Suggestion-Based Assistance of Suboptimal Users in Sequential Decision-Making Tasks

Extended Abstract

Niklas Dieckow
Research Group on Human-Centric Machine Learning,
Hamburg University of Technology
Hamburg, Germany
niklas.dieckow@tuhh.de

Pierre-Alexandre Murena
Research Group on Human-Centric Machine Learning,
Hamburg University of Technology
Hamburg, Germany
pierre-alexandre.murena@tuhh.de

ABSTRACT

AI agents can assist humans by offering suggestions which users may accept or reject. This creates an *asymmetric collaboration* where the user retains full control while the AI lacks direct agency. We demonstrate that merely suggesting task-optimal actions can yield worse outcomes than unassisted performance; effective assistance requires understanding the user's decision-making. To address this, we propose a zero-shot method based on Bayesian estimation of the user's acceptance and fallback behavior, relying on a parametric model rather than prior data. We validate our approach theoretically and empirically on a novel toy environment, assessing its performance against baselines and stability in situations where parameters are incorrectly estimated.

KEYWORDS

Human-AI collaboration; zero-shot assistance; two-agent collaboration

ACM Reference Format:
Niklas Dieckow and Pierre-Alexandre Murena. 2026. Suggestion-Based Assistance of Suboptimal Users in Sequential Decision-Making Tasks: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/DFPG9276

1 INTRODUCTION

The democratization of Large Language Models has fostered applications where users rely on AI assistance in fields ranging from programming [17] to medicine [27] and law [22]. This raises the question of whether such agents provide effective assistance, even when theoretically efficient. We investigate a sequential decision setting where a *user* is advised by an *assistant*. While ideally the user follows suggestions, factors like mistrust, misunderstanding, preference, or incapability may cause them to deviate from recommendations. Our contributions are: (1) a formalization of suggestion-based AI assistance; (2) a demonstration that an optimal task policy is generally not the solution to the assistance problem; and (3) a method to obtain a solution for stationary user behavior, validated experimentally on a novel toy environment.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DFPG9276

2 RELATED WORK

Prior work has demonstrated that human-AI teams often underperform compared to individuals [1, 2, 6, 8, 24, 29], a trend particularly pronounced in decision-making tasks [28]. While some models propose user overrides [31], many works focus on single-decision contexts rather than sequential ones [6]. These works address issues such as classifier updates reducing collaborative performance [2], rejection of optimal advice [8], and bias reinforcement [29]. Unlike approaches the predict acceptance probabilities for single decisions [19] or providing confidence estimates [3, 30], our work addresses the long-term consequences of sequential decisions without placing restrictive assumptions on the drivers behind the user's adherence behavior. Thus, we take a similar perspective as [10], whose adherence-aware MDP framework is tightly related to ours, but does not cover user parameter inference. Shared autonomy [11, 16, 20] mostly investigates robot teleoperation using policy blending [7, 23], although some works also address sequential problems [13, 26]. Cognitive models describe user variability [12] or define team utility functions based on advice acceptance [1, 4]. Others suggest that acceptance depends on self-trust [14] or analyze single-decision contexts [15]. Our work relates closely to [4], where users optimize subjective rewards estimated via inverse reinforcement learning [18]. Alternatively, [5] considers users acting optimally regarding subjective transitions in a Stackelberg setting. We differ by estimating the user's policy directly, allowing for irrational behavior without requiring agent pre-commitment.

3 PRELIMINARIES

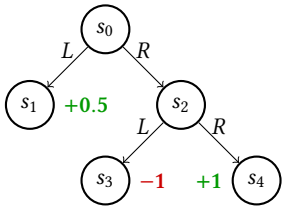


Figure 1: If the right action is not guaranteed in s_2, going to s_1 might be safer.

We model the user's task as an MDP $\mathcal{M} = (S, A, R, T, \gamma)$, utilizing standard definitions [25]. The assistant suggests actions $\tilde{a} \in A$ to a user who accepts with probability $\alpha(h_t)$, where h_t denotes the history at time t, or follows their own policy π_u. This results in the *effective policy* $\pi_\alpha^{(t)}(a \mid h_t) = \alpha(h_t)\pi_{\text{sup}}(a \mid h_t) + (1 - \alpha(h_t))\pi_u(a \mid s)$, whose return we aim to maximize.

Optimal Policy ≠ Optimal Assistance. Suppose that α is constant. Using the deterministic MDP in Figure 1, an optimal policy plays R twice in a row and yields $V^\star(s_0) = 1$. An inexperienced user may instead play L in s_0, and randomly

Algorithm 1 Belief assistant (stationary acceptance)

Input: $\mathcal{M}, \mu, p^{(0)}$, plan, n, k
$s \sim \mu, t \leftarrow 0, \pi^\star \leftarrow \text{plan}(\mathcal{M})$
while s is not terminal **do**
  for $1 \le i \le n$ **do**
    draw $(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i) \sim p^{(t)}$; obtain $\pi^\star_{\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i}$ (plan or cache)
  end for
  suggest $\tilde{a} \sim \frac{1}{n}\sum_{i=1}^{n} \pi^\star_{\boldsymbol{\theta}_i,\boldsymbol{\alpha}_i}(\cdot \mid s)$
  $a \leftarrow$ observed user action; $s \leftarrow$ result of step with a
  update $p^{(t+1)}$ with Eq. (2)
  $t \leftarrow t+1$
end while

choose between both directions if it ends up in s_2. This policy has an expected return of 0.5. If the assistant suggests actions according to $\pi^\star$, the resulting mixture policy has an expected return of $V^{\pi_\alpha}(s_0) = \frac{1-\alpha}{2} + \alpha^2$, which only improves upon the return of the user policy when $\alpha > \frac{1}{2}$, and may be lower than the user alone for $\alpha \in (0, \frac{1}{2})$.

Stationary Acceptance. For stationary α and a user policy parameterized by $\boldsymbol{\theta} \in \Theta$, we define a modified MDP $\mathcal{M}_{\boldsymbol{\theta},\alpha}$ with transitions

$$T_{\boldsymbol{\theta},\alpha}(s' \mid \tilde{a}, s) = \alpha T(s' \mid \tilde{a}, s) + (1-\alpha)\sum_{a \in A} \pi_u(a \mid s) T(s' \mid a, s). \quad (1)$$

Optimizing $\mathcal{M}_{\boldsymbol{\theta},\alpha}$ solves the assistance problem. From now on, we call an optimal policy of $\mathcal{M}$ a *task-optimal policy*, and an optimal policy of $\mathcal{M}_{\boldsymbol{\theta},\alpha}$ a *user-optimal policy*. The following lemma summarizes the theoretical foundation of our approach.

Lemma 3.1. *Let $\boldsymbol{\theta} \in \Theta$ be the parameters of a user policy and $\alpha\colon S \times A \to [0,1]$ an acceptance function. Define π_α as before, with π_{sup} an optimal policy of $\mathcal{M}_{\boldsymbol{\theta},\alpha}$. Then, in any state $s \in S$, we have for the value function of the original MDP, $V^\star(s) \ge V^{\pi_\alpha}(s) \ge V^{\pi_{\boldsymbol{\theta}}}(s)$.*

The proof has been omitted for space reasons but only relies on the definition of the optimal value function. This result requires that the user policy is known. However, in practice, we want to be able to assist a user without any prior knowledge about them, which is also known as *zero-shot assistance* [4].

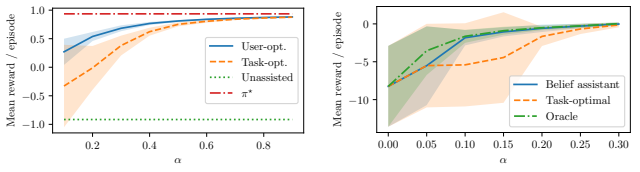
4 METHOD

Previously, we assumed known user parameters; here, we estimate them interactively via Bayesian updates. We assume α is stationary but parameterized by $\boldsymbol{\alpha} \in \mathbb{R}^m$.

We employ a generic planner, plan, to compute an initial optimal policy $\pi^\star$ for $\mathcal{M}$. At each step t, we sample user parameters from the belief $p^{(t)}$ and compute corresponding optimal policies. To reduce computation, we use $\pi^\star$ as a warm-start heuristic and store policies in a cache of size k. For tractable parameter spaces, we pre-train assistants on the set of user parameters. For larger spaces, one can utilize coarse subsets or on-demand training via meta-learning [9]. Given a suggestion $\tilde{a}_t$, the belief is updated using the likelihood of the observed action a_t:

$$p(a_t \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, h_t) = \alpha \mathbf{1}\{a_t = \tilde{a}_t\} + (1-\alpha)\pi_{\boldsymbol{\theta}}(a_t \mid s_t), \quad (2)$$

where $\alpha = \alpha(s_t, \tilde{a}_t)$ has been abbreviated. The procedure is summarized in Algorithm 1.



(a) Exp. 1: Known user (b) Exp. 3: Zero-shot / belief assistant

Figure 2: Mean episode rewards. (a) User-optimal assistance outperforms task-optimal for known users. (b) Belief assistant matches oracle performance for low α.

5 EXPERIMENTS

We evaluate our approach on a grid world navigation task where a user aims to reach a target (reward +1) with step costs of −0.01. The user is influenced by a *preference map* q assigning attraction or repulsion to landmarks. We use value iteration, although RL methods like PPO [21] are applicable. Our code is publicly available on GitHub: https://github.com/ndieckow/aamas-suggestion-assistance.

Exp. 1: Known User. Comparing unassisted, task-optimal, and user-optimal policies across 20 random preference maps, we found that user-optimal assistance yields higher average returns and lower variance than task-optimal assistance, particularly for low acceptance probabilities α (see Figure 2a).

Exp. 2: Sensitivity. Sensitivity analysis reveals that performance degrades most when the assistant significantly overestimates the user's acceptance probability ($\alpha' \gg \alpha$).

Exp. 3: Zero-shot Belief Agent. We evaluated the belief agent (Algorithm 1) on a handcrafted environment with strong repellant zones. For $\alpha \le 0.3$, the belief agent significantly outperformed task-optimal assistance, matching the oracle's performance (Figure 2b).

Exp. 4 & 5: Robustness and Decay. Against a random user, the belief assistant performed comparably to the task-optimal baseline, showing robustness to model mismatch. With exponentially decaying acceptance and the option to suggest nothing, the assistant effectively rationed suggestions to critical states.

6 LIMITATIONS AND FUTURE WORK

The primary computational bottleneck is the potential need to train models at each interaction step. While caching and pre-trained policies mitigate this, further optimization is required. Additionally, our evaluation was restricted to a toy environment. Future work must extend to realistic tasks and human testing to capture complexities of real-world behavior that our current setup may overlook.

Promising directions for future work include improved pre-training and continual learning, using interaction data to update the parameter prior and refine the policy cache for faster fine-tuning. Theoretical questions remain regarding the regret of task-optimal policies—determining when expensive personalization is truly necessary—and potential links to robust reinforcement learning. Finally, extending this framework to settings where not even the assistant can solve the problem alone is an ambitious goal.

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (2021), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[3] Nina Corvelo Benz and Manuel Rodriguez. 2023. Human-Aligned Calibration for AI-Assisted Decision Making. In *37th Conference on Neural Information Processing Systems*.

[4] Sebastiaan De Peuter and Samuel Kaski. 2023. Zero-Shot Assistance in Sequential Decision Problems. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v37i10.26365

[5] Christos Dimitrakakis, David C Parkes, Goran Radanovic, and Paul Tylkin. 2017. Multi-View Decision Processes: The Helper-AI Problem. In *Advances in Neural Information Processing Systems*, Vol. 30. https://doi.org/10.5555/3295222.3295296

[6] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *Proceedings of the 2022 ACM Conference on Fairness, Accountability and Transparency* (2022). https://doi.org/10.1145/3531146.3533221

[7] Anca D Dragan and Siddhartha S Srinivasa. 2013. A policy-blending formalism for shared control. *International Journal of Robotics Research* (2013).

[8] Avshalom Elmalech, David Sarne, Avi Rosenfeld, and Eden Erez. 2015. When Suboptimal Rules. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (2015). https://doi.org/10.1609/aaai.v29i1.9335

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.1703.03400

[10] Julien Grand-Clément and Jean Pauphilet. 2023. The Best Decisions Are Not the Best Advice: Making Adherence-Aware Recommendations. *Management Science* 72, 1 (Jan. 2023), 667–692. https://doi.org/10.1287/mnsc.2023.01851

[11] Ming Li, Yu Kang, Yun-Bo Zhao, Jin Zhu, and Shiyi You. 2022. Shared Autonomy Based on Human-in-the-loop Reinforcement Learning with Policy Constraints. In *Proceedings of the 41st Chinese Control Conference*.

[12] Zhuoran Lu, Syed Hasan Amin Mahmoo, Zhuoyan Li, and Ming Yin. 2024. Mix and match: Characterizing heterogeneous human behavior in ai-assisted decision making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 12. 95–104.

[13] Matthew B. Luebbers, Aaquib Tabrez, Kyler Ruvane, and Bradley Hayes. 2023. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. In *Robotics: Science and Systems*.

[14] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[15] Syed Hasan Amin Mahmood, Zhuoran Lu, and Ming Yin. 2024. Designing Behavior-Aware AI to Improve the Human-AI Team Performance in AI-Assisted Decision Making. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3106–3114.

[16] Brandon J. McMahan, Zhenhao Peng, Bolei Zhou, and Jonathan C. Kao. 2024. Shared Autonomy with IDA: Interventional Diffusion Assistance. In *38th NeurIPS Proceedings*.

[17] Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2025. LLMs for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process* 37, 1 (2025), e2723.

[18] Andrew Y. Ng and Stuart Russell. 2000. Algorithms for Inverse Reinforcement Learning. *Proceedings of the 17th International Conference on Machine Learning* (2000). https://doi.org/10.5555/645529.657801

[19] Gali Noti and Yiling Chen. 2022. Learning When to Advise Human Decision Makers. In *International Joint Conference on Artificial Intelligence*.

[20] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. 2018. Shared Autonomy via Deep Reinforcement Learning. arXiv:1802.01744

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347

[22] Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access* (2025).

[23] Saurav Singh and Jamison Heard. 2023. Probabilistic Policy Blending for Shared Autonomy using Deep Reinforcement Learning. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

[24] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999). https://doi.org/10.1006/ijhc.1999.0252

[25] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction, Second edition.* The MIT Press. https://doi.org/10.5555/3312046

[26] Aaquib Tabrez, Ryan Leonard, and Bradley Hayes. 2025. Single-shot policy explanation to improve task performance via semantic reward coaching. *Neural Computing and Applications* (2025).

[27] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology* 19, 1 (2024), 43.

[28] Michael Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8 (2024), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

[29] Michelle Vaccaro and Jim Waldo. 2019. The Effects of Mixing Machine Learning and Human Judgment. *Commun. ACM* 62, 11 (2019). https://doi.org/10.1145/3359338

[30] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. 2022. Uncalibrated Models Can Improve Human-AI Collaboration. In *36th Conference on Neural Information Processing Systems*.

[31] Mustafa Mert Çelikok, Frans A. Olihoek, and Samuel Kaski. 2022. Best-Response Bayesian Reinforcement Learning with Bayes-adaptive POMDPs for Centaurs. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. https://doi.org/10.48550/arXiv.2204.01160