More Than Irrational: Modeling Belief-Biased Agents

Yifan Zhu, Sammie Katt, Samuel Kaski

Aalto University · The University of Manchester · ELLIS Institute Finland

AAAI-26 / IAAI-26 / EAAI-26

Seemingly irrational decisions can be optimal under biased beliefs from bounded memory and we can infer those bounds online to adapt AI assistance

Problem

Seemingly Irrational Behavior in Sequential Decision-Making

Observation

Humans often take **seemingly sub-optimal** actions.

e.g. forget → re-check

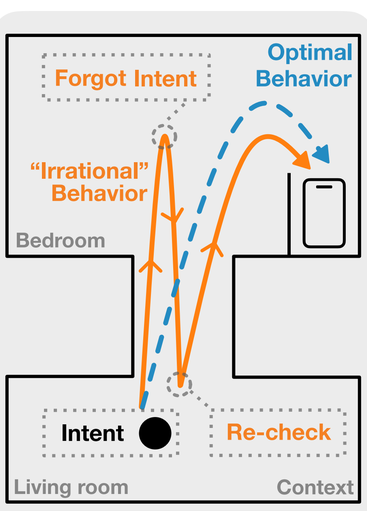
Challenge

To collaborate with such users, the AI needs to

- formalize a cognitively-interpretable model
- infer latent factors online

Insight

Sub-optimal action can be **structurally** modeled as the optimal choice under an internal biased belief induced by bounded memory.



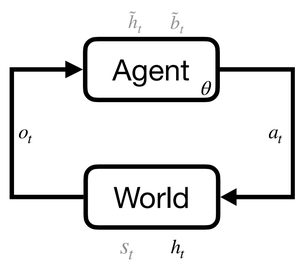
Intent: go to bedroom and get phone

Solved By

Model

Computational Rationality User Modeling

Memory Bound → Corrupted Memory → Biased Belief → Sub-optimal Action

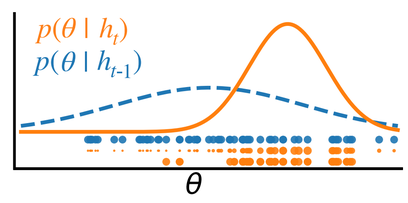


likelihood

Infer

Online Inference Of Bounds and Beliefs

New Data (observation, action) → Nested Particle Filtering → Posterior

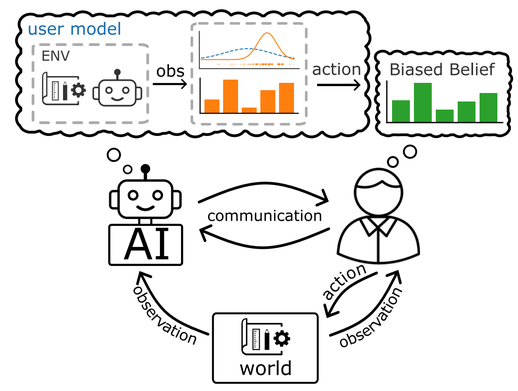


Enables

Assist

Bound-Belief-Aware Adaptive Assistance

Posterior + Cost-sensitive Intervention Policy → Assistance



Methodology

Computational Rationality User Modeling

Action Likelihood for Inference

CR-POMDP: $M_\theta = (S, A, T, R, \Omega, O, \gamma, f_\theta)$.

At time t, the CR agent:

- receives new data, updates **memory** $\tilde{h}_t = (\tilde{o}_t^{:t}, \tilde{a}_t^{:t-1})$:

$$\tilde{h}_t \sim f_\theta(\tilde{h}_{t-1}, o_t, a_{t-1})$$

(θ: user-specific memory bound; f_θ: stochastic process; o_t, a_{t-1}: new obs-action data)

- memory of past observations can **evolve**:

$$\tilde{o}_t^i \neq \tilde{o}_j^i \quad (i \le j < t)$$

($\tilde{o}_t^i$: memory of o_i maintained at time t; $\tilde{o}_j^i$: memory of o_i maintained at time j)

- updates **biased belief** from corrupted memory $\tilde{h}_t$:

$$\tilde{b}_t = p(s_t \mid \tilde{h}_t) \propto \sum_{s_{:t-1}} p(s_0) O(\tilde{o}_t^0 \mid s_0) \prod_{i=1}^{t} O(\tilde{o}_t^i \mid s_i) T(s_i \mid s_{i-1}, \tilde{a}_t^{i-1})$$

($\tilde{b}_t$: a fixed-length summary of $\tilde{h}_t$; the product: computed retroactively)

- takes action from the **optimal policy** conditioned on the biased belief: $a_t \sim \pi_*(\cdot \mid \tilde{b}_t; \theta)$. (action likelihood)

Online Inference of Bounds and Beliefs

Nested Particle Filtering

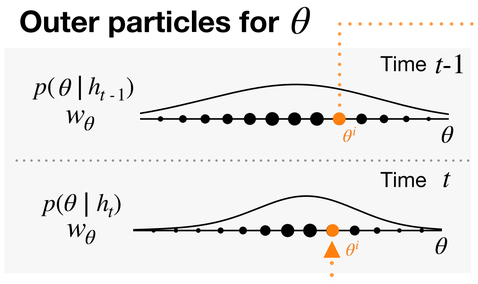
Inverse Problem: infer the user's static bound θ and latent memory $\tilde{h}_{t-1}$ online from observed history $h_t = (o_{:t}, a_{:t-1})$:

$$p(\tilde{h}_{t-1}, \theta \mid h_t) \propto \underbrace{p(a_{t-1} \mid \tilde{h}_{t-1})}_{\text{action likelihood}} \sum_{\tilde{h}_{t-2}} \underbrace{p(\tilde{h}_{t-2}, \theta \mid h_{t-1})}_{\text{recursive}} \underbrace{f_\theta(\tilde{h}_{t-1} \mid \tilde{h}_{t-2}, o_{t-1}, a_{t-2})}_{\text{stochastic memory process}}$$

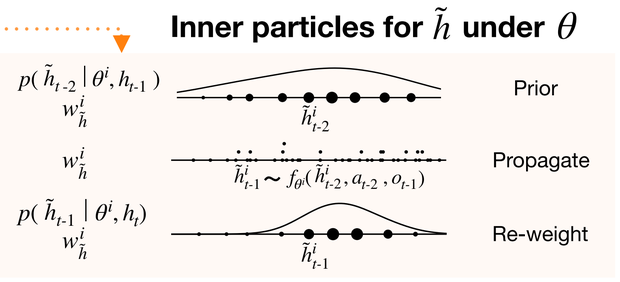
Challenge: exact computation is intractable with complexity $O(|S|^t t!)$.

Idea: approximate $p(\theta, \tilde{h}_{t-1} \mid h_t)$ with nested particle filtering; complexity comes down to $O(N_\theta N_{\tilde{h}} t |S|)$.

Outer particles for θ



Inner particles for $\tilde{h}$ under θ



Bound-Belief-Aware Adaptive Assistance

Decide when & How to assist via inferred bounds and beliefs

Goal: learn an assistive policy s.t.

$$\pi_*^{AI} = \arg\max_{\pi^{AI}} \mathbb{E}\Big[\sum_t \gamma^t \big(r_t - c(a_t^{AI})\big)\Big]$$

(task reward; intervention cost)

Assistive-POMDP:

At time t, the AI:

- observes world and agent: $o_t^{AI} = (o_t, a_{t-1}^H)$
- updates belief via NPF: $b_t^{AI} = p(\theta, \tilde{h}_{t-1} \mid h_t)$
- proposes intervention from $\{\text{DoNothing}, \text{ActionHint}, \text{MemoryHint}\}$: $a_t^{AI} \sim \pi_*^{AI}(\cdot \mid b_t^{AI})$

Behavioral Validation

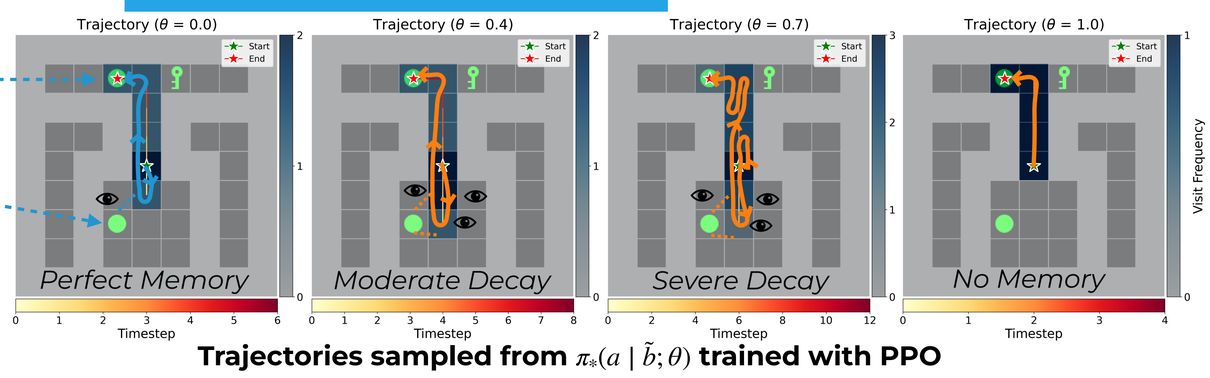
Setup: T-Maze

Goal: reach the correct terminal quickly.

The agent must:

- find clue for direction
- remember the clue
- go to the terminal

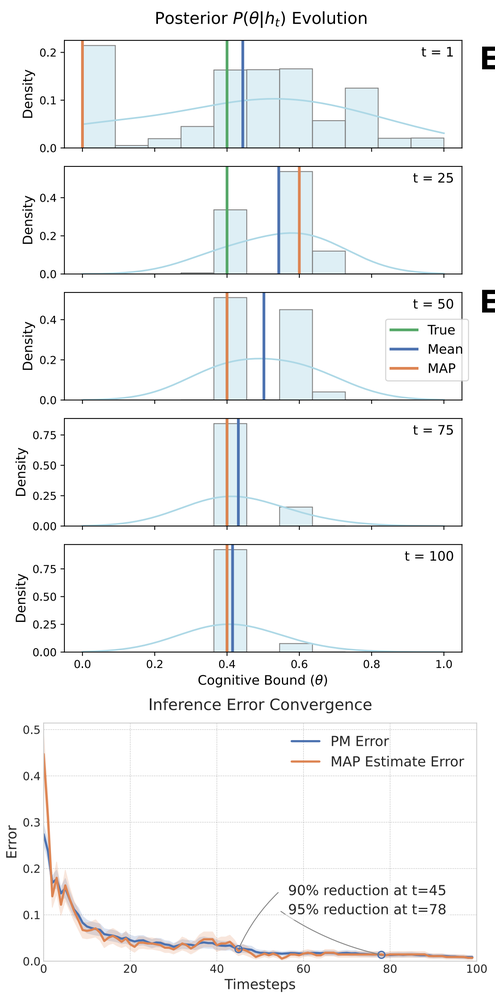
Memory decay f_θ: a stochastic forget process with θ being decay rate.



Trajectories sampled from $\pi_*(a \mid \tilde{b}; \theta)$ trained with PPO

Optimal: shortest path to success · Double checking before committing · Forget-recheck hesitation · Random guess

Inference Validation



Efficacy

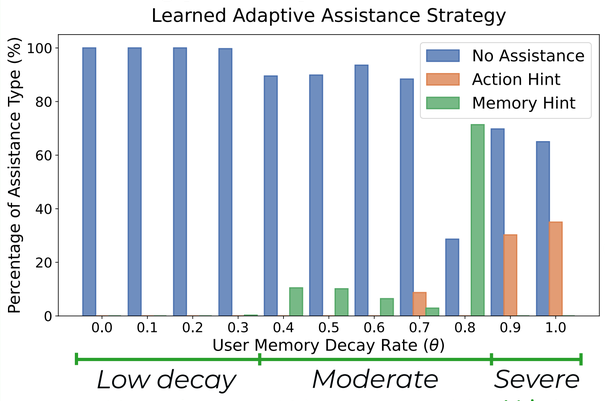
NPF quickly concentrates $p(\theta \mid h_t)$ on a small set of hypotheses.

Efficiency

Posterior-mean error drops 90% by t=45 and 95% by t=78.

Assistance Validation

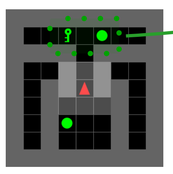
Type adaptation



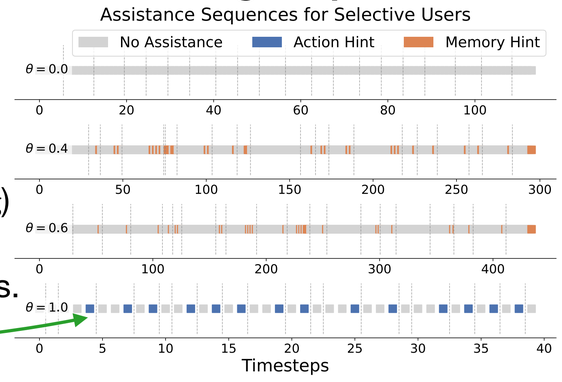
AI rarely intervene · AI intervene with memory hints · AI intervene with action hints

Setup: T-Maze

- Policy learned with PPO using inferred posterior.
- Cost: $c(\text{ActionHint}) > c(\text{MemoryHint}) > c(\text{DoNothing})$
- Critical decision time: last step before terminal states.



Timing adaptation



AI generally learned to intervene near or at critical decision time

Find in the Paper:

- Detailed formulation of assistive-POMDP.
- More experiment and result details.

Full Paper