Theory of Mind in Human-in-the-Loop

Sammie Katt' & Sami Kaski'*

'ELLIS Institute Finland
'Aalto University, Finland
*University of Manchester, United Kingdom

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. 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.

User Model

A user is defined by:

- their belief over the AI p_{AI}, and
- their objective U_u, predefined by the task and *dependent on the AI's state.*

AToM

Assumption: user knows the AI is fitting a model and approximates this process. For example AI fits a posterior based on (X, Y) data.

$$p(f \mid X, Y) \propto p(f) \prod_i \mathcal{N}(y_i; \mu = f(x_i), \sigma = \epsilon) \tag{1}$$

User approximates this behavior, denoted by p_{AI}.

User Policy

User picks feedback *with intention*, denoted by utility U_u. Using Boltzmann utility model:

$$\pi_u(h \mid Q, H, q, f) \propto \exp\Big(\tau \text{U}_u(p_{\text{AI}}(Q \cup q, H \cup h), f)\Big) \tag{2}$$

AI Solution

Preliminary idea: user model as likelihood function:

$$\begin{aligned} p(f \mid Q, H) &= \frac{p(H \mid Q, f)p(f)}{p(H \mid Q)} \propto p(H \mid Q, f)p(f) \\ &= p_f(f) \prod_i \underbrace{\pi_u(h_i \mid Q_{<i}, H_{<i}, q_i, f)}_{\text{user model}} \end{aligned} \tag{3}$$

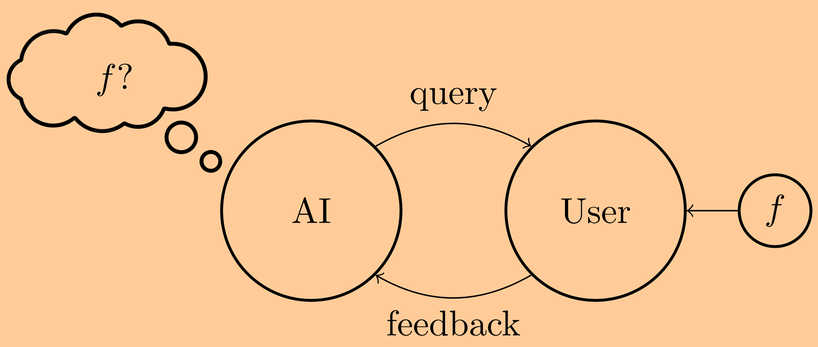
Note: *not i.i.d.*

Aalto University
School of Science

FCAI Finnish Center for Artificial Intelligence

Understanding **Users** in **Human-in-the-Loop** as **Decision-Makers** with **Theory of Mind**

ellis INSTITUTE FINLAND



Current Paradigm

The user is an oracle:

$$y = f(x) + \epsilon$$

+ Easy inference.
− Unrealistic.
− Limited applicability.

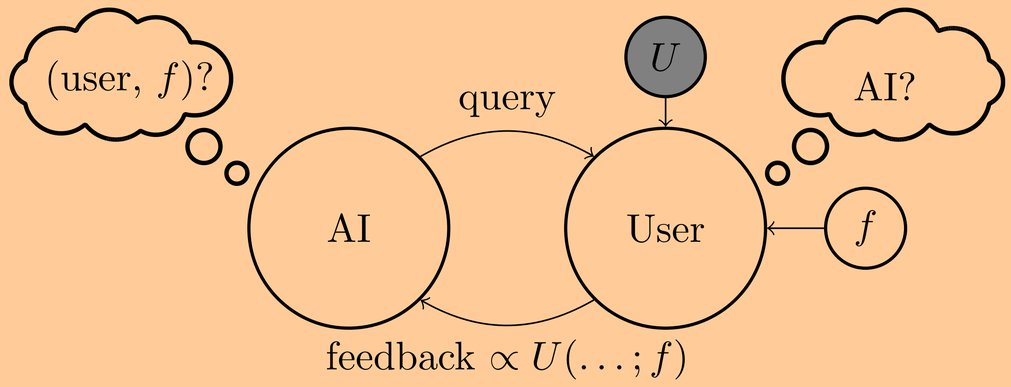
This Work

User is *intentional*:

$$\pi_u(\cdot \mid p_{\text{AI}}) \propto \text{U}_u(\cdot)$$

+ More realistic.
+ Widely applicable.
− Computationally expensive.

(link to paper)

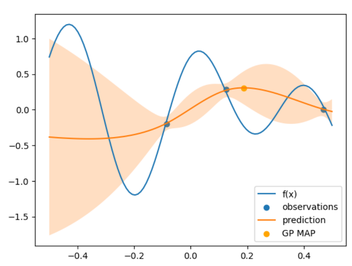


Bayesian Optimization Experiments

Setting: the AI is interested in optimizing scoring function $f : \mathcal{X} \to \mathcal{Y}$ by getting feedback $y \in \mathcal{Y}$ given queries $x \in \mathcal{X}$.

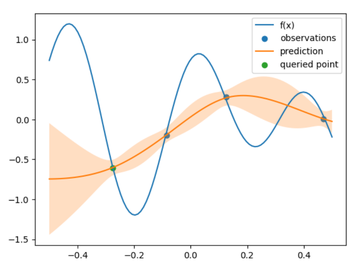
Prior

On the right we see three data points and the resulting prior.



Ground-truth posterior

True $f(x)$ on the query x leads to a undesirable posterior shown on the right.



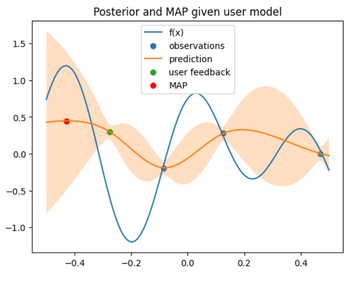
Our User Model

Instead, we argue users steer these processes and, for example, optimize the objective of moving the AI's posterior maximum:

$$\text{U}_u^{\text{argmax-dist}}(p(f), f) = -\left|\arg\max_x f(x) - \arg\max_x \mathbb{E}_{p(f)}[f(x)]\right| \tag{4}$$

User-model posterior

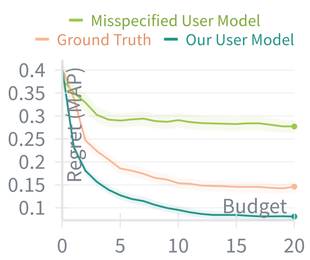
Our proposed user model overestimates the query, leading to a more favorable posterior shown on the right.



Empirical Evaluation

Compare ground truth on plain AI (*UCB* on *GP*) against our user model (i) correct and (ii) misspecified prior over the AI.

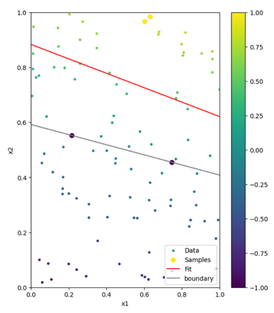
- Well-informed users steer to better solution.
- Misspecified users harm performance.



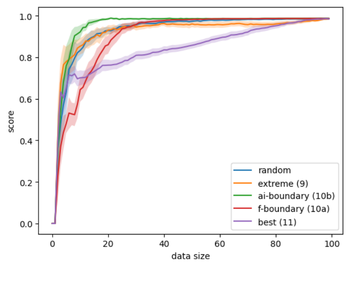
Recommender System Experiments

Setting: a recommendation system estimates a scoring function $f : \mathcal{X} \to \mathcal{Y}$ given labeled data (X, Y) provided by the user.

In real systems, users label *content they receive from the system.* We investigate the effect of several of such strategies.



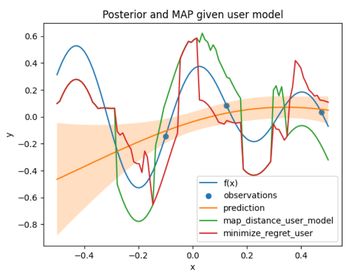
Some strategies, such as random labeling, lead to good performance. For (reasonable!) users who label the worst and best of the provided content, however, the system performs poor!



Discussion

User Model

Despite the promising results so far, behaviors are not universally intuitive and further investigations of the realism of these user models is necessary.



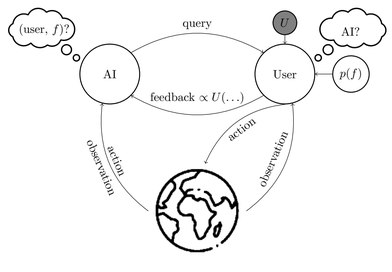
Take-Aways

- Formalized how users may act intentionally in human-in-the-loop systems.
- Developed a framework for designing and injecting such user models.
- Preliminary investigation on the models and their effect on AI performance.

Open questions

- How to approximate inference of AI?
- What are reasonable priors over "user's prior over AI".

Future Work



Sammie Katt and Samuel Kaski. Artificial theory of mind in human-in-the-loop. In *EurIPS 2025 Workshop on Metacognition in Generative AI*, 2025