More Than Irrational: Modeling Belief-Biased Agents

Yifan Zhu[1, 2], **Sammie Katt**[1, 2], **Samuel Kaski**[1, 2, 3]

[1]ELLIS Institute Finland
[2]Department of Computer Science, Aalto University, Finland
[3]Department of Computer Science, University of Manchester, United Kingdom
{yifan.zhu, sammie.katt, samuel.kaski}@aalto.fi

Abstract

Despite the explosive growth of AI and the technologies built upon it, predicting and inferring the sub-optimal behavior of users or human collaborators remains a critical challenge. In many cases, such behaviors are not a result of irrationality, but rather a rational decision made given inherent cognitive bounds and biased beliefs about the world. In this paper, we formally introduce a class of computational-rational (CR) user models for cognitively-bounded agents acting optimally under biased beliefs. The key novelty lies in explicitly modeling how a bounded memory process leads to a dynamically inconsistent and biased belief state and, consequently, sub-optimal sequential decision-making. We address the challenge of identifying the latent user-specific bound and inferring biased belief states from passive observations on the fly. We argue that for our formalized CR model family with an explicit and parameterized cognitive process, this challenge is tractable. To support our claim, we propose an efficient online inference method based on nested particle filtering that simultaneously tracks the user's latent belief state and estimates the unknown cognitive bound from a stream of observed actions. We validate our approach in a representative navigation task using memory decay as an example of a cognitive bound. With simulations, we show that (1) our CR model generates intuitively plausible behaviors corresponding to different levels of memory capacity, and (2) our inference method accurately and efficiently recovers the ground-truth cognitive bounds from limited observations ($\leq$ 100 steps). We further demonstrate how this approach provides a principled foundation for developing adaptive AI assistants, enabling adaptive assistance that accounts for the user's memory limitations.

Code — https://github.com/Yifan-Zhu/More-Than-Irrational-Modeling-Belief-Biased-Agents

Extended version — https://www.arxiv.org/abs/2511.12359

Introduction

In human-AI collaboration, the efficacy of an AI agent depends on its ability to infer the user's goals, beliefs, and future actions from observations of their past behavior. Learning such a user model on the fly is particularly challenging, primarily because human behaviors rarely appear to be fully rational. Computational rationality (CR) is a universal theoretical framework for understanding such human behavior, positing that humans are rational decision-makers as expected utility maximizers (Lewis, Howes, and Singh 2014; Howes et al. 2016), yet our decisions may appear sub-optimal due to latent constraints imposed by subjective utility function, cognitive bounds, and the environment (Oulasvirta, Jokinen, and Howes 2022; Howes, Jokinen, and Oulasvirta 2023; Chandramouli et al. 2024). Such constraints, especially cognitive bounds, can lead to biased beliefs and sub-optimal behavior. This happens in everyday life; for example, when going to fetch a misplaced mobile phone, we might check several seemingly irrelevant places simply because our belief about the phone's location is biased due to a faulty or decayed memory. An AI assistant capable of inferring the user's memory capacity and tracking their belief state, rather than merely assuming irrationality, would be able to provide more effective assistance.

Computational rationality has shown success in many domains, including eye movements (Chen et al. 2015, 2021), pointing (Ikkala et al. 2022), typing (Shi et al. 2024), and driving (Jokinen, Kujala, and Oulasvirta 2021). Surprisingly, little work considers bounded memory resources, and those that do (e.g. (Shi et al. 2024)) are specific to their application. We argue that a significant amount of irrational behavior can simply be explained by missing or fake memories that lead to incorrect beliefs. *Biased beliefs, even when acted upon optimally, lead to sub-optimal behavior.*

This work is motivated by a general research question in human-AI collaboration: *how can an AI assistant effectively collaborate with a user whose behaviors seem irrational due to imperfect memory?* We investigate how such users with memory-related cognitive bounds can be modeled and learned during online interactions in a partially observable environment. To this end, we propose a formal CR framework for explicitly modeling the user's biased belief as the link between latent memory bounds and the user's decisions. We posit that these cognitive bounds lead to a subjective belief state that deviates from the objective world. When the user acts optimally according to this subjective state, they make seemingly sub-optimal decisions. This formalism allows the AI to interpret seemingly irrational actions as rational decisions made upon latent bounds and biased beliefs, providing a fundamental user-specific answer to *why* an ac-

tion was taken and what actions to expect in the future.

Our CR framework is a mathematical solution for inferring the user's latent cognitive bounds and their belief state online; however, it comes with computational challenges. We contribute to tackling this issue by first defining a tractable inference setting where the AI has access to the environment model, while the user-specific cognitive parameter remains unknown. To solve this joint bound-belief estimation problem, we propose an approximate online Bayesian inference method based on sequential Monte-Carlo techniques. We validate our approach in a representative navigation task, the T-maze, using memory decay as an instance of parameterized bounded memory. The simulations validate both the expressive power of our CR model in generating intuitive behaviors across various memory capacities and the ability of our inference method to accurately recover the ground-truth bounds from passive observations. Building on this validation, we formalize an assistive-POMDP framework in the same task to demonstrate the efficacy of our work in building adaptive AI assistants. For example, it can learn to distinguish when a user needs direct action advice due to severe memory decay versus when a subtle memory hint suffices.

Related Work: Computational Rationality

Our work can be best understood from the perspective of computational rationality (CR). This literature views irrationality, such as the behavior of humans, as the result of rational techniques executed under bounded resources. Seminal papers include (Gershman, Horvitz, and Tenenbaum 2015; Oulasvirta, Jokinen, and Howes 2022; Howes, Jokinen, and Oulasvirta 2023; Chandramouli et al. 2024).

A rich body of research in CR has focused on sub-optimal behavior models with latent parameters and their inference. For example, users in (Kwon et al. 2020) maintain an internal model of the environment, though they assume perfect memory given imperfect internal models. Irrationality has also been explained by bounds on the inference budget (Jacob, Gupta, and Andreas 2023). Our work, within this growing field, stands apart as providing a model for bounded memory, which the previous works have not considered much in their general form.

Common in CR, other related work focuses on specific applications, such as gaze prediction (Chen et al. 2021), pointing (Ikkala et al. 2022), menu search (Chen et al. 2015), and traffic behavior (Jokinen, Kujala, and Oulasvirta 2021; Wang et al. 2023). To our understanding, work on memory has been limited, though some memory decay was included in user models for using touch screens (Shi et al. 2024; Jokinen et al. 2021). As opposed to application-specific solutions, we provide a more general approach that facilitates the modeling of bounded memory in multiple settings.

Of course, for more broadly interested readers, there is an inexhaustive list of works attempting to model decision making and facilitate its inverse problem; yet to our best understanding, they do not support explicit general descriptive models of memory (Jarrett, Hüyük, and Van Der Schaar 2021; Lieder and Griffiths 2020; Alanqary et al. 2021).

Preliminaries: Rational Decision Making

We model sequential decision-making under uncertainty as a Partially Observable Markov Decision Process (POMDP) (Kaelbling, Littman, and Cassandra 1998; Sutton and Barto 2018). Within this framework, we define a fully rational (FR) agent that can maintain an accurate belief state over the latent states via Bayesian filtering and acts optimally to maximize a given objective.

Formally, a POMDP is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \Omega, \gamma)$, where $\mathcal{S}, \mathcal{A}, \Omega$ denote the environment state, action, and observation spaces respectively; $\mathcal{T}(s_{t+1}|s_t, a_t), \mathcal{R}(s_t, a_t), \mathcal{O}(o_t|s_t)$ represent the state transition dynamics, reward function, and observation probability distributions respectively. At each timestep t, the FR agent is in a *latent* world state $s_t \in \mathcal{S}$ and observes $o_t \sim \mathcal{O}(o_t|s_t)$, drawn from the observation model $\mathcal{O} : \mathcal{S} \times \Omega \to [0, 1]$. The agent then updates their belief $b_t \in \Delta(\mathcal{S})$ about the environmental states via Bayesian filtering:

$$b_t(s_t) = p(s_t \mid o_t, a_{t-1}, b_{t-1}) \propto \mathcal{O}(o_t \mid s_t) \sum_{s_{t-1}} \mathcal{T}(s_t \mid s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1}). \quad (1)$$

Then, the agent makes an action from its policy $\pi : \Delta(\mathcal{S}) \to \Delta(\mathcal{A})$: $a_t \sim \pi(\cdot \mid b_t)$, and receives reward $r_t \sim \mathcal{R}(s_t, a_t)$; The world then evolves to a new state s_{t+1} stochastically according to the transition dynamics: $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t)$.

Solving a POMDP The goal of this FR agent is to learn the optimal policy π_* for maximizing the expected return: $\pi_*(a \mid b) \propto \exp(\tau Q_*(b, a))$, where τ is the inverse temperature parameter, and learning $Q_*(b, a)$, the optimal action-value function, is done through RL. For solution techniques, see (Sutton and Barto 2018; Shani, Pineau, and Kaplow 2013; Silver and Veness 2010).

In non-trivial problems, the state space is too large to compute the belief exactly. The common solution is to approximate the belief with particles. In *particle filtering* (Gordon, Salmond, and Smith 1993; Doucet et al. 2001; Liu and Chen 1998), a distribution $p(x)$ is estimated with n (potentially weighted) Monte-Carlo particles $\{x^i, w^i\}_i^n$ by $p(x) \approx \sum_i w^i \delta_{x^i}(x)$, where $\delta_x(\cdot)$ denotes the Dirac delta mass function at x.

Computational Rationality User Modeling

In this section, we formally define our computational rationality user modeling framework. Central to our proposed design is an internal memory process f_θ, parameterized by user-specific bounds $\theta \in \Theta$, that explicitly models the agent's memory: an estimate of the observation-action history $h_t \triangleq (\mathbf{o}_{:t}, \mathbf{a}_{:t-1})$. This function captures, for example, how memories of the past decay over time. With this addition, our user model is a POMDP solver using biased beliefs as a result of imperfect memory according to f_θ.

Formally, at time t, we define the internal memory over observations and actions *received* up to time i ($t \geq i$) as:

$$\tilde{h}_t^{:i} = (\tilde{\mathbf{o}}_t^{:i}, \tilde{\mathbf{a}}_t^{:i-1}) = (\tilde{o}_t^1, ..., \tilde{o}_t^i, \tilde{a}_t^1, ..., \tilde{a}_t^{i-1}) \quad (2)$$

where each element $\tilde{o}_t^j \in \Omega$ and $\tilde{a}_t^{j-1} \in \mathcal{A}$ is the agent's internal (corrupted) version at time t, of the true observation o_j and action a_{j-1} originally received at time j. The key insight of our CR model is that the memory of past observations can evolve, i.e. $\tilde{o}_t^j \neq \tilde{o}_{t-1}^j$. For notational simplicity, we denote the complete internal memory $\tilde{h}_t^{:t}$ at time t as $\tilde{h}_t$.

The agent's internal state $\tilde{h}_t$ evolves according to the internal stochastic dynamics function $f_\theta : \Omega^{t-1} \times \mathcal{A}^{t-2} \times \Omega \times \mathcal{A} \to \Delta(\Omega^t \times \mathcal{A}^{t-1})$ that maps the previous internal history $\tilde{h}_{t-1}$ and new observation pair (o_t, a_{t-1}) to a distribution over the current cognitive state:

$$\tilde{h}_t \sim f_\theta(\tilde{h}_{t-1}, o_t, a_{t-1}), \tag{3}$$

which is parameterized by the cognitive bound θ.

For intuitions, recall the phone-searching example. A limited memory capacity can result in forgetfulness, resulting in corrupted memory and, consequently, biased beliefs. For example, humans frequently forget about their last interaction(s) with their phone and, hence, have incorrect beliefs about its location. The function f_θ provides the formal mechanism for this corruption process.

Unlike rational agents, a CR agent computes its belief state $\tilde{b}_t \in \Delta(\mathcal{S})$ via Bayes' rule by conditioning on the corrupted internal memory of the history:

$$\begin{aligned}\tilde{b}_t &\triangleq p(s_t \mid \tilde{\mathbf{o}}_t^{:t}, \tilde{\mathbf{a}}_t^{:t-1}) \propto \sum_{\mathbf{s}_{:t-1}} p(\mathbf{s}_{:t}, \tilde{\mathbf{o}}_t^{:t} \mid \tilde{\mathbf{a}}_t^{:t-1}) \\ &= \sum_{\mathbf{s}_{:t-1}} p(s_0)\mathcal{O}(\tilde{o}_t^0 \mid s_0) \\ &\qquad \times \prod_{i=1}^{t} \mathcal{O}(\tilde{o}_t^i \mid s_i)\mathcal{T}(s_i \mid s_{i-1}, \tilde{a}_t^{i-1}).\end{aligned} \tag{4}$$

Because the Markov property no longer holds for the biased belief, computing it involves marginalizing out the agent's entire, potentially flawed, memory sequence, rather than simply a one-step update based on the previous belief in Equation (1). Namely, when memory modifies $\tilde{o}_t^i$, the belief state shifts retroactively. Equation (4) along with Equation (3) shows that the cognitive bound θ introduces bias into the belief by systematically corrupting the elements in $\tilde{h}_t$ through f_θ. This non-trivial formalism allows for explicitly capturing human-like decision-making that often involves "replaying" or "re-evaluating" memories.

The CR agent is internally rational, acting optimally based on their subjective beliefs:

$$\pi_*(a \mid \tilde{b}; \theta) = \frac{\exp(\tau\, Q_*(\tilde{b}, a; \theta))}{\sum_{a' \in \mathcal{A}} \exp(\tau\, Q_*(\tilde{b}, a'; \theta))}, \tag{5}$$

where τ is the inverse temperature parameter, $Q_*(\cdot; \theta)$ is the optimal action-value function learned based on the agent's biased beliefs. The CR agent's action $a_{\text{CR}} \sim \pi_*(a \mid \tilde{b}; \theta)$ is rational from its own perspective, yet for external observers with access to the objective belief b^*, the exact action may appear sub-optimal (i.e. $Q_*(b^*, a_{\text{CR}}) < Q_*(b^*, a_{\text{FR}})$). This framework thus provides a principled and elegant way to model sub-optimal behavior stemming from biased beliefs as a result of latent cognitive bounds.

Online Inference of Bounds and Beliefs

A direct result of the CR user model above is the ability to simulate agents with bounded and evolving memory. Most useful applications necessitate solving the inverse problem: inferring the latent bound from observed behavior. In this section, we present and address the technical challenge of learning such user models by inferring θ online. We first present a well-defined problem setting, then introduce an online inference approach based on nested particle filtering.

Problem Setting

We are interested in estimating an agents' bound parameter θ online given a passively observed action-observation history ($h_t \triangleq (\mathbf{o}_{:t}, \mathbf{a}_{:t-1})$). Given that we have a likelihood model — the user model described above (Equation (5)) — it is natural to assume a uniform prior over the parameter of interest $p(\theta)$ and consider the Bayesian task of inferring its posterior. It is natural to consider the joint distribution over parameter and internal state (here $\tilde{h}_{t-1}$), as the following derivation shows:

$$\begin{aligned}&p(\tilde{h}_{t-1}, \theta \mid h_t) \\ &= p(\tilde{h}_{t-1}, \theta \mid h_{t-1}, a_{t-1}, o_t) \\ &\propto \sum_{\tilde{h}_{t-2}} p(\tilde{h}_{t-1}, \tilde{h}_{t-2}, a_{t-1}, \theta \mid h_{t-1}, o_t) \\ &= p(a_{t-1} \mid \tilde{h}_{t-1}) \\ &\qquad \times \sum_{\tilde{h}_{t-2}} p(\tilde{h}_{t-1} \mid \tilde{h}_{t-2}, h_{t-1}, \theta) p(\tilde{h}_{t-2}, \theta \mid h_{t-1}) \\ &= \overbrace{p(a_{t-1} \mid \tilde{h}_{t-1})}^{\text{user action likelihood}} \sum_{\tilde{h}_{t-2}} \Big[\overbrace{p(\tilde{h}_{t-2}, \theta \mid h_{t-1})}^{\text{recursive}} \\ &\qquad \times f_\theta(\tilde{h}_{t-1} \mid \tilde{h}_{t-2}, o_{t-1}, a_{t-2}) \Big].\end{aligned} \tag{6}$$

Let us consider these three terms. First, the memory model f_θ is assumed to be given (for example memory decay, as used in our experiments). The "recursive" term is the belief quantity computed at the previous timestep and thus given. The first term, the "user action likelihood", reflects the core idea of this work and represents the probability that a CR agent takes action a given a (biased) belief resulting from corrupted memory $\tilde{h}_{t-1}$. In practice, this is computed by deriving the biased belief $\tilde{b}$ given the corrupted memory and computing the optimal action given this belief (recall Equations (4) & (5)), i.e. $p(a_{t-1} \mid \tilde{h}_{t-1}) = \pi_*(a_{t-1} \mid \tilde{b}_{t-1}; \theta)$ with $\tilde{b}_{t-1} = p(s \mid \tilde{h}_{t-1})$. To compute this, we assume knowledge of the system dynamics ($\mathcal{T}$ & $\mathcal{O}$) and the user's bounded memory model family f_θ (though not their latent parameter θ!). In practice, we precompute these policies.

This derivation formally exposes the non-trivial computational structure of the inference problem. The computational intractability of the belief update is clear: the summation, $\sum_{\tilde{h}_{t-2}}$, is over the high-dimensional space of (corrupted) memory sequences and grows in time. This, we address next.

Online Inference via Nested Particle Filtering

In light of the computation challenges and the problem structure of estimating a static parameter θ alongside a dynamic latent state $\tilde{h}_{t-1}$, we propose to approximate the posterior over $(\tilde{h}_{t-1}, \theta)$ with *nested* particle filtering (NPF). We find this technique particularly effective when a joint distribution is best represented as a marginal and conditional distribution: $p(\tilde{h}_{t-1}, \theta) = p(\theta)p(\tilde{h}_{t-1}|\theta)$. Additionally, since in our setting the parameter space Θ is small and the policy likelihood $\pi_*(\cdot; \theta)$ is expensive to learn and thus realistically must be precomputed, NPF suits naturally here over other inference methods, such as Particle MCMC (Andrieu, Doucet, and Holenstein 2010), where the latter samples novel θ and requires computing policies online.

In practice, NPF maintains N_θ (potentially weighted) particles $\{\theta^i\}_{i=1}^{N_\theta}$ and, *for each particle* θ^i, $N_{\tilde{h}}$ conditional particles $\{\tilde{h}_{t-1}^{(i,j)}\}_{j=1}^{N_{\tilde{h}}}$. The posterior, following standard Monte-Carlo formalisms, is then approximated as follows:

$$p(\theta \mid h_t) \approx \sum_{i=1}^{N_\theta} w^i \delta_{\theta^i}(\theta) \tag{7}$$

$$p(\tilde{h}_{t-1} \mid h_t, \theta^i) \approx \sum_{j=1}^{N_{\tilde{h}}} w^{(i,j)} \delta_{\tilde{h}_{t-1}^{(i,j)}}(\tilde{h}_{t-1}) \tag{8}$$

We utilize particle filtering to update the posterior given a new action-observation pair. The key step in this process is the weight update: the weight of each outer particle, w^i, is updated according to the likelihood of explaining the user's most recent action a_{t-1}, estimated by the inner particle filter associated with θ^i. Our contribution, regarding inference methodology, is the specific formulation of this likelihood computation within our CR framework. As detailed in Algorithm 1, this likelihood computation requires simulating the user's internal decision-making process for each particle (i, j): first, each internal state particle $\tilde{h}^{(i,j)}$ is updated using our cognitive process f_{θ^i}, and the corresponding biased belief $\tilde{b}^{(i,j)}$ is computed (line 4 and 5). Then the weights of all the particles are updated according to their likelihood of producing the observed action $\pi_*(a \mid \tilde{b}^{(i,j)}; \theta^i)$ (line 6). Lastly, the weights are re-normalized (line 10 to 12).

With NPF estimation, it is possible to infer a user's bounds and track their biased belief online: the computational cost of exact inference in Equation (6) is $O(|\mathcal{S}|^t t!)$, while our inference method only costs $O(N_\theta N_{\tilde{h}} t|\mathcal{S}|)$ (biased belief computation in Equation (4) costs $O(t|\mathcal{S}|)$), which allows for real-time inference.

Algorithm 1: Nested Particle Filter Update $p(\tilde{h}_{t-1}, \theta \mid h_t)$

Input: Previous Particles: $\left\{\theta^i, w^i, \{\tilde{h}_{t-2}^{(i,j)}, w^{(i,j)}\}_{j=1}^{N_{\tilde{h}}}\right\}_{i=1}^{N_\theta}$
History h_t: $(\mathbf{o}_{:t}, \mathbf{a}_{:t-1})$

1: // update particles:
2: **for** $i = 1, \ldots, N_\theta$ **do**
3: **for** $j = 1, \ldots, N_{\tilde{h}}$ **do**
4: Sample $\tilde{h}_{t-1}^{(i,j)} \sim f_{\theta^i}(\tilde{h}_{t-2}^{(i,j)}, a_{t-2}, o_{t-1})$
5: Compute $\tilde{b}_{t-1}^{(i,j)} \leftarrow p(s \mid \tilde{h}_{t-1}^{(i,j)})$
6: Evalute likelihood $L^{(i,j)} \leftarrow \pi_*(a_{t-1} \mid \tilde{b}_{t-1}^{(i,j)}; \theta^i)$
7: **end for**
8: **end for**
9: // update and normalize weights:
10: $w^{(i,j)} \leftarrow w^{(i,j)} L^{(i,j)}$
11: $w^i \leftarrow \frac{w^i \sum_{j=1}^{N_{\tilde{h}}} w^{(i,j)}}{\sum_{i'} w^{i'} \sum_{j=1}^{N_{\tilde{h}}} w^{(i',j)}}$
12: $w^{(i,j)} \leftarrow \frac{w^{(i,j)}}{\sum_{j'} w^{(i,j')}}$
13: **return** updated particles $\left\{\theta^i, w^i, \{\tilde{h}_{t-1}^{(i,j)}, w^{(i,j)}\}_{j}^{N_{\tilde{h}}}\right\}_{i=1}^{N_\theta}$

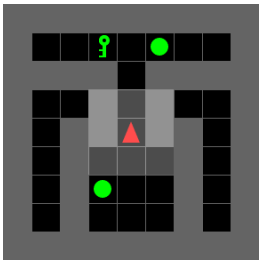


Figure 1: Rendering of the T-maze simulation task. The agent (red triangle) must explore the bottom room to find the target object (ball in this case), memorize it, and navigate to the corresponding terminal state (one of the arms of the "T", here the right-side) to succeed.

Experiments

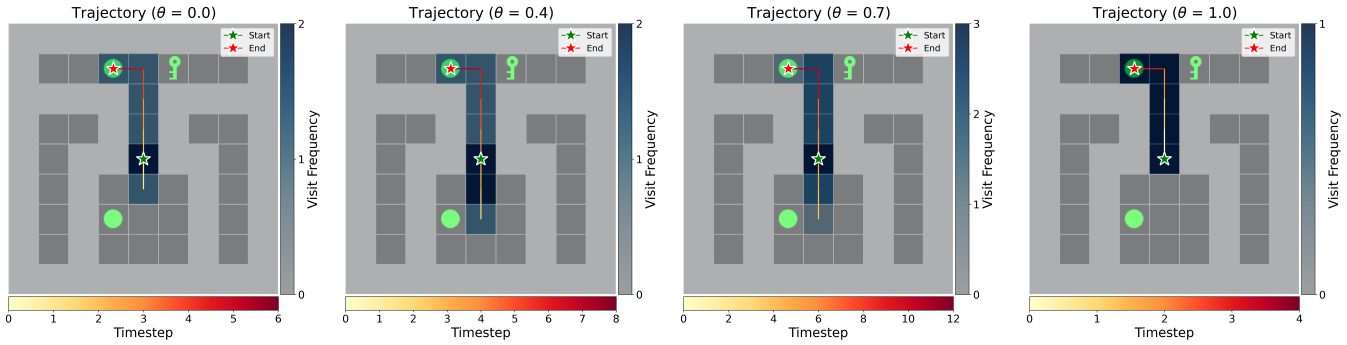
In this section, we present a series of simulation-based experiments designed to provide a foundational proof-of-concept validation for our core claims. Specifically, we aim to answer the following key questions through simulations:

1. Can our proposed CR user model generate a spectrum of intuitively plausible, sub-optimal behaviors that correspond to different levels of cognitive bounds?
2. Can our online inference method accurately and efficiently infer the user's latent cognitive bounds and state from only passive observations?
3. Does the inferred cognitive bound serve as useful information for a downstream adaptive assistance task?

To this end, we take a typical navigation task that requires memory use as a simple clear testbed for our validation.

Simulation Task: T-maze

All our experiments are performed on a simulated grid-world partial observable navigation task based on the T-maze, a classical test for working memory. As illustrated in Figure 1, the agent (red triangle) starts in a hallway and is tasked with navigating to the goal, one of the two terminating locations at the top. The optimal behavior first explores the bottom room, observes and memorizes the object (here a ball), and then proceeds through the hallway to reach the location with the same object as seen. The hidden state contains both the agent's current position and the target object; The agent can choose to go to any of the four direc-



(a) Perfect Memory ($\theta = 0.0$) (b) Imperfect Memory ($\theta = 0.4$) (c) Imperfect Memory ($\theta = 0.7$) (d) No Memory ($\theta = 1.0$)

Figure 2: Representative trajectories generated by our CR model for agents with different memory bound parameters θ. The color of the trajectory indicates the temporal order of the steps (brighter is earlier), and the colored grids indicate the visit frequency. The agent starts in the hallway (green star) and reaches one of the terminal states (red star). (a) With perfect memory, the agent acts optimally by going down to find the object and directly proceeds to the correct terminal state. (b) With a 40% chance of losing memory, the agent learned to collect more observations on the object. (c) With a 70% chance of losing memory, the agent learned to recheck the found position after it appeared to have forgotten. (d) With no memory, the agent learned to take a random guess without exploring the environment.

tions, or simply stay in place; The observation is the 3 by 3 grid around the agent; The transition function is deterministic; reaching the correct terminal state generates a reward of $1 - 0.9 \times \frac{\text{timesteps}}{\text{maxsteps}}$, in all other scenarios the reward is 0. This task is representative in our case as it creates a clear credit assignment problem that can only be solved by maintaining a specific piece of information in memory over a period of time, and we expect varying behaviors depending on the quality of the memory model.

Validation of the CR User Model

Setup We start by studying the expressive power of our CR user model. We instantiate a CR agent with a cognitive process f_θ designed to capture the gradual decay of memory. We implement this as having a $p = \theta$ chance of corrupting the observation of the target object. Specifically, at each step t, after the new observation (o_t, a_{t-1}) is added to the buffer, for each observation of the target object in $\tilde{h}_t$, there is a probability of θ that the seen object is replaced by a default value. The parameter θ here thus represents the memory decay rate, where an agent with $\theta = 0.0$ has perfect memory and the one with $\theta = 1.0$ forgets the object seen immediately. To evaluate the behavior trajectories generated by our CR user model in this case, we pre-trained optimal policies $\pi_*(a \mid \tilde{b}; \theta)$ for a range of θ values using Proximal Policy Optimization (PPO) (Schulman et al. 2017) and visualized the corresponding trajectories. The goal is to verify that our model can generate a range of distinct and cognitively plausible behaviors by varying the memory decay rate θ.

Results We present 4 representative trajectories in Figure 2, illustrating the behavioral pattern of CR agents having perfect memory, imperfect memory, and no memory:

- Perfect Memory ($\theta = 0.0$): with no decay, the agent learned to act optimally by taking the shortest path to complete the task successfully: one step down to observe the object, then proceed to the correct terminal state.
- Imperfect Memory ($\theta \in \{0.4, 0.7\}$): agents with various levels of memory decay exhibit classic sub-optimal but highly intuitive behavioral patterns. With a moderate decay rate $\theta = 0.4$, the agent learned to collect enough observations for robust memorization. With worse memory capability ($\theta = 0.7$), the agent shows a "forget-recheck" pattern: they forgot the seen object before making the turning decision, and went down to re-check the object.
- No Memory ($\theta = 1.0$): with no memory, the agent is unable to solve the task strategically. It does not even waste time exploring; instead, it makes a random turn.

This result shows that our CR model is qualitatively capable of generating a series of diverse and intuitively plausible behavioral patterns given different memory-bound values. The simple yet interpretable mechanism of memory decay is sufficient to generate a spectrum of behaviors that are not merely random and irrational, but are actually sub-optimal in a structured and plausible way.

Validation of Online Bound-Belief Inference

Setup To evaluate our online inference method, we run Algorithm 1 on action-observation trajectories of a CR agent in 100 timesteps across multiple episodes. For each simulation, we sample a ground-true θ_{true} uniformly from $\{0.0, 0.1, ..., 1.0\}$ for the CR agent. Then, we assume same uniform prior over θ_{true} and use our method to infer the joint posterior $p(\tilde{h}_{t-1}, \theta \mid h_t)$ with N_θ and $N_{\tilde{h}}$ particles.

We evaluate the performance of our inference method with two metrics: the Posterior Mean (PM) error ($|\mathbb{E}[\theta \mid h_t] - \theta_{\text{true}}|$), which is the absolute difference between the mean of the estimated posterior $p(\theta \mid h_t)$ and θ_{true}, and the Maximum a Posteriori (MAP) error ($|\arg\max_\theta p(\theta \mid h_t) - \theta_{\text{true}}|$), which is the absolute difference between the

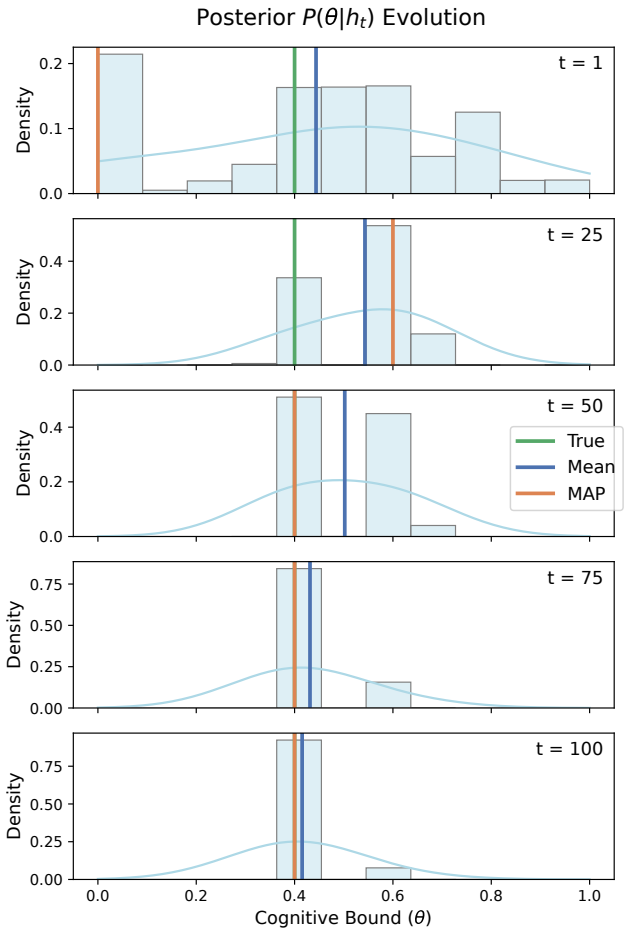


Figure 3: Visualization of $p(\theta|h_t)$ evolving over time where $\theta_{\text{true}} = 0.4$ and $\tau = 3.0$. Each panel shows the estimated posterior at representative timesteps, where the histogram represents the distribution of the weighted particles and the curve is the kernel density estimate used to approximate the posterior. The posterior mean (blue line) and MAP estimate (red line) rapidly converge towards the true value (green line) as more observations are made.

most probable estimated value and θ_{true}. All results are reported as means and standard errors over runs with all θ_{true} and 5 different random seeds, and sensitivity analysis of hyperparameter $N_{\tilde{h}}$ and τ is available in the extended version, which can be found in our code repository.

Results We present a qualitative plot of the inference updates and a quantitative convergence plot to demonstrate that our method is both accurate and efficient for identifying the user's latent parameter θ from passive observations.

Figure 3 visualizes the evolution of the posterior distribution $p(\theta \mid h_t)$ at representative timesteps (steps 1, 25, 50, 75, and 100) where $\theta_{\text{true}} = 0.4$. After one-step inference from a uniform prior, the estimated posterior shifts dramatically, indicating that the user's initial action is highly informative, which may depend on the task design. With more observations, the estimation concentrates around plausible hypotheses, i.e. $\{0.4, 0.6\}$, showing the challenge of identifiability: different cognitive bounds can lead to similar actions. As evidence accumulates, the estimation converges gradually, and our method identifies the true bound from competing hypotheses. The figure shows how our method reduces uncertainty in identifying θ over time with more observations of the environment and the user's actions.

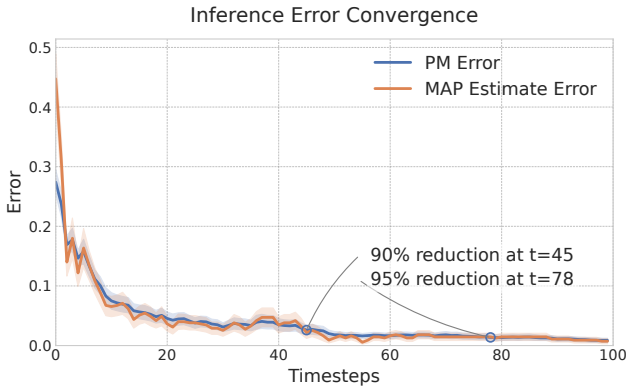


Figure 4: Inference error convergence: the PM error and MAP error over time (mean $\pm$ standard error over 5 seeds and all θ_{true}). Relative to the initial error at $t = 1$, the PM error decreased by 90% and 95% at $t = 45$ and $t = 78$ respectively, demonstrating the inference efficiency.

To quantify the accuracy and efficiency of the process, we aggregated the results across all ground-truth θ conditions and 5 random seeds where $\tau = 3.0$. Figure 4 plots the evolution of the PM error and MAP error over 100 steps, confirming two key properties of our method:

1. Accuracy: Both error metrics rapidly converge to near-zero, with the final PM error being 0.0087 ± 0.0035 (standard error), demonstrating method effectiveness.
2. Data Efficiency: The majority of the error reduction occurs within the first 20-30 steps ($\sim$ 2-3 episodes) of observations, and the PM error is reduced by 90% after 45 steps and by 95% after 78 steps (baseline being PM error at $t = 1$), demonstrating the applicability potential of our approach in real-time assistance.

Together, the qualitative visualization of the posterior evolution in Figure 3 and the quantitative error analysis in Figure 4 provide strong empirical evidence in support of our central claim. They demonstrate that the distinct behavioral patterns generated by our CR model contain sufficient information for inference, and that our NPF-based method can effectively utilize this information, which supports our claim that the user's latent cognitive bound is practically identifiable within our proposed framework.

Application Demonstration: Assistive-POMDP

Setup To demonstrate how our CR user model and inference method can benefit developing adaptive AI assistants, we designed an AI assistant aimed at maximizing the collaborative return while minimizing intervention cost. Here, the

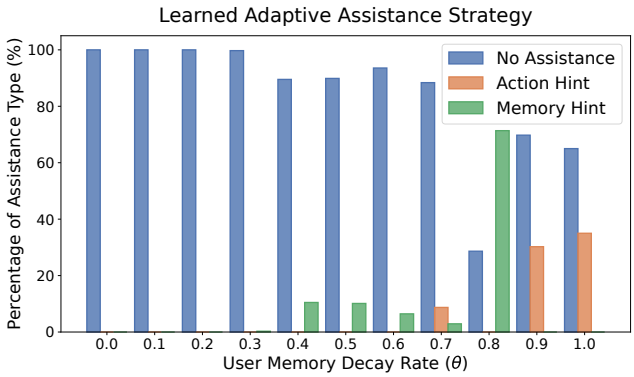


Figure 5: The learned adaptive assistance policy shows the distribution of assistance types provided by the AI assistant to simulated users with different ground-truth memory decay rates (θ). The AI learned not to intervene for users with good memory ($\theta \leq 0.3$), provide timely memory hints for moderately forgetful users, and provide more direct action hints for users with severe memory decay ($\theta \geq 0.9$).

assistant observes the CR agent's actions and the task environment while the CR agent solves the navigation task, and is tasked with helping it. The AI is rewarded when the CR agent finds the goal. The AI can do nothing or, for a small cost, remind the CR agent of a previous critical observation or, for a higher cost, suggest an action directly. The AI is assumed to have access to the state of the environment, and thus knows where the goal is, but does not know the internal state $\tilde{h}$ or cognitive bounds θ of the user. We formalize the AI's problem as a POMDP, where the internal state and parameters of the CR agent are hidden. We solve this POMDP by using our proposed inference method to maintain a belief over those hidden quantities, and optimize the belief-based policy with PPO. The detailed formalization of this assistive-POMDP framework and experiment setup, as well as the assistance policy learning algorithm, can be found in the extended version from our code repository.

Results We evaluated our solution on simulated users with various memory bounds. The results demonstrate that the AI learned a highly adaptive policy that tailors both the type and timing of interventions to the inferred user cognitive bounds. Figure 5 summarizes the learned policy by showing the distribution of each assistance type across the full spectrum of users, showing an intuitively adaptive strategy:

- For users with low memory decay ($\theta \leq 0.3$), the AI correctly infers that the user is memory competent and learned to rarely intervene.
- For users with moderate memory decay ($\theta \in [0.4, 0.8]$), the AI identifies their need for cognitive support and provides significantly more memory hints and action hints.
- For users with severe memory decay ($\theta \geq 0.9$), the AI learns that simple memory hints are insufficient as the user is likely to forget them immediately, thus providing more direct action hints at critical moments.

To investigate the assistance timing, we plot an intervention

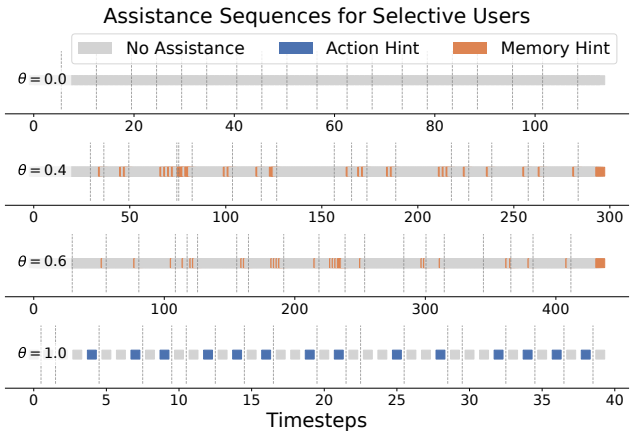


Figure 6: The sequences show the timing of the learned assistance. Each panel shows the assistance sequence over multiple episodes for a user with memory bound θ; The vertical dashed lines indicate episode boundaries. For moderately forgetful users ($\theta \in \{0.4, 0.6\}$), memory hints are occasionally provided around the end of episodes, right before making the critical decision; For users with no memory $\theta = 1.0$, only action hint is provided on the last step.

sequence of a representative run in Figure 6, which shows that the AI provides memory hints for moderately forgetful users ($\theta \in \{0.4, 0.6\}$), and action hints for severely forgetful users ($\theta = 1.0$) at the end of each episode and at moments that are critical or easy-to-forget. This demonstrates that the AI learned to intervene at moments of maximum utility, and that our online inference method provides a *principled* foundation for learning an adaptive assistance policy.

Discussion and Conclusion

We have proposed a model, based on the concept of computational rationality, to explain irrational behavior as a result of imperfect memory. Crucially, this model is flexible in its choice of memory model f_θ, and allows for flexibility in the form of latent variables. As part of our contribution, we also described an efficient inference algorithm based on nested particle filtering, demonstrating the effectiveness and interpretability of the model and inference technique in a non-trivial domain that reflects critical decision-making, with both qualitative and quantitative analysis.

There are limitations that, when addressed in future work, would further strengthen this research direction. We want to highlight two of those: First, for simplicity, we have assumed that both the user model and inference method have access to the underlying dynamics of the problem. Second, while our choice of memory model in the experiments is reasonable, it is likely insufficient to capture realistically sophisticated agents. Nevertheless, we believe this work is an important step toward a *general* model for prediction and inference over the irrational behavior of agents due to imperfect memory.

Acknowledgments

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

Alanqary, A.; Lin, G. Z.; Le, J.; Zhi-Xuan, T.; Mansinghka, V. K.; and Tenenbaum, J. B. 2021. Modeling the mistakes of boundedly rational agents within a Bayesian theory of mind. *arXiv preprint arXiv:2106.13249*.

Andrieu, C.; Doucet, A.; and Holenstein, R. 2010. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3): 269–342.

Chandramouli, S.; Shi, D.; Putkonen, A.; De Peuter, S.; Zhang, S.; Jokinen, J.; Howes, A.; and Oulasvirta, A. 2024. A workflow for building computationally rational models of human behavior. *Computational Brain & Behavior*, 7(3): 399–419.

Chen, X.; Acharya, A.; Oulasvirta, A.; and Howes, A. 2021. An adaptive model of gaze-based selection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Chen, X.; Bailly, G.; Brumby, D. P.; Oulasvirta, A.; and Howes, A. 2015. The emergence of interactive behavior: A model of rational menu search. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 4217–4226.

Doucet, A.; De Freitas, N.; Gordon, N. J.; et al. 2001. *Sequential Monte Carlo methods in practice*, volume 1. Springer.

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

Gordon, N. J.; Salmond, D. J.; and Smith, A. F. 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, 107–113. IET.

Howes, A.; Jokinen, J. P. P.; and Oulasvirta, A. 2023. Towards machines that understand people. *AI Magazine*, 44(3): 312–327.

Howes, A.; Warren, P. A.; Farmer, G.; El-Deredy, W.; and Lewis, R. L. 2016. Why contextual preference reversals maximize expected value. *Psychological review*, 123(4): 368.

Ikkala, A.; Fischer, F.; Klar, M.; Bachinski, M.; Fleig, A.; Howes, A.; Hämäläinen, P.; Müller, J.; Murray-Smith, R.; and Oulasvirta, A. 2022. Breathing life into biomechanical user models. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 1–14.

Jacob, A. P.; Gupta, A.; and Andreas, J. 2023. Modeling boundedly rational agents with latent inference budgets. *arXiv preprint arXiv:2312.04030*.

Jarrett, D.; Hüyük, A.; and Van Der Schaar, M. 2021. Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning*, 4755–4771. PMLR.

Jokinen, J.; Acharya, A.; Uzair, M.; Jiang, X.; and Oulasvirta, A. 2021. Touchscreen typing as optimal supervisory control. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.

Jokinen, J. P.; Kujala, T.; and Oulasvirta, A. 2021. Multitasking in driving as optimal adaptation under uncertainty. *Human factors*, 63(8): 1324–1341.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2): 99–134.

Kwon, M.; Daptardar, S.; Schrater, P. R.; and Pitkow, X. 2020. Inverse rational control with partially observable continuous nonlinear dynamics. *Advances in neural information processing systems*, 33: 7898–7909.

Lewis, R. L.; Howes, A.; and Singh, S. 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2): 279–311.

Lieder, F.; and Griffiths, T. L. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43: e1.

Liu, J. S.; and Chen, R. 1998. Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443): 1032–1044.

Oulasvirta, A.; Jokinen, J. P. P.; and Howes, A. 2022. Computational Rationality as a Theory of Interaction. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–14. ACM.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shani, G.; Pineau, J.; and Kaplow, R. 2013. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1): 1–51.

Shi, D.; Zhu, Y.; Jokinen, J. P.; Acharya, A.; Putkonen, A.; Zhai, S.; and Oulasvirta, A. 2024. CRTypist: Simulating touchscreen typing behavior via computational rationality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Silver, D.; and Veness, J. 2010. Monte-Carlo planning in large POMDPs. *Advances in neural information processing systems*, 23.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book. ISBN 0262039249.

Wang, Y.; Srinivasan, A. R.; Jokinen, J. P.; Oulasvirta, A.; and Markkula, G. 2023. Modeling human road crossing decisions as reward maximization with visual perception limitations. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, 1–6. IEEE.