Artificial Theory of Mind in Human-in-the-Loop

Sammie Katt
ELLIS Institute Finland
Department of Computer Science
Aalto University, Finland
sammie.katt@aalto.fi

Samuel Kaski
ELLIS Institute Finland
Department of Computer Science,
Aalto University, Finland
Department of Computer Science,
University of Manchester, United Kingdom
samuel.kaski@aalto.fi

Abstract

Thanks to the advances in artificial intelligence (AI), interactive human-AI applications are growing explosively. A common assumption in these systems is that the humans provide ground-truth (oracle) data during interactions. This is seen, for example, when fine-tuning large language models with human feedback or in personalized recommendation-systems. However, it is well-known that human users often do not act like oracles which implies they, instead, should be represented more realistically instead. In this work, we propose a preliminary framework for user models for human-in-the-loop (HITL) problems. In particular, we first define a general decision-making problem statement for HITL which explicitly includes user models, with a focus on how they may reason about the AI they are interacting with. We then derive user models for HITL from simple but powerful assumptions about the user, and show the implications empirically in a Bayesian optimization and recommendation system setting. Through this lens, we discuss how assuming humans are oracles can lead to bias under several concrete settings.

1 Introduction

Thanks to recent developments in artificial intelligence (AI) and machine learning (ML), human-AI interactive systems are becoming more widespread and powerful. A well-known subclass of these systems, called "human-in-the-loop" (HITL), considers humans as a source of data (fig. 1). Examples include the well-known personalized recommendation systems [33], active learning or optimization of human preferences [28, 19, 30], and also precision medicine systems [35]. Most notably, LLM fine-tuning with reinforcement learning from feedback also falls in this category [10]. The underlying assumption in these approaches is that the human is an *oracle* and provides ground-truth labels.

However, even when in rather trivial instances and instructed to provide true labels, oracles are poor models of human users. First, humans may simply not know the right answer or, more likely, are biased [24]. More relevant for us, even when that is accounted for, users adjust their answers on purpose to guide the interaction [2, 11]. For example, it is known that users maintain a mental model of the AI with which they make predictions of the behavior of the system [9, 37], a phenomenon we propose to call *artificial theory of mind*.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Metacognition in Generative AI.

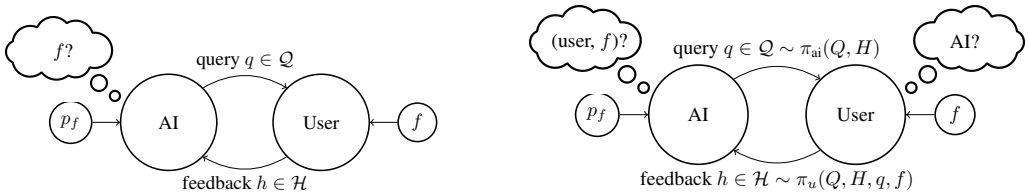

Figure 1: Model of human-in-the-loop systems. The AI system communicates by giving queries, whereas the human gives feedback. Left shows the typical setting. Right is our proposed modification: we argue the user reasons about the AI's state and gives feedback to intentionally change it.

Example: let us say we are interested in finding the best hotel for a user, and do this by asking the user to score hotels. The assumption behind this work is that users will provide scores that (they believe) move the system towards some desired state, rather than necessary giving the ground-truth labels. For instance, a user may give a mediocre score to a poorly performing hotel, *if the user believes it will 'help'*: maybe the hotel is undesirable, due to some specific property, but *similar* hotels have high scores and the user wants to avoid steering the system away from those options.

This work makes steps towards methods that include more realistic user models for human-in-the-loop problems (fig. 1, right). Starting from a general view of human-in-the-loop setting, we formalize a framework for explicitly defining, deriving, and inserting user models. From there, we focus on modeling users who consider the state of the system, motivated by ideas grounded in modeling users as computationally rational agents. Then, we consider Bayesian optimization and recommendation systems as concrete settings, and investigate the behavior of our proposed user models and their effect on the system.

2 Preliminaries

We consider a minimal setup for human-AI collaboration where there is no state or dynamics except for the interaction between the human and the AI, as depicted in fig. 1. This is a sequential process in which the user and AI iteratively communicate, where the only assumption is that there is *some* function $f \in \mathcal{F}$, $f : \mathcal{X} \to \mathcal{Y}$, of interest driving the interaction.

This process is defined by the AI's query space $\mathcal{Q}$, the user's feedback space $\mathcal{H}$, and some function of interest f according to which the user gives feedback $h \in \mathcal{H}$ on some query $q \in \mathcal{Q}$. Generally, we assume there is a (mutual) objective $\mathrm{U} : (\Delta\mathcal{F}, \mathcal{F}) \to \mathbb{R}$, which assigns values (score) to the AI's belief over the function $p_f(f)$, given the actual function. The ultimate goal, and performance metric, is determined by this utility. The AI's belief corresponds to the posterior given the likelihood of the observed query-label data, (Q, H), and the prior p_f:

$$p_f(f \mid Q, H) \propto p_f(f) p(H \mid Q; f) = \underbrace{p_f(f)}_{\text{prior}} \prod_i \underbrace{p(h_i \mid q_i; f)}_{\text{likelihood}}, \tag{1}$$

where the likelihood typically Gaussian around the ground truth $p(h \mid q; f) = \mathcal{N}(f(x), \sigma)$ (e.g. [19, 30, 14]); the assumption that we explicitly challenge in this work.

2.1 Common Concrete Instances

While minimal, this setting represents several important settings, including learning human preferences, maximizing them, learning in recommendation systems, as well as the recently popular (reward learning in) reinforcement learning from human feedback. Here, we discuss two settings used in the empirical evaluation.

Bayesian optimization of Human Preferences: in this setting, we are interested in finding elements $x \in \mathcal{X}$ that optimizes the user's preferences f. Since these are typically impossible for users to define mathematically, they must be learned (and maximized) from data. In Bayesian optimization (BO),

the AI queries data points $\mathcal{Q}_{\text{BO}} \triangleq \mathcal{X}$, which the user then scores $\mathcal{H}_{\text{BO}} \triangleq \mathcal{Y}$. The utility of the AI's belief is the true value of the AI's best guess of the optimal:

$$\text{U}_{\text{BO}}(p(f), f) \triangleq f(\arg\max_x \mathbb{E}_{f' \sim p(f)}\left[f'(x)\right]) \tag{2}$$

A solution in BO is a method for tracking the posterior over f, commonly done with Gaussian processes [36], combined with an *acquisition function* that — when optimized — determines which query to pick. Typical acquisition functions are upper confidence bound (UCB) and expected improvement (EI) [16].

Learning in Recommendation Systems: in recommendation systems, such as those employed by Spotify, the AI typically *does not query*: $\mathcal{Q}_{\text{RS}} \triangleq \emptyset$ [33]. Instead, the user is free to provide any labels on any element $\mathcal{H}_{\text{RS}} \triangleq (\mathcal{X}, \mathcal{Y})$, where we assume the labels are binary ("like/not like", i.e. $\mathcal{Y} \triangleq \{-1, 1\}$) given the user's internal scoring function f. The utility is some classification performance metric, such as accuracy over test data X:

$$\text{U}_{\text{RS}}(p(f), f) \triangleq \sum_{x \in X} \mathbb{E}_{f' \sim p(f)}\left[\mathbb{I}_{f(x)}\left(f'(x)\right)\right] \tag{3}$$

A typical solution to this problem is often not Bayesian, such as logistic regression[22].

2.2 Problem: User Modeling

When inferring the latent function f from data in human-AI systems, one must make a choice on how to model the user's decision process. For example, a Bayesian approach requires a likelihood function. A common solution is to "ignore" the user component, and pretend we are directly modeling the underlying function. In reinforcement learning from preferences, for example, we assume the user's response — ranking of possible outputs — follows the ground truth preference of the user. Similarly, when maximizing for user preferences with Bayesian optimization, we assume the user scores queries according to the Boltzmann distribution given the true utility function. We argue this is *not* what happens in practice, leading to a misspecification and thus estimation errors.

Examples Colella et al. [11] show that human users do not faithfully report the (true) scores $f(x)$ when interacting with a Bayesian optimization system. Instead, users *steer* the system by providing feedback that, presumably, (they believe) moves (the maximum of) the posterior more closely to that of (the maximum of) the true function.
Similarly, we can expect users to not provide arbitrary labels in a recommendation system. Instead, they provide those that (they expect) create some meaningful change in the recommendations.

Concretely, the feedback likelihood in eq. (1) which, universally is assumed to follow f with Gaussian noise, is wrong. Instead, the *users are autonomous agents* in this interactive setting, and their feedback can be considered actions that optimize for some utility given their beliefs over the system that they are interacting with.

Computational Rationality To formalize this intuition, we adopt the idea of computational rationality (CR), which proposes that humans can be modeled as (bounded) utility maximizers [18, 26]. In practice, CR formalizes the user's task as a decision process and compute the values of its optimal solution (with e.g. reinforcement learning). To create a user model, it is typically assumed the user picks actions $h \in \mathcal{H}$ proportional to their utility U_u using the Boltzmann-rational model [23, 38]:

$$\pi_u(h) = \frac{\exp\Big(\tau \text{U}_u(h)\Big)}{\sum_{h' \in \mathcal{H}} \exp\Big(\tau \text{U}_u(h')\Big)} \propto \exp\Big(\tau \text{U}_u(h)\Big), \tag{4}$$

where temperature τ controls stochasticity. Computational rationality has already seen success in human-AI interactive systems [31]; our contribution is its realization to the human-in-the-loop setting.

3 Proposal: Solution for HITL with Artificial ToM

Our solution is two-fold. First, we introduce a new likelihood that models how users take actions with the intention of influencing the (AI) system. Second, we derive the novel (Bayesian) inference problem and resulting policy of the AI.

3.1 User Model

We introduce two components (visualized on the right in fig. 1). First, we assume the user knows that the AI maintains a posterior over "the function" $p_f(f \mid Q, H)$, given some prior over f. We describe the prior which *the user believes is assigned to the AI* with p_{AI}, and abuse notation by denoting $p_{\text{AI}}(f \mid \dots)$ as a posterior given prior p_{AI} . Second, we assume the user has some objective $\mathrm{U}_u : (\Delta F, F) \to \mathbb{R}$ that they optimize for, with respect to this posterior. This objective depends on the task, and can be infeasible to compute in practice (both for computers and humans) and, so, we allow for myopic approximations and heuristics. We adopt the CR approach and model the user as optimizer of expected utility U_u given their model of the AI p_{AI} with the Boltzmann model (eq. (4)):

$$\pi_u(h \mid Q, H, q, f) \propto \exp\Big(\tau \mathrm{U}_u(p_{\text{AI}}(Q \cup q, H \cup h), f)\Big) \tag{5}$$

To summarize, we formalize a user with the tuple $(f, p_{\text{AI}}, \mathrm{U}_u)$, where p_{AI} is the user's estimation of the AI's belief $p(f)$, and eq. (5) is the user feedback likelihood given objective U_u and replaces the likelihood in eq. (1).

3.2 AI Inference

As originally, the AI's aims to maximize the original task U, which includes learning f. Originally, this means computing the Bayesian posterior, assuming a Gaussian process prior and Gaussian likelihood (eq. (1)). Given the new likelihood derived above, the posterior now becomes

$$\begin{aligned} p(f \mid Q, H) &= \frac{p(H \mid Q, f)p(f)}{p(H \mid Q)} \quad \propto \quad p(H \mid Q, f)p(f) \\ &= p_f(f) \prod_i \underbrace{\pi_u(h_i \mid Q_{<i}, H_{<i}, q_i, f)}_{\text{eq. (5)}} \end{aligned} \tag{6}$$

The key observations are that we do not assume each data-point is i.i.d. (compare eqs. (1) and (6)) and that the posterior no longer has a closed-form solution.

A solution to the human-AI problem is an AI policy π_{ai} that maximizes the utility function U given a posterior over the user's internal latent function eq. (6). In this preliminary work, we propose to adopt acquisition functions associated with the original task $\alpha : (\Delta\mathcal{F}, \mathcal{Q}) \to \mathbb{R}$, such as UCB for BO, and extend these to our novel posterior:

$$\pi_{\text{AI}}(H, Q) = \arg\max_q \alpha(\underbrace{p(f \mid H, Q)}_{\text{eq. (6)}}, q) \tag{7}$$

4 Empirical Evaluation: Proof of Concept

This work is still in progress and the empirical evaluation is limited. However, below we present empirical insights in two settings. For each setting, we develop user models that reason about the system and give feedback with intentions of achieving the collaborative goal, and show they generate reasonable (human) behavior. We show that the performance of typical (AI) solutions may vary widely, depending on the accuracy of the theory of mind of these user models. This means, in practice, the accuracy of our deployed systems may similarly be suffering.

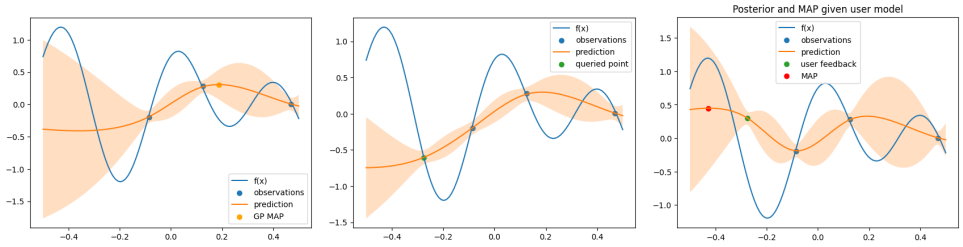

Figure 2: (**left**) Prior in a 1-dimensional Bayesian optimization problem. The true function f is shown in blue, whereas the orange is the posterior (mean and uncertainty) with orange dot being the maximum of the posterior mean, given the observations in blue. (**center**) Posterior of next time step given ground truth label. (**right**) Posterior of next time step given label from user model. This label leads to a more desirable posterior, which we claim simulates behavior of an expert user.

4.1 Bayesian Optimization

Setting We consider the running example of Bayesian optimization where the AI observes the user's score $\mathcal{H}_{\text{BO}} = \mathcal{Y}$ given queries $\mathcal{Q}_{\text{BO}} = \mathcal{X}$, and is interested in finding the optimizing value $\arg\max_x f(x)$ (recall section 2.1). We consider two (myopic) utilities that the user may optimize:

1. To model a user who is interested in explicitly steering the posterior, we model them minimizing the distance between the true maximum and that of the AI's posterior mean.

$$\mathrm{U}_u^{\text{argmax-dist}}(p(f), f) = -\left|\arg\max_x f(x) - \arg\max_x \mathbb{E}_{p(f)}\left[f(x)\right]\right| \tag{8}$$

2. To represent a user focused on pure performance, we propose a model that maximizes the true value of the maximum of the posterior mean:

$$\mathrm{U}_u^{\text{regret}}(p(f), f) = f\left(\arg\max_x \mathbb{E}_{p(f)}\left[f(x)\right]\right) \tag{9}$$

Results First, we study the behavior of our first user model (eq. (8)) in a salient situation (fig. 2). This setting is one where we argue users may recognize the result of steering to positively guide the posterior to a more useful posterior, one which assigns higher value to the true maximum of f. The left and center figure shows the prior, and the posterior assuming oracle feedback. The right figure shows the response of our user model. This scenario showcases that our user model behaves *naturally*, in that it returns false feedback at the benefit of positively steering the optimization process.

In our second experiment, we evaluate UCB with Gaussian processes on the function plotted in fig. 2. Figure 3 (left) plots the regret given oracle or user model feedback (eq. (2)), where we initiate our user model with either a correct ("our user model") or incorrect ("misspecified user model") prior p_{AI}. We see that our user model is able to improve performance when it has the correct assumptions on the system it is interacting with, but harms this process when provided with faulty assumptions. This showcases the type of behavior we expect and would like to resolve in real world scenarios.

Lastly, fig. 3 (right) shows the exact output of the two proposed user models, eqs. (8) and (9), for the whole query space. Despite the promising results so far, their behavior is not universally intuitive which highlights that additional work is necessary to investigate the realism of these user models.

4.2 Learning in Recommendation Systems

Setting Here, we emulate a recommendation system that is interested in modeling a user's preference given their binary feedback. The feedback is generated by an underlying scoring function f), with a threshold to return either -1 or 1 (recall section 2.1 and see left in fig. 4). Notably, the AI is *unable to query the user*, but passively receives query-label pairs $\mathcal{H}_{\text{RS}} = (\mathcal{X}, \{-1, 1\})$. We argue that the user *picks* which queries to label intentionally, given their running estimate of the system. We simplify the AI and focus on a non-Bayes setting: logistic regression. In particular, we assume that the base model fits a logistic regression model on the data and the user is aware of this. Hence, we

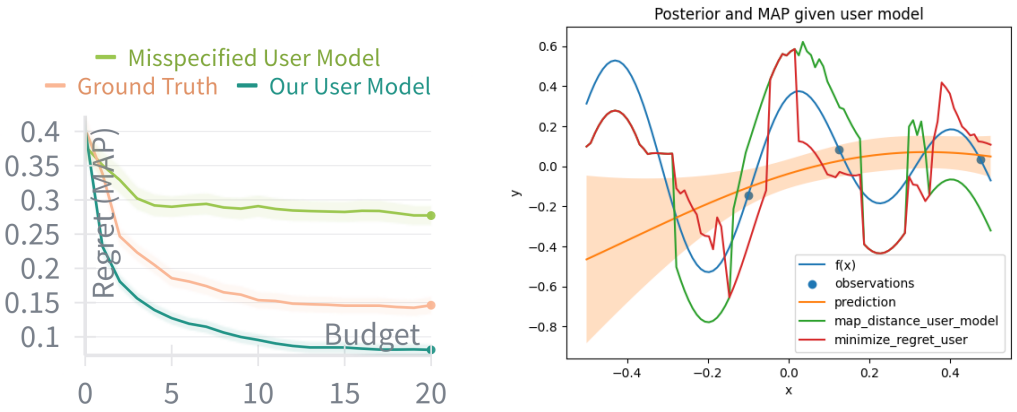

Figure 3: (left) Regret of BO (UCB) given "ground truth" or "ours" helpful user model. Our user model is able to give feedback such that the regret is much lower than ground truth. (right) Feedback of several user models given x. The blue line here is the "ground truth" model, whereas red and green are respectively the regret-minimizer and argmax-dist user models defined in eqs. (8) and (9). More work must be done to investigate how realistic these models are.

skip the user definition of the pair $(p_{\text{AI}}, \text{U}_u)$, and directly model the user's policy π_u (eq. (5)). We compare the following (probabilistic) behaviors:

1. We argue users may attempt to be explicit and pick the picks the best and worst examples:
$$\pi_u^{\text{extreme}}(x) \propto |f(x)| \tag{10}$$
2. We model users who may try to explicitly help identify the decision boundary by picking those around it. This decision boundary can either be the user's (f) or the AI's:
$$\pi_u^{\text{f-boundary}}(x) \propto \frac{1}{|f(x)|}, \quad \pi_u^{\text{ai-boundary}}(x) \propto \frac{1}{\mathbb{E}_{p(f)}\left[|f(x)|\right]} \tag{11}$$
3. Lastly, we model user who label the best and "best of worst". This models users who label the *best and worst content that is shown by the recommendation system* and, we argue, reflects realistic labeling behavior.
$$\pi_u^{\text{best}}(x) \propto \text{eq. (10) if } f(x) > 0, \text{ else eq. (11)} \tag{12}$$
4. The baseline is a random labeling strategy: $\pi_u^{\text{random}}(x) = \frac{1}{N}$

Results We test the accuracy of a typical classification (logistic regression) system given the proposed labeling strategies in fig. 4 (right). While some strategies lead to performance similar to the (typically assumed) random labeling strategy, some labeling strategies severely affect the sample efficiency. In particular, users that label the best and worst content shown to them (eq. (12)), a strategy we believe is quite reasonable, has particularly poor performance.

5 Related Work

Closely related to this work is student-teacher settings, in which an artificial agent is teaching or learning from a human user [32, 20]. A very interesting work looks at higher-order models for AI learning from human feedback in an active learning setting [25]. Under some additional assumptions, their work could be seen as a specific instance of ours, in particular when the objective is one agent *learning* a model from the other.

Work that shows evidence for human steering interactions is crucial to our motivation. In addition to the Bayesian optimization setting [11] previously mentioned, "COACH" [29] shows users adapt their advice to the current policy of a reinforcement learning agent. Research has also shown that (not how) users learn mental models of their assistant in AI-assisted human decision making [6].

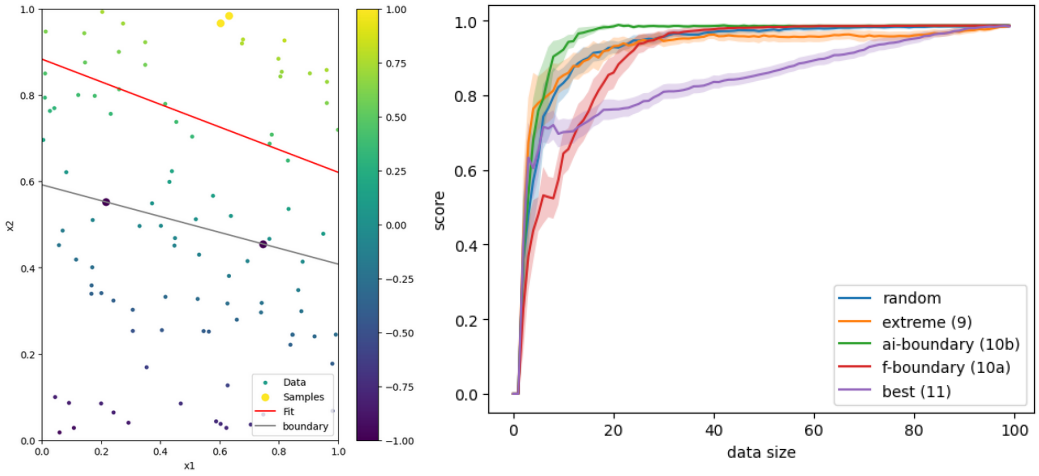

Figure 4: (**left**) Classification problem of 2-dimensional problem; colored small dots is (unlabeled) data set and their true scores, whereas bigger yellow ($y = 1$) & purple ($y = -1$) are labeled data. The AI's logistic regression boundary is displayed in red, while the gray line is the true boundary line. (**right**) Accuracy of trained classifiers given different labeling strategies, average over multiple run (shaded areas are 2 times standard error estimates). Computes the percentage of correctly labeled data points (y-axis) given number of labeled data (x-axis), given different user model labeling strategies.

Work on modeling agents in multi-agent settings typically take a more abstract view — they rarely care for modeling *human* agents — but nevertheless are a rich source of ideas [1, 20], including those on theory-of-mind [12, 21]. These have been successfully applied to, for example, human-robot navigation [8, 27]. Their work is interesting, but typically simplifies the aspects of theory of mind that are highlighted here — highly complex theory of mind models — and instead focus other complicating dimensions of sequential decision making and non-collaborative settings.

Other related topics include methods for designing HITL systems [15, 13, 34], inference over rational agents given their behavior [4, 39], and computational rationality as explanation for human behavior [18, 26, 31]. Lastly, literature that shows Bayesian reasoning in humans is particularly important motivation for modeling in this work [7, 5, 17, 3]. These works do not directly tackle human-in-the-loop settings, but have similar motivations and ideas in modeling humans.

6 Conclusion

Applications involving human-in-the-loop interactions are expanding rapidly, thanks to advancements in generative AI. However, a key challenge remains: how to effectively simulate, model, and interpret the most crucial component—the users themselves. In this work, we explore the concept of *artificial* theory of mind, where users explicitly reason about the (consequences of their feedback on the) systems they interact with. We argue that users provide feedback with the intent to benefit the overall system. We formalize this phenomenon in the form of a decision process and derive concrete instances on Bayesian optimization and recommendation systems. Finally, we examine empirically how different user models lead to drastically different behaviors and system performance, confirming our concerns regarding user misspecification and highlighting the importance of our proposed solution.

Acknowledgments and Disclosure of Funding

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, Grant 359207), ELISE Networks of Excellence Centres (EU Horizon:2020 grant agreement 951847), and UKRI Turing AI World-Leading Researcher Fellowship (EP/W002973/1). We acknowledge the research environment provided by ELLIS Institute Finland. We also acknowledge the computational resources provided by the Aalto Science-IT Project from Computer Science IT and CSC–IT Center for Science, Finland.

References

[1] S. V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[2] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[3] D. Arumugam, M. K. Ho, N. D. Goodman, and B. Van Roy. Bayesian reinforcement learning with limited cognitive load. *arXiv preprint arXiv:2305.03263*, 2023.

[4] A. Atrash and J. Pineau. A bayesian reinforcement learning approach for customizing human-robot interfaces. In *International Conference on Intelligent User Interfaces*, pages 355–360, 2009.

[5] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

[6] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11, 2019.

[7] A. Borji and L. Itti. Bayesian optimization explains human active search. *Advances in Neural Information Processing Systems*, 26, 2013.

[8] C. Brooks and D. Szafir. Building second-order mental models for human-robot interaction. *arXiv preprint arXiv:1909.06508*, 2019.

[9] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, and D. Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017.

[10] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] F. Colella, P. Daee, J. Jokinen, A. Oulasvirta, and S. Kaski. Human strategic steering improves performance of interactive optimization. In *Conference on User Modeling, Adaptation and Personalization*, pages 293–297, 2020.

[12] P. Doshi, X. Qu, A. Goodie, and D. Young. Modeling recursive reasoning by humans using empirically informed interactive pomdps. In *Autonomous Agents and MultiAgent Systems*, pages 1223–1230, 2010.

[13] A. Dubey, K. Abhinav, S. Jain, V. Arora, and A. Puttaveerana. Haco: a framework for developing human-ai teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference)*, pages 1–9, 2020.

[14] A. M. Elkahky, Y. Song, and X. He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th international conference on world wide web*, pages 278–288, 2015.

[15] J. A. Fails and D. R. Olsen Jr. Interactive machine learning. In *International Conference on Intelligent User Interfaces*, pages 39–45, 2003.

[16] P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[17] T. Gao, C. L. Baker, N. Tang, H. Xu, and J. B. Tenenbaum. The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive science*, 43(8):e12775, 2019.

[18] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

[19] J. González, Z. Dai, A. Damianou, and N. D. Lawrence. Preferential bayesian optimization. In *International Conference on Machine Learning*, pages 1282–1291. PMLR, 2017.

[20] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[21] T. Hedden and J. Zhang. What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, 85(1):1–36, 2002.

[22] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[23] H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

[24] D. Kahneman, P. Slovic, and A. Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

[25] O. Keurulainen, G. Alcan, and V. Kyrki. The role of higher-order cognitive models in active learning. *arXiv preprint arXiv:2401.04397*, 2024.

[26] R. L. Lewis, A. Howes, and S. Singh. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2):279–311, 2014.

[27] Y. Liao, M. Cao, X. Xu, and L. Xie. Atom: Adaptive theory-of-mind-based human motion prediction in long-term human-robot interactions. *arXiv preprint arXiv:2502.05792*, 2025.

[28] Y.-C. Liao, J. J. Dudley, G. B. Mo, C.-L. Cheng, L. Chan, A. Oulasvirta, and P. O. Kristensson. Interaction design with multi-objective bayesian optimization. *IEEE Pervasive Computing*, 22(1):29–38, 2023.

[29] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the International Conference on Machine Learning*, pages 2285–2294. Proceedings of Machine Learning Research, 2017.

[30] P. Mikkola, O. A. Martin, S. Chandramouli, M. Hartmann, O. Abril Pla, O. Thomas, H. Pesonen, J. Corander, A. Vehtari, S. Kaski, et al. Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis*, 19(4):1129–1161, 2024.

[31] A. Oulasvirta, J. P. Jokinen, and A. Howes. Computational rationality as a theory of interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[32] T. Peltola, M. M. Çelikok, P. Daee, and S. Kaski. Machine teaching of active sequential learners. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] I. Portugal, P. Alencar, and D. Cowan. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97:205–227, 2018.

[34] G. Ramos, C. Meek, P. Simard, J. Suh, and S. Ghorashi. Interactive machine teaching: a human-centered approach to building machine-learned models. *Proceedings of the ACM on Human-Computer Interaction*, 35(5-6):413–451, 2020.

[35] I. Sundin, T. Peltola, L. Micallef, H. Afrabandpey, M. Soare, M. Mamun Majumder, P. Daee, C. He, B. Serim, A. Havulinna, et al. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, 34(13):i395–i403, 2018.

[36] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[37] R. Williams, H. W. Park, and C. Breazeal. A is for artificial intelligence: the impact of artificial intelligence activities on young children's perceptions of robots. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[38] T. Yamagata, T. Oberkofler, T. Kaufmann, V. Bengs, E. Hüllermeier, and R. Santos-Rodriguez. Relatively Rational: Learning Utilities and Rationalities Jointly from Pairwise Preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, volume 2024, Vienna, Austria, 2024. PMLR.

[39] T. Zhi-Xuan, J. Mann, T. Silver, J. Tenenbaum, and V. Mansinghka. Online bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems*, 33:19238–19250, 2020.