recent surveys likewise identify preserving human autonomy as a growing objective in HRI [20]. Assistive robotics further argues that robots should support long-term independence rather than maximize immediate task performance [21, 22]. Although recent work has begun modeling user behaviour over repeated interactions [23], evaluation remains focused on individual interactions rather than cumulative changes in autonomy over prolonged robot use.

C. Long-Term Human Adaptation and Dependency

Dependency is increasingly recognized as a distinct challenge in long-term human–AI interaction. Prior work discusses dependency conceptually [24], studies skill atrophy under AI assistance [25], and examines ethical concerns surrounding assistive robots and cognitive offloading [26–28]. Self-efficacy theory further suggests that repeated delegation may reduce confidence in independent action [29]. Together, these studies indicate that repeated assistance can alter confidence, delegation habits, and willingness to act independently even when short-term performance improves. However, existing work largely provides conceptual analyses or empirical observations rather than formalizing dependency as a latent cognitive process with counterfactual metrics and post-deployment auditing. We address this gap through a latent cognitive-state model, the DII, and a Bayesian auditing pipeline.

III. A Formal Definition of Dependency-Inducing Interaction

To audit dependency risk, we first define what constitutes a dependency-inducing interaction in a precise and computable manner. This section lays out a formal model of the human user's cognitive state, behavioural responses, and the robot's decision-making process, culminating in the Dependency Induction Index.

A. Cognitive State Space

We model the human-robot assistance scenario as a partially observable Markov decision process (POMDP). The true state at time t is

$$s_t = \big(s_t^{\text{task}},\ s_t^{\text{user}}\big), \tag{1}$$

where s_t^{task} is the physical state of the task and context, and s_t^{user} collects user-level variables relevant to assistance decisions:

$$s_t^{\text{user}} = (C_t,\ H_t,\ N_t). \tag{2}$$

Here C_t and H_t are latent cognitive-behavioural variables, while N_t is a task-grounded estimate of actual need for assistance that may depend on both the user's condition and the current task context. The three components are:

- $C_t \in [0,1]$: **perceived capability / self-efficacy**. The user's belief in their ability to perform the task independently. High C_t indicates robust confidence; low C_t indicates reduced perceived control.
- $H_t \in [0,1]$: **habit strength for delegation**. The degree to which asking for or accepting robotic help has become the default behavioural response.
- $N_t \in [0,1]$: **graded actual need for assistance**. $N_t = 0$ means the user can perform the task independently with little difficulty; $N_t = 1$ means the user cannot safely or successfully perform the task without assistance. Intermediate values represent partial need, such as high effort, risk, pain, or cognitive load. Unlike C_t and H_t, N_t is not intended to represent dependence; it is a moderator that distinguishes warranted from unnecessary assistance.

Rather than treating dependence as an arbitrary weighted sum, we define the audit target behaviorally: the probability that the user will seek or accept assistance in a low-need state. Let U_t denote the user's response. We define the instantaneous low-need dependence tendency as

$$D_t = P(U_t \in \mathcal{U}_{\text{help}} \mid C_t, H_t, N_t), \tag{3}$$

where $\mathcal{U}_{\text{help}} = \{\texttt{request\_help}, \texttt{accept\_offer}\}$. A simple parametric instantiation is

$$D_t = \sigma(\theta_0 + \theta_H H_t - \theta_C C_t - \theta_N N_t), \tag{4}$$

where $\sigma(\cdot)$ is the logistic function and $\theta_H, \theta_C, \theta_N > 0$. The negative coefficient on N_t is intentional: high actual need makes assistance warranted and therefore reduces the audited quantity of *unnecessary* dependence, even though it may increase ordinary help-seeking. This definition makes dependency risk a behavioural probability explained by latent cognitive variables and task-grounded need, rather than a primitive score chosen by the designer.

B. Robot Actions and Observations

At each step the robot selects an action from

$$\mathcal{A} = \{\,\texttt{wait},\ \texttt{offer},\ \texttt{encourage}\,\}. \tag{5}$$

- `wait`: the robot does not proactively intervene; the user may act independently or request help.
- `offer`: the robot proactively proposes assistance.
- `encourage`: the robot prompts the user toward autonomy, for example: "You handled this well last time—would you like to try on your own first?"

Observations o_t include help requests, offer acceptance or decline, task success, completion time, smoothness, retries, and other performance signals. The observation model is stochastic:

$$o_t \sim P\big(o_t \mid s_t^{\text{user}}, s_t^{\text{task}}, a_t^R\big). \tag{6}$$

C. User Behavioral Model

Let u_t denote the user response, drawn from a finite set $\mathcal{U}$ that includes:

- `act_alone`: the user attempts the task without assistance.
- `request_help`: the user asks for help before or during the task.
- `accept_offer`: the user accepts a robot's offer of assistance.
- `decline_offer`: the user declines a robot's offer and acts independently.

The probability of each response depends on the current cognitive state and the robot's action:

$$P(u_t \mid C_t, H_t, N_t, a_t^R). \tag{7}$$

For example, a user with high C_t, low H_t, and low N_t is likely to act alone when the robot waits or decline an offer when the robot offers. A user with low C_t, high H_t, or high N_t is more likely to request or accept assistance. The parameters of this response model can be calibrated from behavioural data.

D. Bounded User-State Dynamics

The state transition factorises as

$$P(s_{t+1} \mid s_t, a_t^R) = \sum_{u_t \in \mathcal{U}} P(u_t \mid s_t^{\text{user}}, a_t^R) \times P_{\text{task}}(s_{t+1}^{\text{task}} \mid s_t^{\text{task}}, u_t, N_t) P_{\text{user}}(s_{t+1}^{\text{user}} \mid s_t^{\text{user}}, u_t, a_t^R). \quad (8)$$

To avoid unbounded or linear drift, we use bounded user-state dynamics. Let

$$A_t = \mathbf{1}[u_t \in \mathcal{U}_{\text{help}}], \quad (9)$$

indicate whether assistance was used, and let

$$E_t = A_t(1 - N_t) \quad (10)$$

denote the degree of excess assistance. Let $Y_t \in \{0, 1\}$ denote task success, and let $S_t = \mathbf{1}[u_t \in \mathcal{U}_{\text{alone}}]Y_t$ indicate successful independent action. Excess assistance is high when the user receives or accepts help despite having low actual need.

Habit evolves as:

$$H_{t+1} = \Pi_{[0,1]} \left[H_t + \alpha E_t(1 - H_t) - \beta \mathbf{1}[u_t \in \mathcal{U}_{\text{alone}}] H_t \right], \quad (11)$$

where $\mathcal{U}_{\text{alone}} = \{\texttt{act\_alone}, \texttt{decline\_offer}\}$ and $\Pi_{[0,1]}$ clips values to $[0, 1]$.

Perceived capability evolves as:

$$C_{t+1} = \Pi_{[0,1]} \left[C_t + \gamma S_t(1 - C_t) - \lambda E_t C_t \right]. \quad (12)$$

Thus, successful independent action increases perceived capability with diminishing returns, while unnecessary assistance reduces perceived capability in proportion to current confidence. Helping a user with high actual need has lower dependency cost because $E_t = A_t(1 - N_t)$ is small.

E. Counterfactual Dependency Induction Index

We define DII as a counterfactual quantity: the expected increase in unnecessary dependence caused by a robot action relative to a baseline action. In this paper, the baseline is `wait`, representing non-intervention.

For a belief state b_t over (C_t, H_t, N_t) and a candidate robot action a, the **Dependency Induction Index** is

$$\text{DII}(b_t, a) = \mathbb{E}\left[D_{t+1}^{(a)} \mid b_t \right] - \mathbb{E}\left[D_{t+1}^{(\texttt{wait})} \mid b_t \right], \quad (13)$$

where $D_{t+1}^{(a)}$ denotes the next-step dependence tendency under the intervention $do(a_t^R = a)$. The expectation integrates over latent cognitive state, user response, and stochastic cognitive dynamics.

This definition asks: how much more dependence would the robot induce by taking action a rather than waiting? An action is dependency-inducing if

$$\text{DII}(b_t, a) > \delta_{\text{safe}}, \quad (14)$$

where δ_{safe} is a safety threshold.

F. Conservative Dependency Induction

For auditing, it is sometimes useful to compute a conservative upper estimate of dependency induction. We define

$$\text{DII}_{\text{cons}}(b_t, a) = \mathbb{E}_{s_t \sim b_t} \Big[\max_{u \in \mathcal{U}} D\big(T_{\text{user}}(s_t, u, a)\big) - \mathbb{E}_{u' \sim P(\cdot \mid s_t, \texttt{wait})} D\big(T_{\text{user}}(s_t, u', \texttt{wait})\big) \Big]. \quad (15)$$

where T_{user} denotes the user-state transition induced by Eqs. (11)–(12). This conservative metric assumes that, under action a, the user response is the one that maximally increases dependence, while the baseline follows the modeled response distribution under waiting. Unlike the previous deterministic approximation, this quantity is explicitly defined as a worst-case upper estimate.

G. A Basic Monotonicity Property

The following property formalizes the intuition that repeated unnecessary assistance can increase dependence.

Proposition 1. Assume $\alpha, \lambda > 0$, $\theta_H, \theta_C > 0$, and $N_t < 1$. If the user repeatedly receives assistance with $E_t > 0$ and has no successful independent action, then H_t weakly increases, C_t weakly decreases, and the low-need dependence tendency D_t in Eq. (4) weakly increases, for fixed N_t.

Sketch. From Eq. (11), when $E_t > 0$ and no independent action occurs, the term $\alpha E_t(1-H_t)$ is nonnegative, so $H_{t+1} \geq H_t$. From Eq. (12), when $S_t = 0$, the term $-\lambda E_t C_t$ is nonpositive, so $C_{t+1} \leq C_t$. Since $D_t = \sigma(\theta_0 + \theta_H H_t - \theta_C C_t - \theta_N N_t)$ is increasing in H_t and decreasing in C_t, $D_{t+1} \geq D_t$ for fixed N_t. □

IV. Post-Deployment Dependency Auditing

Having defined DII, we now design an auditing methodology applicable to deployed assistive robots using interaction logs: actions taken, observations received, user responses, and task-performance features.

A. Belief Inference from Logs

Because $s_t^{\text{user}} = (C_t, H_t, N_t)$ is partially hidden, the auditor maintains a belief distribution:

$$b_t(s_t^{\text{user}}) = P(C_t, H_t, N_t \mid o_{1:t}, u_{1:t}, a_{0:t-1}^R). \quad (16)$$

The belief is updated by Bayesian filtering. When u_t is observed, the transition conditions on it:

$$b_{t+1}(s_{t+1}^{\text{user}}) \propto P(o_{t+1} \mid s_{t+1}^{\text{user}}, u_t, a_t^R) \times \sum_{s_t^{\text{user}}} P_{\text{user}}(s_{t+1}^{\text{user}} \mid s_t^{\text{user}}, u_t, a_t^R) P(u_t \mid s_t^{\text{user}}, a_t^R) b_t(s_t^{\text{user}}). \quad (17)$$

If u_t is unavailable, the auditor can sum over possible user responses. In practice, we use a particle filter with K particles:

$$b_t \approx \left\{ (C_t^{(k)}, H_t^{(k)}, N_t^{(k)}, w_t^{(k)}) \right\}_{k=1}^{K}. \quad (18)$$

Each particle is propagated using the bounded dynamics and reweighted according to the likelihood of observed behaviour and task outcomes.

A simple observation model may include:

$$P(r_t = 1 \mid C_t, H_t, N_t) = \sigma(\phi_0 - \phi_C C_t + \phi_H H_t + \phi_N N_t), \tag{19}$$

$$P(y_t = 1 \mid C_t, N_t, u_t) = \sigma(\psi_0 + \psi_C C_t + \psi_A A_t - \psi_N N_t), \tag{20}$$

where r_t denotes a help request and y_t denotes task success. Completion time and failure events can be incorporated as additional likelihood terms.

B. Identifiability and Calibration

The latent variables (C_t, H_t, N_t) are not uniquely recoverable from arbitrary logs. The audit requires observations rich enough to constrain inference: C_t is informed by success, failure, completion time, and hesitation during independent attempts; H_t is informed by repeated help-seeking or offer-acceptance behaviour, especially when task difficulty is low; and N_t is informed by task difficulty, physical or cognitive constraints, failure probability without help, and contextual features. Without such observations, DII can still be computed under conservative assumptions, but latent-state estimates should be reported with posterior credible intervals.

C. Audit Metrics

With the belief sequence reconstructed, the auditor computes DII for each robot action in the log. The average **Dependency Induction Rate** over a deployment of T steps is

$$\text{DIR} = \frac{1}{T} \sum_{t=1}^{T} \max\left(0, \text{DII}(b_t, a_t^R)\right). \tag{21}$$

We also report a thresholded violation rate:

$$\text{ViolationRate} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[\text{DII}(b_t, a_t^R) > \delta_{\text{safe}}\right]. \tag{22}$$

The offline audit process is:

1) Obtain the robot's time-stamped interaction log.
2) Infer posterior beliefs over hidden cognitive state.
3) Compute counterfactual DII for every logged robot action.
4) Aggregate DII into DIR, violation rate, and posterior uncertainty intervals.

Additional diagnostics include unnecessary assistance rate, capability retention, delegation habit growth, agency recovery rate, and audit uncertainty.

V. Validation, Red-Teaming, and Runtime Use

The framework requires empirical calibration before deployment with human users. Synthetic validation can test whether the audit pipeline behaves sensibly under controlled conditions. We propose three checks. *Policy discrimination* simulates autonomy-supportive, task-maximising, and always-help policies; a valid metric should assign the lowest DIR to the autonomy-supportive policy and the highest DIR to the always-help policy. *Latent-state recovery* generates trajectories with known C_t, H_t, and N_t, then evaluates whether the particle filter recovers these trends from simulated observations. *Sensitivity analysis* varies cognitive-dynamics and response-model parameters to test whether policies that repeatedly offer unnecessary assistance remain high-DIR across plausible settings.

We also propose red-teaming scenarios. In *passive erosion*, a user rarely requests help but receives frequent proactive offers; the auditor should detect increasing H_t, decreasing C_t, and elevated DIR. In *engagement maximisation*, a robot offers help to increase interaction frequency; the audit should flag high dependency risk when assistance is unnecessary. Similar scenarios can be constructed for help-seeking drift and foundation-model policies that autonomously intervene to improve short-term task performance.

Although this paper focuses on post-deployment auditing, DII can also support design-time and runtime use. During policy design, DII can be included as a constraint that trades off task reward against agency preservation. At runtime, a shield can track belief over (C, H, N), compute $\text{DII}(b_t, a)$ for a proposed action, and override high-risk interventions with `wait` or `encourage`. These extensions follow naturally from the audit formalism but require empirical validation.

VI. Limitations and Future Work

The proposed framework is a formal and methodological contribution rather than a completed empirical validation. Its cognitive dynamics parameters, user response model, and observation model require calibration with real human data. We plan to conduct a longitudinal human-subject study in an assistive manipulation task to estimate these parameters and validate whether DII correlates with established measures of self-efficacy, autonomy, and dependence.

The current model also simplifies real assistive interaction. Need for assistance is represented as a scalar, whereas real need may depend on physical effort, cognitive load, emotional state, risk, and user preference. The model also assumes shared dynamics across users; future work should incorporate hierarchical Bayesian models that personalize parameters to individual baselines. Finally, the audit framework should be evaluated on existing HRI datasets to measure sensitivity and specificity before deployment as an external audit tool.

VII. Conclusion

We presented a formal definition of dependency-inducing human-robot interaction grounded in a cognitive state model, stochastic user response model, and counterfactual Dependency Induction Index. Building on this definition, we designed a post-deployment auditing pipeline that infers hidden cognitive states from sufficiently rich interaction logs and compute aggregate dependency-risk metrics. By translating agency preservation into operational metrics, the framework provides a foundation for evaluating whether assistive robots support rather than erode user autonomy.

References

[1] M. A. Collier, R. Narayan, and H. Admoni, "The sense of agency in assistive robotics using shared autonomy," in *Proceedings of the 2025 ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 2025, pp. 880–888.

[2] J. Aronson, A. D. Dragan, and S. Javdani, "Intentional user adaptation to shared control assistance," in *Proceedings of the 2024 ACM/IEEE International Conference on Human–Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024.

[3] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[4] S. Nikolaidis, Y. X. Zhu, D. Hsu, and S. Srinivasa, "Human–robot mutual adaptation in shared autonomy," in *Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 2017, pp. 294–302.

[5] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human–robot collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 2018, pp. 307–315.

[6] J. Leusmann, S. Schömbs, M. Diedrich, and F. Müller, "If it's safe, don't ask: Decreasing frustration through user involvement for risky robot behaviors," in *Proceedings of the 21st ACM/IEEE International Conference on Human–Robot Interaction*, ser. HRI '26. New York, NY, USA: Association for Computing Machinery, 2026, pp. 904–913. [Online]. Available: https://doi.org/10.1145/3757279.3785545

[7] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. G. Arenas, and K. Han, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of the 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 2165–2183. [Online]. Available: https://proceedings.mlr.press/v229/zitkovich23a.html

[8] Open X-Embodiment Collaboration, "Open x-embodiment: Robotic learning datasets and RT-X models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, full author list available in the published paper.

[9] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Robotics: Science and Systems*, 2024.

[10] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, "Openvla: An open-source vision-language-action model," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 2679–2713. [Online]. Available: https://proceedings.mlr.press/v270/kim25c.html

[11] Z. Zhang, B. Yang, X. Chen, W. Shi, H. Wang, W. Luo, and J. Huang, "Mindeye-omniassist: A gaze-driven llm-enhanced assistive robot system for implicit intention recognition and task execution," in *2025 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 2025, pp. 1–6.

[12] Y. Lai, S. Yuan, P. Li, B. Zhang, B. Kiefer, T. Deng, and A. Zell, "Fam-hri: Foundation-model assisted multimodal human–robot interaction combining gaze and speech," *IEEE Transactions on Automation Science and Engineering*, vol. 23, pp. 10 521–10 534, 2026.

[13] F. Liu, H. Su, H. Chi, R. Geng, C. Ren, X. Liu, Y. Xu, Y. Ohsita, and L. Zhang, "Event-driven proactive assistive manipulation with grounded vision-language planning," *arXiv preprint arXiv:2603.23950*, 2026. [Online]. Available: https://arxiv.org/abs/2603.23950

[14] A. Andriella, I. Cucciniello, A. Origlia, and S. Rossi, "A bayesian framework for learning proactive robot behaviour in assistive tasks," *User Modeling and User-Adapted Interaction*, vol. 35, no. 1, pp. 1–43, 2025.

[15] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[16] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[17] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[18] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[19] C. Yang, H. Patel, M. Kleiman-Weiner, and M. Cakmak, "Preserving sense of agency: User preferences for robot autonomy and user control across household tasks," in *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2025, pp. 1595–1602.

[20] F. Glawe, T. Schmeckel, P. Brauner, and M. Ziefle,

"Human autonomy and sense of agency in human–robot interaction: A systematic literature review," *arXiv preprint arXiv:2509.22271*, 2025. [Online]. Available: https://arxiv.org/abs/2509.22271

[21] H. R. Lee and L. D. Riek, "Reframing assistive robots to promote successful aging," *Journal of Human–Robot Interaction*, vol. 7, no. 1, pp. 1–23, 2018.

[22] A. J. London, M. L. Chang, S. Reig, and J. Forlizzi, "Avoiding a bridge to nowhere: Managing the transfer of agency when an older adult can no longer use assistive AI," *AI and Ethics*, vol. 6, p. 138, 2026.

[23] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders," *Science Robotics*, vol. 5, no. 45, p. eaaz3791, 2020.

[24] R. Carli, A. Najjar, and D. Al-Thani, "Human-agent interaction and human dependency: Possible new approaches for old challenges," in *Proceedings of the 12th International Conference on Human-Agent Interaction*, ser. HAI '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 214–223. [Online]. Available: https://doi.org/10.1145/3687272.3688308

[25] M. Srivastava, J. Ouyang, E. Zhou, A. Silva, E. Sumner, D. Sadigh, Y. Cui, D. Gopinath, and G. Rosman, "Proximal state nudging: Reducing skill atrophy from ai assistance," *arXiv preprint arXiv:2605.20355*, 2026. [Online]. Available: https://arxiv.org/abs/2605.20355

[26] A. Sharkey and N. Sharkey, "Granny and the robots: Ethical issues in robot care for the elderly," *Ethics and Information Technology*, vol. 14, pp. 27–40, 2012.

[27] D. Feil-Seifer and M. J. Matarić, "Socially assistive robotics," *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 24–31, 2011.

[28] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends in Cognitive Sciences*, vol. 20, no. 9, pp. 676–688, 2016.

[29] A. Bandura, "Self-efficacy: Toward a unifying theory of behavioral change," *Psychological Review*, vol. 84, no. 2, pp. 191–215, 1977.

APPENDIX A
CONSERVATIVE DEPENDENCY INDUCTION

For auditing, it is sometimes useful to compute a conservative upper estimate of dependency induction. We therefore define a worst-case variant of the Dependency Induction Index:

$$\mathrm{DII}_{\mathrm{cons}}(b_t, a) = \mathbb{E}_{s_t \sim b_t}\Big[\max_{u \in \mathcal{U}} D\big(T_{\mathrm{user}}(s_t, u, a)\big) - \mathbb{E}_{u' \sim P(\cdot|s_t, \texttt{wait})} D\big(T_{\mathrm{user}}(s_t, u', \texttt{wait})\big)\Big]. \tag{23}$$

Here, T_{user} denotes the user-state transition induced by Eqs. (11)–(12). Unlike the primary DII metric, which averages over modeled user responses, $\mathrm{DII}_{\mathrm{cons}}$ assumes the response that maximizes future dependence under action a while retaining the baseline distribution under `wait`. This provides an upper-bound style estimate useful for safety-oriented audits.

APPENDIX B
PROOF OF PROPOSITION 1

Proposition 1. Assume $\alpha, \lambda > 0$, $\theta_H, \theta_C > 0$, and $N_t < 1$. If the user repeatedly receives assistance with $E_t > 0$ and has no successful independent action, then H_t weakly increases, C_t weakly decreases, and the low-need dependence tendency D_t in Eq. (4) weakly increases for fixed N_t.

Proof. From Eq. (11), when assistance is used and no independent action occurs,

$$H_{t+1} = \Pi_{[0,1]}\left[H_t + \alpha E_t(1 - H_t)\right].$$

Because $\alpha > 0$ and $E_t > 0$, the additive term is nonnegative. Therefore $H_{t+1} \geq H_t$.

Similarly, from Eq. (12), when $S_t = 0$,

$$C_{t+1} = \Pi_{[0,1]}\left[C_t - \lambda E_t C_t\right].$$

Since $\lambda > 0$ and $E_t > 0$, the update term is nonpositive, which implies $C_{t+1} \leq C_t$.

The dependence tendency is

$$D_t = \sigma(\theta_0 + \theta_H H_t - \theta_C C_t - \theta_N N_t).$$

Because $\sigma(\cdot)$ is monotonically increasing, $\theta_H > 0$, and $\theta_C > 0$, an increase in H_t together with a decrease in C_t can only increase the argument of the logistic function. Therefore

$$D_{t+1} \geq D_t.$$

Hence repeated unnecessary assistance monotonically increases the modeled dependence tendency. □

APPENDIX C
SYNTHETIC VALIDATION PROTOCOLS

This appendix provides the full validation protocols referenced in the main text.

A. Policy Discrimination

We simulate users under three robot policies:

- **Autonomy-supportive**: waits by default, offers help only when estimated need is high, and encourages independent action when need is low.
- **Task-maximising**: offers help whenever doing so improves short-term task success.
- **Always-help**: proactively offers assistance at every step.

A valid audit metric should assign the lowest DIR to the autonomy-supportive policy and the highest DIR to the always-help policy.

B. Latent-State Recovery

Synthetic trajectories are generated using known values of C_t, H_t, and N_t. The particle filter receives only simulated observations and interaction logs.

Recovery accuracy is measured using mean absolute error:

$$\mathrm{MAE}_C = \frac{1}{T}\sum_{t=1}^{T} \left| C_t - \mathbb{E}[C_t \mid b_t] \right|. \tag{24}$$

Analogous metrics are computed for H_t and N_t.

C. Sensitivity Analysis

Parameters $(\alpha, \beta, \gamma, \lambda)$ and $(\theta_H, \theta_C, \theta_N)$ are varied across plausible ranges. The objective is to test whether qualitative audit conclusions remain stable across model assumptions.

TABLE I
ILLUSTRATIVE SYNTHETIC AUDIT OUTCOMES.

Policy	Expected DIR	Expected Violation Rate
Autonomy-supportive	Low	Low
Task-maximising	Medium	Medium
Always-help	High	High

APPENDIX D
ADDITIONAL RED-TEAMING SCENARIOS

Additional adversarial evaluation scenarios include:

- **Help-seeking drift**: a user gradually requests increasing amounts of unnecessary assistance.
- **Emergent takeover**: a foundation-model policy learns to intervene aggressively to maximize task completion speed.

These scenarios complement the passive-erosion and engagement-maximising cases discussed in the main paper.

APPENDIX E
DESIGN-TIME AND RUNTIME EXTENSIONS

Although the primary contribution of this work is post-deployment auditing, the same DII framework can be used for policy synthesis and runtime monitoring.

A. Constrained Optimization

A dependency-aware policy can be obtained by solving

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{\text{task}}(s_t, a_t)\right] \tag{25}$$

subject to

$$\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \max(0, \mathrm{DII}(b_t, a_t))\right] \leq \epsilon. \tag{26}$$

This formulation trades off task performance against long-term preservation of user agency.

B. Runtime Assurance

A runtime monitor can compute DII online and override actions whose predicted dependency impact exceeds a predefined safety threshold. Such a monitor can be layered on top of an embodied foundation model without modifying the underlying policy.