When and Why Hyperbolic Discounting Matters for Reinforcement Learning Interventions

Ian M. Moore, Eura Nofshin, Siddharth Swaroop, Susan Murphy, Finale Doshi-Velez, Weiwei Pan

Keywords: Hyperbolic discounting, Human-AI interaction, Agent-based modeling of humans

Summary

In settings where an AI agent sends interventions to nudge a human agent toward a goal, the AI's ability to quickly learn a high-quality policy depends on how well it models the human. Despite behavioral evidence that humans hyperbolically discount future rewards, we continue to model human agents as Markov Decision Processes (MDPs) with exponential discounting because of its mathematical properties. In this work, we derive an exponential discount factor that will never miss a necessary intervention–and minimizes unnecessary extra interventions–even when the real human is hyperbolic. In addition, we demonstrate that when the dynamics are unknown, using our exponential alternative outperforms correctly modeling the human, even when the human's true hyperbolic discount is known.

Contribution(s)

1. Using theory, we connect model misspecification of a hyperbolic human agent as an exponential one to errors in the downstream AI intervention policy.
   Context: Prior work in human-AI settings has not studied how misspecifications of the human agent's discount affect AI policies. Our analysis is in the context of absorbing state MDPS (discrete state / action spaces with absorbing reward states) and on interventions of the human agent's discount factor. We make simplifying assumptions– about the stochasticity of the transitions, intermediate rewards, and noise in the human policy– which *we relax* in our empirical experiments. All humans in our experiments are simulated agents modeled using a Markov Decision Process (MDP).
2. We prove that the exponential mean hazard rate, γ_{mhr}, guarantees no false negatives in the AI policy. However, it does not minimize AI false positives.
   Context: The AI policy is the optimal policy for an MDP in which the actions are interventions, delivered by an artificial agent, on a human agent's MDP parameters. The mean hazard rate (MHR) is an established method for approximating hyperbolic human agents as exponential ones (Rambaud & Torrecillas, 2005; Sozou, 1998; 2009). Previously, there were no formal guarantees on how the MHR affects error when used to model human agents in a human-AI setting. The same context from contribution 1 (about absorbing-state MDPs, theoretical assumptions), apply.
3. We derive a fixed exponential discount rate, γ_{safe}, for approximating hyperbolic agents.
   Context: Our theoretical justification relies on the same assumptions as contribution 1. However, γ_{safe} is as broad as γ_{mhr} and is applicable to settings beyond the ones considered in this paper.
4. In empirical experiments (on small tabular MDPs), we demonstrate that (biased) exponential approximations using a fixed discount parameter outperform several different (unbiased) methods of approximating the hyperbolic discount when the transitions are learned online.
   Context: Prior work had not considered how the choice of discount model for the human agent affects the AI policy. We found that the hyperbolic approximations are unexpectedly sensitive to online learning. Our experiments are in small, tabular MDP settings.
5. Empirically, we characterize situations where a fixed exponential discount model with γ_{safe} is preferable to a fixed one with γ_{mhr}; we do the same for γ_{safe} vs. updating γ online.
   Context: None.

When and Why Hyperbolic Discounting Matters for Reinforcement Learning Interventions

Ian M. Moore[1, †], **Eura Nofshin**[1, †], **Siddharth Swaroop**[1], **Susan Murphy**[1], **Finale Doshi-Velez**[1], **Weiwei Pan**[1]
ian_moore@college.harvard.edu, eurashin@g.harvard.edu

[1]**Department of Computer Science, Harvard University, USA**

[†] indicates equal contribution

Abstract

In settings where an AI agent nudges a human agent toward a goal, the quality of the AI's policy depends on how well it models the human. Despite behavioral evidence that humans hyperbolically discount future rewards, the RL community continues to model humans as Markov Decision Processes (MDPs) with exponential discounting. This is because planning is difficult with non-exponential discounts. In this work, we investigate whether the performance benefits of modeling humans as hyperbolic discounters outweigh the computational costs. We focus on AI interventions that change the human's discounting (i.e. decreases the human's "nearsightedness" to help them toward distant goals). We derive a fixed exponential discount factor that can approximate hyperbolic discounting, and prove that this approximation guarantees the AI will never miss a necessary intervention. We also prove that our approximation causes fewer false positives (unnecessary interventions) than the mean hazard rate, another well-known method for approximating hyperbolic MDPs as exponential ones. Surprisingly, our experiments demonstrate that exponential approximations outperform hyperbolic ones in online learning, even when the ground-truth human MDP is hyperbolically discounted.

1 Introduction

In AI-assisted behavior change, an AI agent intervenes on human agents to influence them toward a goal state. For example, in digital interventions, a mobile health application may encourage users to do their daily physical therapy. Prior literature has found it useful to model a human agent's policy using a Markov Decision Process (as in Nofshin et al., 2024; Yu & Ho, 2022; Evans et al., 2016; Mintz et al., 2023). In our paper, we consider AI interventions that change the human's discount, or the degree to which they prioritize a faraway goal (Scholten et al., 2019). For example, the app may remind the user that adhering to physical therapy will enable them to return to a favorite sport. In this setting, the AI must model the human MDP well enough to plan high-quality interventions.

The human MDP includes a choice of discount function, which models how humans trade off future and immediate rewards. Behavioral science has overwhelmingly found that humans discount hyperbolically , $d_{\text{hyp}}(t) = 1/1+kt$, where k controls the level of discounting (e.g., Myerson & Green, 1995; Rachlin et al., 1991; Madden et al., 1999). Despite this, in reinforcement learning (RL), works on human-AI interaction continue to model humans as exponential discounters, $d_{\text{exp}}(t) = \gamma^t$, where γ controls the level of discounting (e.g., Giwa & Lee, 2021; Nofshin et al., 2024; Aswani et al., 2019; Mintz et al., 2023; Peysakhovich, 2019; Shah et al., 2019; Knox & Stone, 2012). This is because planning with exponential discounting is mathematically convenient; it lets us leverage the

majority of RL tools that depend on the Bellman Equation. On the other hand, planning with hyperbolic discounting is generally intractable and incurs significant computational costs to approximate. For example, a hyperbolic MDP may be approximated as the average of exponential MDPs (Fedus et al., 2019; Kurth-Nelson & Redish, 2009), but this requires re-solving for the optimal policy under several different exponential discount rates, γ. Unfortunately, no works have explored whether the policy improvements obtained by representing humans as hyperbolic discounters are worth the increase in model complexity, both computational and mathematical.

In this work, we ask whether there are alternatives to using hyperbolic discounting; in particular, can we cleverly select an exponential discount rate γ such that we still get a high-quality AI policy? We theoretically derive an exponential discount rate, γ_{safe}, which ensures the AI never misses a necessary intervention when modeling hyperbolic humans in a class of discrete, goal-oriented MDPs. Notably, while setting γ_{safe} requires knowledge of the human's hyperbolic discount rate k (which there are surveys to estimate (Kirby et al., 1999; Reynolds & Schiffbauer, 2004)), it does not require any information about the environment and can be used in practice when the transition dynamics are unknown. Furthermore, it incurs fewer false positives– unnecessary interventions to the user– compared to the well-known method of using an exponential discount model with *mean hazard rate* to approximate hyperbolic discount models (Rambaud & Torrecillas, 2005; Sozou, 1998; 2009).

Interestingly, when the AI learns the environment dynamics, we found that an AI planning with an exponential discount model always outperforms the hyperbolic one, even when the true human is hyperbolic. Despite predicting more accurate human Q-values, hyperbolic discounting causes more false negatives in the downstream AI policy. Furthermore, we found that learning γ online had worse performance than fixing it to γ_{safe}, especially in early episodes with less data. This work highlights the importance of carefully selecting a human discount model in human-AI settings, as different models impose trade-offs in AI planning. We demonstrate that defaulting to a hyperbolic discount is unnecessary since a well-chosen exponential discount rate can outperform a hyperbolic one while avoiding its computational and mathematical complexities.

2 Related Works

Evidence that humans are hyperbolic discounters. Behavioral science has shown that human discounting is better modeled with a hyperbolic, rather than exponential function on a wide range of tasks (e.g., Myerson & Green, 1995; Rachlin et al., 1991; Madden et al., 1999; Story et al., 2014). This is because hyperbolic functions can capture people's tendency to perform "preference reversal" (Myerson & Green, 1995); people who originally prefer a smaller reward sooner "flip" to preferring a larger reward later when asked the same question on a more distant timescale. Some work seeks to reduce this present bias (e.g., Callaway et al., 2022; Lieder et al., 2019; Muslimani et al., 2023). However, most of these studies formalize discounting in "one-off" decision settings and do not embed these discount models within a sequential decision-making framework (e.g., MDPs). Thus, it remains unclear whether, within the full MDP framework, modeling humans as hyperbolic discounters leads to better human-AI interactions. This question is underexplored due to the mathematically challenging nature of using non-exponential discount functions in RL planning.

Reinforcement learning with hyperbolic discounting. Planning with non-exponential discount functions is challenging because the Bellman equation no longer holds, and standard dynamic programming solutions cease to apply (Fedus et al., 2019). Despite this, recent works attempt to optimize value functions under non-exponential discounts through approximation (Fedus et al., 2019; Ali, 2023; Ali et al., 2024; Kurth-Nelson & Redish, 2009; Schultheis et al., 2022). For example, Fedus et al. (2019) and Kurth-Nelson & Redish (2009) approximate a hyperbolic Q-function by averaging over several exponential Q-functions. Recently, Schultheis et al. (2022) proposed an iterative, gradient-based solution to learn optimal values for continuous control. All these methods require parameter tuning (either the number of samples or the gradient parameters) to approximate the hyperbolic Q-function, yet no work has formally studied how these approximations impact the quality of the downstream AI policy in human-AI settings.

Approximating human agents as exponential discounters. RL literature largely models humans as exponential discounters (e.g., Giwa & Lee, 2021; Nofshin et al., 2024; Aswani et al., 2019; Mintz et al., 2023; Peysakhovich, 2019; Shah et al., 2019; Knox & Stone, 2012). This approach requires specifying a discount rate, γ, a priori. Some works learn a fixed γ from a batch of data (Aswani et al., 2019; Mintz et al., 2023), while others fix γ to one that simulates realistic behaviors (Peysakhovich, 2019). In contrast, we propose a fixed γ that depends only on the human's hyperbolic discount rate k, and no other data or domain knowledge. The economics literature defaults to the *mean hazard rate* (MHR), (e.g., Rambaud & Torrecillas, 2005; Sozou, 1998; 2009). We challenge the MHR as the default γ in our setting because it leads to AI policies that *over-intervene*. Instead of fixing γ a-priori, other methods learn it online (e.g., Nofshin et al., 2024; Yu & Ho, 2022; Yu et al., 2024; Zhou et al., 2018; Evans et al., 2016). However, these methods do not explore how misspecification of the discount model (i.e. the assumption of exponential discounting) affects the AI policy.

3 Background

Hazards: Relating Hyperbolic and Exponential Discounting. A Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, R, P, d \rangle$ is a tuple of states $\mathcal{S}$, actions $\mathcal{A}$, a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. The discount function $d(t)$ devalues future rewards and can be interpreted as the probability of surviving to timestep t. Exponential discounting takes the form $d_{\text{exp}}(t) = \gamma^t$ for $\gamma \in [0, 1]$, and hyperbolic discounting takes the form $d_{\text{hyp}}(t) = {}^1\!/\!_{(1+kt)}$ for $k \in [0, \infty)$. Note: in exponential discounting, a smaller γ is more myopic, but in hyperbolic, a larger k is more myopic. Exponential and hyperbolic discounting can be related through the *hazard rate*, a concept from reliability engineering that describes how one's probability of survival changes over time: $h(t) = -\frac{d}{dt} \ln d(t)$, where a high hazard corresponds to a sharply decreasing probability of survival. A constant hazard rate is equivalent to exponential discounting:

$$h(t) = \lambda \to d(t) = \exp(-\lambda t) \to d(t) = \gamma^t, \text{ where } \gamma = \exp(-\lambda).$$

On the other hand, Sozou (1998) proved that if hazard follows an exponential distribution, such that $p(\lambda) = ({}^1\!/\!_k) \exp(-{}^\lambda\!/\!_k)$, then this is equivalent to hyperbolic discounting:

$$d(t) = \int_{\lambda=0}^{\infty} \exp(-\lambda t) p(\lambda) d\lambda = \frac{1}{1+kt}. \tag{1}$$

The hazard rate provides a natural way to approximate hyperbolic discounting with an exponential one. By setting γ to the mean of the exponential distribution on the hazard ($\mathbb{E}[\lambda] = k$), we recover the well-known *mean hazard rate* (mhr): $\gamma_{\text{mhr}} = \exp(-k)$. Finally, Eq. (1) provides an estimate of hyperbolic Q-values as an expectation over exponential ones:

$$Q_{\text{hyp}}(s, a; k) = \mathbb{E}_{\gamma \sim \text{Beta}({}^1\!/\!_k, 1)} \left[Q^*_{\text{exp}}(s, a; \gamma) \right], \tag{2}$$

where $Q^*_{\text{exp}}(s, a; \gamma)$ is the optimal value at state s and action a for an exponentially-discounted MDP.

Behavior Model RL: an AI Agent that Intervenes on a Human Agent's MDP. To study settings where an AI agent guides human agents to a goal state, we use the behavior model RL (BMRL) framework from Nofshin et al. (2024), where AI actions change the human agent's MDP parameters, as shown in Fig. 1. Throughout, unless subscripted with "AI," entities belong to the human agent. In BMRL, the AI agent is an MDP $\mathcal{M}^{AI} = \langle \mathcal{S}^{AI}, \mathcal{A}^{AI}, R^{AI}, P^{AI}, \gamma^{AI} \rangle$. **AI actions** $\mathcal{A}^{AI}$ are interventions that cause *temporary* changes to the human agent's MDP parameters; following an AI intervention, the human MDP changes from $\mathcal{M}$ to $\mathcal{M}'$, then reverts to $\mathcal{M}$ the next time-step. We consider a binary action space: the AI either intervenes on the human's discounting to make them more farsighted ($a^{AI} = 1$) or does nothing ($a^{AI} = 0$). Explicitly, if the AI intervenes $a^{AI} = 1$, then the human's MDP changes from $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, R, P, k \rangle$ to $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, R, P, k' \rangle$, where the only difference is to make the the hyperbolic discount factor more farsighted, from k to $k' = k - \delta_k$. Here, δ_k is the change to the discount factor.

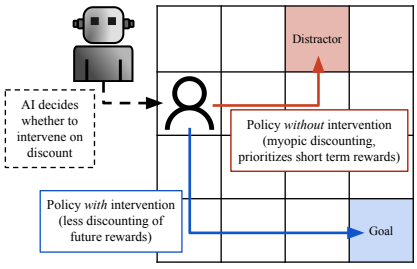

Figure 1: BMRL: the AI nudges the human toward a goal by altering their MDP (e.g., increasing their discount).

Because the AI agent's actions change the human's MDP, the human's MDP (and discount model) is part of the AI environment. Formally, the **states** $\mathcal{S}^{AI}$ are the same as the human states, but also include the human's action from the *last* timestep α, so that $s^{AI} = [s, \alpha]$. This causes the **transitions** factorize into two distributions: $P^{AI}(s'^{AI}|s^{AI}, a^{AI}) = P(s'|s, \alpha')\pi(\alpha'|s, a^{AI})$, where $\pi(\alpha'|s, a^{AI})$ is the effect of the AI intervention on the human's action, and $P(s'|s, \alpha')$ is the effect of the human action on the next state. Note that π depends on the human MDP $\mathcal{M}'$ that results from an AI intervention a^{AI}.

4 Problem Setting and Formulation

Setting: Absorbing state MDPs. Our human agents act in a discrete class of MDPs that represent the behavior change setting. There are N absorbing states, and $s^{(n)}$ refers to the nth absorbing state. One of the absorbing states, $s^{(N)}$, is the "goal" state (e.g. doing physical therapy). The remaining $s^{(1)}, \ldots, s^{(N-1)}$ absorbing states are "distractors" (e.g. watch TV instead). The reward at the goal state $r^{(N)} = 1$ is larger than all others $r^{(1)}, \ldots, r^{(N-1)} \in (0, 1)$. Even though the goal reward is largest, the human agent may still choose the distractor for its proximity. Finally, a per-timestep reward $r_b < 0$ represents the burden of behavior change (Baumeister & Vohs, 2007; Nofshin et al., 2024), and incentivizes the human agent to settle for nearby absorbing states.

Absorbing state MDPs are general and encapsulate several environments from the literature, such as those from Evans et al. (2016); Peysakhovich (2019); Ankile et al. (2023); Nofshin et al. (2024). See Appendix D.3 for in-depth examples.

Problem formulation: Approximating Hyperbolic Human MDPs for High-Quality AI Interventions. Following behavioral science, our true human agents discount hyperbolically. The AI agent intervenes on the human agent's MDP to help them reach the goal. We focus on AI interventions that target the human's *discounting*, so that when $a_{\text{AI}} = 1$, γ increases to $\gamma' = \gamma + \delta_\gamma$ or k decreases to $k' = k - \delta_k$, depending on whether the AI assumes a hyperbolic or exponential discount. Note: $\delta_k, \delta_\gamma > 0$ are changes to the discount factor, also called the *intervention effect*.

We aim to understand how misspecifications of the discount model impact optimal AI policies. Optimal AI policies solve the AI MDP defined in section 3 where actions are interventions. Specifically, we study differences in AI policies arising from approximating hyperbolic human agents as exponential ones; cases where $\pi^{AI}_{\text{exp}}(s; \gamma) \neq \pi^{AI}_{\text{hyp}}(s; k)$, for some $s \in \mathcal{S}^{AI}$. Here, $\pi^{AI}_{\text{exp}}(\cdot; \gamma)$ is an AI policy that uses an exponential discount model of the human with parameter γ.

In our AI assisted behavior change setting, false negatives (missing necessary interventions) are more harmful than false positives (delivering unnecessary interventions). Missing an intervention means the human will not accomplish their behavioral goal, while excessive interventions ensure goal achievement but annoy the user. Formally, false negatives are cases where the hyperbolic AI policy intervenes but the exponential AI policy withholds:

$$FN_s(\gamma, k) = \mathbb{I}\left\{\pi^{AI}_{\text{hyp}}(s; k) = 1 \text{ and } \pi^{AI}_{\text{exp}}(s; \gamma) = 0\right\}. \tag{3}$$

Here, $\mathbb{I}\{\cdot\}$ is the indicator function. Likewise, false positives are cases where the hyperbolic AI agent withholds intervention but the exponential AI agent intervenes: $FP_s(\gamma, k) = \mathbb{I}\left\{\pi^{AI}_{\text{hyp}}(s; k) = 0 \text{ and } \pi^{AI}_{\text{exp}}(s; \gamma) = 1\right\}$.

We have two goals:

1. Identify an exponential approximation of the hyperbolic human that guarantees no false negatives and minimizes false positives in AI policy. In section 5, we identify γs that solve the following

optimization problem:

$$\min_{\gamma} \sum_{s \in \mathcal{S}} FP_s(\gamma, k), \quad \text{s.t.} \sum_{s \in \mathcal{S}} FN_s(\gamma, k) = 0. \tag{4}$$

2. Identify the best approximation (exponential or otherwise) of the human's discounting function when learning the AI policy online. In most real-life settings, the transition dynamics of the behavior change setting are unknown. In section 6, we perform empirical experiments that compare approximate discount models when transitions are learned online.

5 Theoretical Analysis

We identify solutions to the optimization problem in Eq. (4), which ensures the AI will not miss necesssary interventions while minimizing uncessesary ones. First, we characterize which γs guarantee no false negatives in the AI intervention policy. Then, we prove that the larger the γ, the fewer the false positives. We use this fact to propose two solutions for γ that require different levels of knowledge. One is a state-specific γ, which relies on knowledge of the environment transitions and human's hyperbolic discount rate k. The other solution still requires k but does not assume knowledge of the environment transitions. [1] For brevity, our analysis focuses on the choice between the goal state and only a single distractor state. When multiple distractors exist, only the highest-valued distractor is relevant, so this reduces to the same pairwise comparison.

5.1 Guaranteeing No False Negatives

We characterize exponential discount rates that guarantee no false negatives. Intuitively, to prevent false negatives, we want our exponential approximation to be "conservative," meaning it *underestimates* the human's preference for the goal state. This way, we never miss an intervention by incorrectly assuming that the human agent's policy will reach the goal without intervention. Definition 1 formally defines a "conservative" exponential approximation; whenever the hyperbolic agent values a distractor state over the goal state, the exponential agent must also prefer the distractor.

Definition 1 (Conservative exponential approximation)**.** Let $\pi^{(n)}$ refer to a policy whose actions lead to absorbing state $s^{(n)}$ and $V^{(n)}$ refer to the value of following this policy. Suppose the human agent is hyperbolic with discount rate k. An exponential approximation of the agent is *conservative* if, for all states $s \in \mathcal{S}$ where the hyperbolic agents prefers the distractor state $V^{(n)}_{\text{hyp}}(s; k) \geq V^{(N)}_{\text{hyp}}(s; k)$, the exponential agent also prefers the distractor state $V^{(n)}_{\text{exp}}(s; \gamma) \geq V^{(N)}_{\text{exp}}(s; \gamma)$.

In order for a conservative exponential approximation to guarantee no false negatives, the AI must also assume that the *intervention effect*, δ_γ, is sufficiently large. Under a conservative γ, the AI always recognizes when the human prefers the distractor and thus never misses opportunities to intervene. However, it may still withhold intervention if the effect is too weak to alter the outcome. The simplest way to ensure δ_γ is sufficiently large is to assume maximal effectiveness; $\delta_\gamma = 1 - \gamma$. In Theorem 2, we prove that an exponential approximation using a conservative γ and $\delta_\gamma = 1 - \gamma$ implies no false negatives in the AI intervention policy.

Theorem 2 (Conservative means no false negatives)**.** *Let the true human agent discount hyperbolically with parameter k and that AI interventions reduce this parameter by δ_k. If the AI agent plans using an exponential approximation with conservative γ (under Definition 1) and maximal intervention effect $\delta_\gamma = 1 - \gamma$, then there are no false negatives: $\sum_{s \in \mathcal{S}} FN_s(\gamma, k) = 0$.*

Proof. In Appendix A.1, we present a proof by contradiction.

[1] In psychology, there are known ways to estimate a human's k, such as the Monetary Choice Questionnaire survey Kirby et al. (1999)

Solving for a conservative exponential discount rate γ. We now characterize what γ's are conservative (and by extension of Theorem 2, what γ's guarantee no false negatives). To facilitate our theoretical characterization, we make three assumptions: the transitions of the MDP are deterministic, there is no burden (intermediate rewards), and human policies are deterministic. Under these assumptions, we derive closed-form solutions to value functions in absorbing state MDPs. Later, in Section 6, we demonstrate that our results hold empirically when the assumptions are relaxed.

Let $\ell^{(n)}$ refer to the length of the path from state s to absorbing state $s^{(n)}$ under deterministic policy $\pi^{(n)}$. Then the value functions for hyperbolic and exponential discounting are:

$$V_{\text{hyp}}^{(n)}(s;k) = \frac{1}{1+k(\ell^{(n)}-1)} r^{(n)} \qquad V_{\text{exp}}^{(n)}(s;\gamma) = \gamma^{\ell^{(n)}-1} r^{(n)}. \tag{5}$$

In Theorem 3, we leverage these closed-form solutions to derive conditions under which γ guarantees a conservative approximation.

Theorem 3 (Characterizing conservative γ)**.** *Suppose the true human agent discounts hyperbolically with parameter k. Suppose the distractor state $s^{(n)}$ has reward $r^{(n)}$. Let $\ell^{(n)}$ refer to the length of the deterministic path from state $s \in \mathcal{S}$ to $s^{(n)}$. Let $\Delta = \ell^{(N)} - \ell^{(n)} \geq 1$ refer to the difference in distance between the goal and distractor state. If the exponential agent uses a discount rate of γ satisfying the following, then the exponential agent is a conservative approximation:*

$$\gamma \leq \left(\frac{1 + k\ell^{(n)} - k}{1 + k\ell^{(n)} + k\Delta - k} \right)^{\frac{1}{\Delta}}, \tag{6}$$

Proof. Algebraic derivation in Appendix A.2.

5.2 Minimizing False Positives

Ruling out false negatives alone does not ensure good AI policies. We must also minimize false positives, which occur when the AI intervenes even though the human agent would have reached the goal state without intervention. This scenario involves three conditions (details in Appendix C.3):

C1 The true hyperbolic agent with discount k prefers the goal.

C2 The exponential approximation with discount γ prefers the distractor.

C3 The exponential approximation *under intervention* with discount $\gamma + \delta_\gamma$ prefers the goal.

Using the above, we can formalize which parameters will cause a false positive. C2 implies $V_{\text{exp}}^{(n)}(s;\gamma) \geq V_{\text{exp}}^{(N)}(s;\gamma)$. C3 implies that $V_{\text{exp}}^{(n)}(s;\gamma+\delta_\gamma) < V_{\text{exp}}^{(N)}(s;\gamma+\delta_\gamma)$. C1 can be ignored because the exponential approximation does not affect it. Together, C2 and C3 imply: $(\gamma+\delta_\gamma)^{\ell^{(n)}-\ell^{(N)}} \leq {}^{r^{(N)}}\!/_{r^{(n)}} \leq \gamma^{\ell^{(n)}-\ell^{(N)}}$. Our choice of γ affects how often this condition is met. We define the "broadness" of the condition as a function of γ, which we call the false positive range:

$$FP_{\text{range}}(\gamma) = \gamma^{\ell^{(n)}-\ell^{(N)}} - (\gamma+\delta_\gamma)^{\ell^{(n)}-\ell^{(N)}}. \tag{7}$$

A larger FP_{range} means more false positives, since more reward pairs $(r^{(n^*)}, r^{(N)})$ will satisfy the condition. In Theorem 4, we show that FP_{range} decreases with γ, so larger γ reduce false positives.

Theorem 4. *Let $\ell^{(n)}$ and $\ell^{(N)}$ refer to the length to distractor and goal state from state s. Consider two exponential approximations, which use discount rates γ_1 and γ_2. Both approximations assume the same intervention effect, δ_γ. If $\gamma_1 > \gamma_2$, then $FP_{range}(\gamma_1) < FP_{range}(\gamma_2)$.*

Proof. In Appendix A.4, we take the derivative of FP_{range}.

5.3 Solutions

In Section 5.1, we proved γ must be small enough to avoid false negatives, and in Section 5.2 we proved that larger γ result in fewer false positives. This implies a natural solution to our optimization

problem in Eq. (4); we set γ to be the largest value in Eq. (6), so that $\gamma_s = \left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(n)}+k\Delta-k}\right)^{\frac{1}{\Delta}}$. However, setting γ_s relies on distances to the goal and distractor, which are derived from the transition dynamics. However, we may not have access to the transition dynamics for real-world applications.

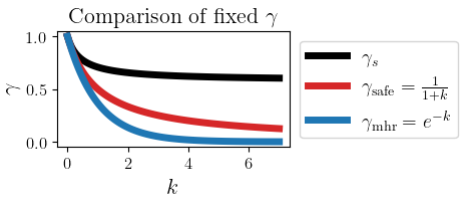

Figure 2: Comparison of the state-specific γ_s, the mean hazard rate γ_{mhr}, and our proposed γ_{safe} as a function of k. MHR is always smaller than ours.

An exponential approximation, γ_{safe}, that only requires k. Instead requiring access to transitions, we lower bound γ_s by assuming the "worse-case" values of $\Delta = 1$ and $\ell^{(n)} = 1$. This reduces to an exponential discount rate of

$$\gamma_{\text{safe}} = \frac{1}{1+k}. \tag{8}$$

Since $\gamma_{\text{safe}} \leq \gamma_s$, it is conservative and guarantees no false negatives (see Appendix A.3).

The mean hazard rate yields more false positives. Our theory allows us to analyze an exponential approximation with $\gamma_{\text{mhr}} = \exp(-k)$. Since $\gamma_{\text{mhr}} < \gamma_{\text{safe}}$ (see Appendix C.2), it avoids false negatives but incurs more false positives.

6 Empirical Analysis

When learning online, our decision to approximate hyperbolic humans as exponential reduces variance but increases bias, which means that the AI can guide users to their goals faster at the long-term cost of sending more interventions. Our experiments test whether this trade-off is worthwhile, especially compared to the low bias, high variance alternative of using a hyperbolic approximation, which will take longer to help users reach their goals. Furthermore, our experiments relax assumptions of the theory to test its generalizability: the transitions are unknown, the efficacy of the AI intervention varies for each human, there is burden, and humans are not deterministic. Results with stochastic transitions are in Appendix D.4.1 (omitted because main results remain unchanged).

Experimental setup. The experiments are in randomly sampled absorbing state MDPs with $8-20$ states and 2 actions. The deterministic transitions are sampled from a categorical distribution. We filter for valid transitions where every absorbing state is reachable from any state. No generality is lost by considering a binary action space; adding more actions would only increase transition complexity, which we already vary by adjusting the state space size. The range of $8-20$ states allows us to observe results across a class of small tabular MDPs. There is one distractor state with reward $r^{(n)} \sim \text{Unif}(0.1, 0.5)$, a goal state with reward $r^{(N)} = 1$ and burden $r_b \sim$ Unif(-0.5, -0.01). Human agents are hyperbolic with discount $k \sim \text{Unif}(0.1, 5)$. Interventions decrease this by $\delta \sim (0.09, k)$. Following precedent (e.g., Reddy et al., 2018; Laidlaw & Dragan, 2022), our human agents are "Boltzmann rational," actors who follow stochastic softmax policies with a small temperature $\tau = 0.02$, which ensures they can reach the goal under the oracle AI policy (if the human is too random, even good AI policies will not help). The AI agent receives a reward of 1 when the human agent enters the goal, -1 at a distractor, and -0.1 for intervening (further details on the AI MDP in appendix D.1). Episodes start in states where the optimal AI policy intervenes; otherwise, outcomes wouldn't depend on the policy. This setup highlights differences between methods.

Baselines for modeling the human agent. Table 1 describes our baselines. All methods except the oracle estimate the transitions, by normalizing the observed counts of transitions (MLE). AI actions are selected according to an ϵ-greedy policy with $\epsilon = 0.1$; interventions are random 10% of the time and maximize the optimal value function under the estimated transitions for the remaining (certainty-equivalence RL). The hyperbolic baseline in our main experiments use Monte-Carlo estimation with 500 samples to approximate the expectation in Eq. (2), but we also compare alternate approximation methods in our experiments. We include the fixed-γ_{big} baseline to demonstrate what happens when

Table 1: Experimental baselines, which differ in their model of the human's discount

Baseline	Transitions	Discount model	Intervention
oracle	True T	True hyperbolic k	True δ_k
hyp-mcmc (Fedus et al., 2019)	Learned	Hyperbolic approx. Eq. (2)	True δ_k
fixed-γ_{safe}	Learned	Exponential, fixed to $1/1+k$	Max
fixed-γ_{mhr}	Learned	Exponential, fixed to $\exp(-k)$	Max
fixed-γ_{big}	Learned	Exponential, fixed to $\gamma_{\text{safe}} + 0.2$	Max
learning-γ (Nofshin et al., 2024)	Learned	Exponential, γ is learned	δ_γ learned
always-intervene	–	–	–

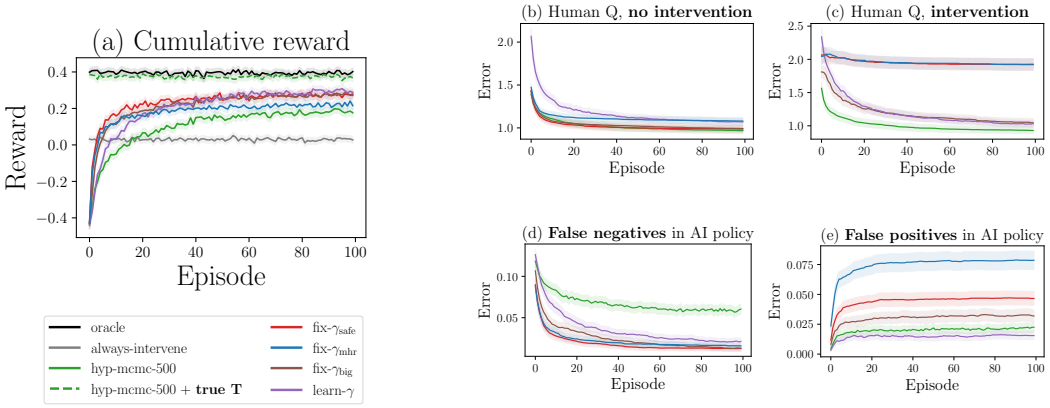

Figure 3: **Hyperbolic approximation (green) predicts the human value well, but leads to more false negatives and worse AI policies.** Error bars are 95% CI over 5000 trials (1000 random MDPs, 5 runs each). First row is error in human value prediction, second row is error in AI policy.

an exponential model with a discount factor larger than ours is used (the value of γ is clipped at 0.99). Always-intervene is a naive strategy that intervenes every timestep.

6.1 Results

Approximation error in the hyperbolic method means AI policies fail to intervene when needed. Surprisingly, in Fig. 3a, the exponential methods outperform the hyperbolic approximation, even though the true human is hyperbolic. The poor performance of the hyperbolic approximation is due to the variance of learning the transitions; with true transitions, its performance matches the oracle. What causes the hyperbolic approximation to have low rewards when learning the transitions? In Fig. 3b and Fig. 3c, the hyperbolic approximation better predicts human value functions, which means it is generally better at anticipating user intentions. But, in Fig. 3d, we see the few user misunderstandings it does make lead to more false negatives in the AI policy (i.e. assumes the user prefers the goal state when they do not).

Our γ_{safe} strikes the right balance of minimizing false negatives and avoiding false positives in the AI policy. Our theory indicates that γ_{safe} and γ_{mhr} will prevent false negatives in the AI policy when the transitions are known, meaning an AI policy that uses γ_{safe} or γ_{mhr} will intervene enough to get the user to the goal state. Fig. 3d and Fig. 3e shows that our theory generalizes to when the transitions are learned; fixed-γ_{safe} and fixed-γ_{mhr} learn AI policies with the fewest false negatives– meaning they help user reach the goal more consistently– but fixed-γ_{safe} has fewer false positives– meaning it is less likely to annoy the user. Naturally, one might wonder how fixed-γ_{safe} and fixed-γ_{mhr} compare to the strategy of always intervening, which has a false negative rate of 0. Though not shown in Fig. 3e for visualization reasons, the always-intervene baseline has a false-positive

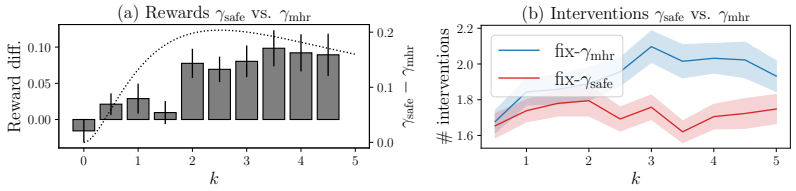

Figure 4: **Gap between fixed-γ_{safe} and fixed-γ_{mhr} is bigger for larger k (humans are more myopic).** Fig. 4a shows reward difference, averaged over all episodes, increases with k. Dotted line is $\gamma_{\text{safe}} - \gamma_{\text{mhr}}$. Fig. 4b shows fixed-$\gamma_{\text{mhr}}$ intervenes more than fixed-γ_{safe} .

rate of 0.8, much higher than even that of fixed-γ_{mhr} at 0.075. As a result of over-intervening, the always-intervene baseline *overburdens* the human and has low overall reward in Fig. 3a.

Finally, fixed-γ_{big} demonstrates that fixed-γ_{safe} is *not too big*. The fixed-γ_{big} baseline incurs more false negatives than fixed-γ_{safe} and lower overall reward in Fig. 3 (this difference is more apparent when humans are optimal in Appendix D.4.2). Thus, we see that γ_{safe} is "just right"; it is conservative enough to intervene on the human when the goal is at stake, but big enough to avoid over-intervening.

The fixed-γ_{mhr} baseline over-intervenes more severely on human agents that are more myopic. Our theory indicates that γ_{mhr} is always smaller than γ_{safe} for humans with the same k, meaning fixed-γ_{mhr} will intervene more on a given user. This is why fixed-γ_{safe} outperforms fixed-γ_{mhr} in Fig. 3a. This performance gap increases for larger k (see Fig. 4a), as the difference in γs– and therefore the difference in false-positives– also increases between the two methods (see Fig. 4b). For small enough k in Fig. 4a, fixed-γ_{mhr} outperforms fixed-γ_{safe} . However, this describes a setting in which the human is already far-sighted, which is less relevant in practice, since far-sighted users are unlikely to need help prioritizing faraway goals.

Fixing γ is better than learning it when the inductive bias aligns with the true environment. When a small γ accurately models the human's behavior, fixed-γ_{safe} outperforms learning-γ by avoiding the cost of learning. For example, a small discount such as γ_{safe} is plausible when the goal is close, because the human agent must be more myopic to prefer the distractor reward. Fig. 5b confirms that the likelihood of the γ_{safe} is higher when the goal is close, and correspondingly, the advantage of fixed-γ_{safe} is more pronounced in Fig. 5a. The fact that fixed-γ_{safe} does worse as the goal grows more distant (and γ_{safe} no longer models the MDP well) suggests that a good strategy may be to use γ_{safe} as a prior, and then learn a more likely γ as more data becomes available.

When the transitions are unknown, regardless of approximation method, the hyperbolic model has worse performance and worse computational efficiency than exponential. Modeling hyperbolic agents requires approximating the expectation over exponential Q-values from Eq. (2). Thus far, our experiments have demonstrated that an MCMC approximation with 500 samples is insufficient for good performance. Fig. 6 further illustrates that fixed-γ_{safe} outperforms a hyperbolic approximation *regardless* of the method used. Because we did not observe substantial differences among different estimation methods in this ablation, in our other experiments we only considered MCMC estimation. Fig. 6b demonstrates that increasing the number of samples does not help, because the hyperbolic methods are still sensitive to the quality of estimated transitions. Overall, hyp-mcmc consumes several orders of magnitude more computation (Fig. 6c) while still failing to meet the performance of fixed-γ_{safe} managed with *no prior engineering effort*.

7 Discussion and Future Work

Estimating k. In this work, we investigated the impact of approximating hyperbolic humans as exponential discounters on AI intervention policy. We proposed an exponential discount rate, γ_{safe}, whose initialization *does not* depend on knowledge of an absorbing-state MDP's transitions, but

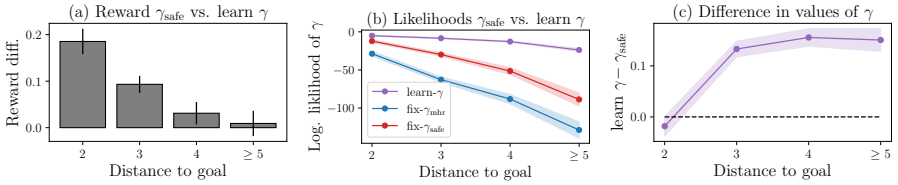

Figure 5: **Gap between fixed-γ_{safe} and learning-γ is bigger when goal is close; γ_{safe} provides appropriate inductive bias.** Fig. 5a shows reward differences, averaged over first 40 episodes, decreases when goal is farther. Fig. 5b shows smaller γ's, e.g. γ_{safe} (red) and γ_{mhr} (blue), are more likely when goal is close. Fig. 5c shows that γ from learning-γ is bigger than γ_{safe} when goal is far.

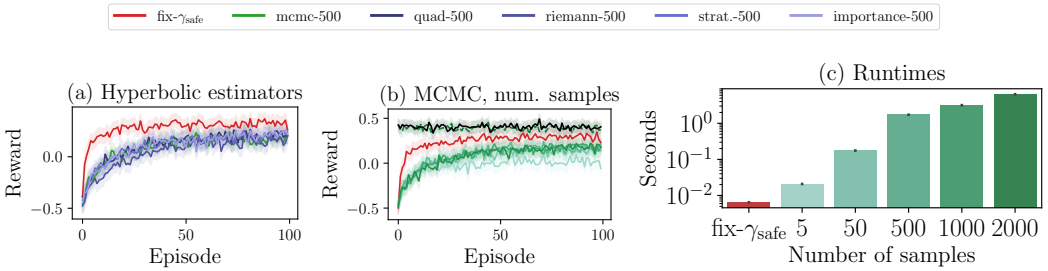

Figure 6: **Hyperbolic approximations, regardless of method and sample size, perform worse than fixed-γ_{safe} and are orders of magnitude more computationally expensive.** In Fig. 6, no approximation method (defined in Appendix D.2) matches the performance of fixed-γ_{safe} . In Fig. 6b, increasing the number of samples has diminishing returns when learning transitions (dashed line is given true transitions). Fig. 6b shows the runtime cost of increasing samples per timestep.

does require knowledge of the human's hyperbolic discount rate, k. In practice, k can be estimated using known surveys (Kirby et al., 1999), and an interesting future direction to study the extent to which surveys can provide accurate measures for k for AI agent planning. Furthermore, we note that needing an estimate for k is not a unique limitation of our method – estimating k is necessary even when using a fully hyperbolic model or the mean hazard rate.

Generalization to other human-AI interaction paradigms. In our AI intervention setting, we found that exponential methods outperformed the hyperbolic approximators, even when human agents were truly hyperbolic. This raises questions about whether careful selection of the exponential discount γ can match– or even surpass– the performance of hyperbolic approximation in other human-AI interaction settings. For example, in inverse reinforcement learning, the goal is to infer the human's other MDP parameters, such as the reward. Recent work has started to explore inverse learning under non-exponential discounts (Yao et al., 2024), but it is worth considering whether there is an exponential discount rate that would suffice.

Beyond absorbing state MDPs. Our results are on absorbing state MDPs, where there is one absorbing goal state and multiple distractor states. Although this class of MDPs covers several worlds considered in recent literature, they do not encompass all the behavior settings we might want to study. It would be interesting to see how our proposed $\gamma = 1/1+k$, which we derived specifically for absorbing state MDPs, generalizes to worlds outside of this class, such as ones with more complex intermediate rewards than burden.

Preference reversal. While we considered hyperbolic discount in our MDPs, we did not include preference reversal in our formalization. To do so, we would have to incorporate replanning, since preference reversal occurs because the agent has a *time dependent policy*; the policy in one timestep

(i.e., looking far into the future) is different from the policy in the other (i.e., considering the "now"). For example, Yu & Ho (2022) implement replanning by changing the definition of value functions; they account for value at a current and future timestep. Modeling pre-commitment would allow us to consider more AI interventions, such as pre-commitment, where humans are encouraged to "pre-commit" to a goal-preferring policy (e.g., Yi et al. (2019)). It is unclear whether it is possible to plan pre-commitment interventions when the AI uses an exponential human model.

Conclusion. In this paper, we addressed a mismatch in how human decisions are modeled in behavioral science (as hyperbolic discounters) and RL (as exponential discounters). We examined the extent to which humans' hyperbolic discounting is approximated by a carefully chosen exponential discount model. In our intervention setting, we found that hyperbolic approximations of the human agent led to worse AI policies than an exponential one using out theoretically-justified discount rate, γ_{safe}. We also showed that γ_{safe} is as general as the well-known γ_{mhr}, but with fewer false positives, which decreases unnecessary interventions. Our work highlights that defaulting to a hyperbolic model is not the best strategy, particularly given its additional computational costs, and we encourage AI researchers who work with human agents to evaluate the trade-offs between different exponential models (including γ_{safe}) and a hyperbolic one in their specific applications.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391 and the NIH/NIBIB and OD P41EB028242. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This paper describes work performed at Harvard University and is not associated with Amazon.

A Appendix

A.1 Proof for Theorem 2: conservative γ means no false negatives

We proceed by contradiction. Let γ be a conservative exponential approximation. By definition of the conservative exponential approximation, we have that $V^{(n)}_{\text{hyp}}(s;k) \geq V^{(N)}_{\text{hyp}}(s;k)$ under no intervention, implies $V^{(n)}_{\text{exp}}(s;\gamma) \geq V^{(N)}_{\text{exp}}(s;\gamma)$– i.e. when the ground-truth hyperbolic agent prefers the distractor state, so does the exponential approximation.

Suppose that the exponential approximation by γ results in a false negative at s (in Eq. (3)). By the definition, we must have that: $\pi^{AI}_{\text{hyp}}(s) = 1$ and $\pi^{AI}_{\text{exp}}(s) = 0$. It follows from the assumption that γ is conservative that $\pi^{AI}_{\text{hyp}}(s) = 1 \implies V^{(n)}_{\text{hyp}}(s;k) \geq V^{(N)}_{\text{hyp}}(s;k) \implies V^{(n)}_{\text{exp}}(s;\gamma) \geq V^{(N)}_{\text{exp}}(s;\gamma)$. There are two cases that $\pi^{AI}_{\text{exp}}(s) = 0$ could be true:

1. Suppose $\pi^{AI}_{\text{exp}}(s) = 0$ because $V^{(n)}_{\text{exp}}(s;\gamma+\delta_\gamma) \geq V^{(N)}_{\text{exp}}(s;\gamma+\delta_\gamma)$. But, by assumption we have $\gamma + \delta_\gamma = 1$. This means that $V^{(n)}_{\text{exp}}(s;\gamma+\delta_\gamma) \geq V^{(N)}_{\text{exp}}(s;\gamma+\delta_\gamma) \implies r^{(n)} \geq r^{(N)}$, noting that by assumption we have that $\gamma + \delta_\gamma = 1$, so therefore in fact $V^{(n)}_{\text{exp}}(s;\gamma+\delta_\gamma) = r^{(n)}$ and $V^{(N)}_{\text{exp}}(s;\gamma+\delta_\gamma) = r^{(N)}$. However, by our problem formulation in Section 4, we must have $r^{(n)} < r^{(N)}$. So, this case does not hold.
2. Suppose $\pi^{AI}_{\text{exp}}(s) = 0$ because $V^{(n)}_{\text{exp}}(s;\gamma) < V^{(N)}_{\text{exp}}(s;\gamma)$. Recall that we had $V^{(n)}_{\text{exp}}(s;\gamma) \geq V^{(N)}_{\text{exp}}(s;\gamma)$. Thus, we have a contradiction, and this case does not hold.

Both cases cannot hold, thus it must be that $\pi^{AI}_{\text{exp}}(s) = 1$.

A.2 Proof for Theorem 3: characterizing conservative γ

By construction, the ground truth hyperbolic agent prefers the distractor state, i.e. $V^{(n)}_{\text{hyp}}(s) \geq V^{(N)}_{\text{hyp}}(s)$. Using the definition of hyperbolic value functions in Eq. (5), we solve the inequality

for a constraint on the reward at the distractor state, $r^{(n)}$:

$$V_{\text{hyp}}^{(n)}(s) \geq V_{\text{hyp}}^{(N)}(s) \implies \frac{r^{(n)}}{1+k\ell^{(N)}-k} \geq \frac{r^{(N)}}{1+k\ell^{(n)}-k} \implies r^{(n)} \geq r^{(N)}\left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(N)}-k}\right). \quad (9)$$

Similarly, we solve for the constraint on the distractor state reward in exponential value functions:

$$V_{\text{exp}}^{(n)}(s) \geq V_{\text{exp}}^{(N)}(s) \implies \gamma^{\ell^{(n)}-1}r^{(n)} \geq \gamma^{\ell^{(N)}-1}r^{(N)} \implies r^{(n)} \geq \gamma^{\ell^{(N)}-\ell^{(n)}}r^{(N)}. \quad (10)$$

We want a γ such that a hyperbolic agent's preference of the distractor state (Eq. (9)) implies that the exponential agent will prefer the same (Eq. (10)):

$$r^{(n)} \geq r^{(N)}\left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(N)}-k}\right) \implies r^{(n)} \geq r^{(N)}\gamma^{\ell^{(N)}-\ell^{(n)}}.$$

It suffices to show that $r^{(N)}\left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(N)}-k}\right) \geq r^{(N)}\gamma^{\ell^{(N)}-\ell^{(n)}}$. Solving this inequality for γ:

$$r^{(N)}\left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(N)}-k}\right) \geq r^{(N)}\gamma^{\ell^{(N)}-\ell^{(n)}} \quad (11)$$

$$\implies \gamma^{\ell^{(N)}-\ell^{(n)}} \leq \left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(N)}-k}\right) \qquad \Delta = \ell^{(N)}-\ell^{(n)} \quad (12)$$

$$\implies \gamma \leq \left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(n)}+k\Delta-k}\right)^{\frac{1}{\Delta}}. \quad (13)$$

A.3 Proof that $\gamma_{\text{safe}} \leq \gamma_s$

We show that $\gamma_{\text{safe}} \leq \gamma_s$, meaning that γ_{safe} is conservative (i.e. guarantees no false negatives). First, note that γ_s is *increasing* with respect to $\ell^{(n)}$. The derivative of γ_s with respect to $\ell^{(n)}$ is:

$$\underbrace{\frac{k}{\Delta}}_{(a)}\underbrace{\left(\frac{1+k(\Delta+\ell^{(n)}-1)}{1+k(\ell^{(n)}-1)}\right)^{\left(1-\frac{1}{\Delta}\right)}}_{(b)}\underbrace{\left(\frac{\Delta k}{(1+k(\Delta+\ell^{(n)}-1))^2}\right)}_{(c)} \quad (14)$$

Part $(a) > 0$ because $k \geq 0$ and $\Delta > 0$. Part $(b) > 0$ because $\ell^{(n)} \geq 1$ and all the other terms are positive. Part $(c) > 0$ for the same reason. So, we know the function is increasing with respect to $\ell^{(n)}$. Since γ_s is increasing with respect to $\ell^{(n)}$, we can lower bound it by substituting the lowest possible value of $\ell^{(n)} = 1$. Note that if $\ell^{(n)} = 0$, then the agent would be in an absorbing state.

The value of γ_s is then: $\gamma_s \geq \left(\frac{1}{1+k\Delta}\right)^{\frac{1}{\Delta}}$. Again, the derivative shows that this *increases* with Δ:

$$\frac{\partial}{\partial\Delta}\left(\frac{1}{1+k\Delta}\right)^{\frac{1}{\Delta}} = -\underbrace{\frac{1}{\Delta^2}}_{(a)}\underbrace{\left(\frac{1}{1+k\Delta}\right)^{\frac{1}{\Delta}+1}}_{(b)}\underbrace{(k\Delta+(1+k\Delta)\log(1/1+k\Delta))}_{(c)}.$$

Since $k \geq 0$ and $\Delta > 0$, term (b) is positive and term (a) negative. So, we show that term $(c) \leq 0$:

$$k\Delta+(1+k\Delta)\log(1/1+k\Delta) \leq k\Delta+(1+k\Delta)\left(\frac{1}{1+k\Delta}-1\right) = 0 \quad (15)$$

So, γ_s increases with Δ. Again, we can fill in the smallest possible $\Delta = 1$, so that $\gamma_s \geq \frac{1}{1+k}$. Thus,

$$\gamma_s = \left(\frac{1+k\ell^{(n)}-k}{1+k\ell^{(n)}+k\Delta-k}\right)^{\frac{1}{\Delta}} \geq \left(\frac{1}{1+k\Delta}\right)^{\frac{1}{\Delta}} \geq \frac{1}{1+k} = \gamma_{\text{safe}}. \quad (16)$$

A.4 Proof for Theorem 4: FP_{range} is a decreasing function of γ

We want to show that FP_{range} is decreasing over $\gamma \in [0, 1]$. Consider the derivative:

$$FP'_{\text{range}}(\gamma) = (\ell^{(n)} - \ell^{(N)})\gamma^{\ell^{(n)}-\ell^{(N)}-1} - (\ell^{(n)} - \ell^{(N)})(\gamma + \delta_\gamma)^{\ell^{(n)}-\ell^{(N)}-1} \tag{17}$$

$$= \underbrace{(\ell^{(n)} - \ell^{(N)})}_{\text{(a)}} \underbrace{\left(\frac{1}{\gamma^{\ell^{(N)}-\ell^{(n)}+1}} - \frac{1}{(\gamma + \delta_\gamma)^{\ell^{(N)}-\ell^{(n)}+1}}\right)}_{\text{(b)}}. \tag{18}$$

Part (a) is negative because $\ell^{(N)} > \ell^{(n)}$ by definition. Part (b) is positive because the left side denominator is smaller than right one ($\delta_\gamma > 0$ by definition). So, the derivative $FP'_{\text{range}}(\gamma) < 0$, i.e. the size of the false-positive range *decreases* as γ increases.

References

Raja Farrukh Ali. Non-exponential reward discounting in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 16111–16112, 2023.

Raja Farrukh Ali, John Woods, Esmaeil Seraj, Kevin Duong, Vahid Behzadan, and William Hsu. Hyperbolic discounting in multi-agent reinforcement learning. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024.

Lars Lien Ankile, Brian Ham, Kevin Mao, Eura Shin, Siddharth Swaroop, Finale Doshi-Velez, and Weiwei Pan. Discovering user types: Characterization of user traits by task-specific behaviors in reinforcement learning. In *First Workshop on Theory of Mind in Communicating Agents*, 2023. URL https://openreview.net/forum?id=XO3WwkIDzk.

Anil Aswani, Philip Kaminsky, Yonatan Mintz, Elena Flowers, and Yoshimi Fukuoka. Behavioral modeling in weight loss interventions. *European journal of operational research*, 272(3):1058–1072, 2019.

Roy F Baumeister and Kathleen D Vohs. Self-regulation, ego depletion, and motivation. *Social and personality psychology compass*, 1(1):115–128, 2007.

Frederick Callaway, Yash Raj Jain, Bas van Opheusden, Priyam Das, Gabriela Iwama, Sayan Gul, Paul M. Krueger, Frederic Becker, Thomas L. Griffiths, and Falk Lieder. Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences*, 119(12):e2117432119, 2022. DOI: 10.1073/pnas.2117432119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2117432119.

Owain Evans, Andreas Stuhlmüller, and Noah Goodman. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

Babatunde H Giwa and Chi-Guhn Lee. Estimation of discount factor in a model-based inverse reinforcement learning framework. In *Bridging the Gap Between AI Planning and Reinforcement Learning Workshop at ICAPS*, 2021.

KN Kirby, NM Petry, and WK Bickel. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General*, 128(1):78–87, March 1999. DOI: 10.1037//0096-3445.128.1.78.

W Bradley Knox and Peter Stone. Reinforcement learning from human reward: Discounting in episodic tasks. In *2012 IEEE RO-MAN: The 21st IEEE international symposium on robot and human interactive communication*, pp. 878–885. IEEE, 2012.

Zeb Kurth-Nelson and A David Redish. Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10):e7362, 2009.

Zeb Kurth-Nelson and A. David Redish. A reinforcement learning model of precommitment in decision making. *Frontiers in Behavioral Neuroscience*, 4, December 2010. ISSN 1662-5153. DOI: 10.3389/fnbeh.2010.00184. URL https://www.frontiersin.org/articles/10.3389/fnbeh.2010.00184.

Cassidy Laidlaw and Anca Dragan. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. *arXiv preprint arXiv:2204.10759*, 2022.

Falk Lieder, Frederick Callaway, Yash Raj Jain, Paul M Krueger, Priyam Das, and Sayan Gul. A cognitive tutor for helping people overcome present bias. 2019.

Gregory J Madden, Warren K Bickel, and Eric A Jacobs. Discounting of delayed rewards in opioid-dependent outpatients: exponential or hyperbolic discounting functions? *Experimental and clinical psychopharmacology*, 7(3):284, 1999.

Yonatan Mintz, Anil Aswani, Philip Kaminsky, Elena Flowers, and Yoshimi Fukuoka. Behavioral analytics for myopic agents. *European Journal of Operational Research*, 310(2):793–811, 2023.

Calarina Muslimani, Saba Gul, Matthew E. Taylor, Carrie Demmans Epp, and Christabel Wayllace. Tutor: Helping people learn to avoid present bias during decision making. In Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (eds.), *Artificial Intelligence in Education*, pp. 733–738, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-36272-9.

Joel Myerson and Leonard Green. Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, 64(3):263–276, 1995.

Eura Nofshin, Siddharth Swaroop, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Reinforcement learning interventions on boundedly rational human agents in frictionful tasks. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, pp. 1482–1491, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

Alexander Peysakhovich. Reinforcement learning and inverse reinforcement learning with system 1 and system 2. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 409–415, 2019.

Howard Rachlin, Andres Raineri, and David Cross. Subjective probability and delay. *Journal of the experimental analysis of behavior*, 55(2):233–244, 1991.

Salvador Cruz Rambaud and María José Muñoz Torrecillas. Some considerations on the social discount rate. *Environmental Science & Policy*, 8(4):343–355, 2005.

Sid Reddy, Anca Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. *Advances in Neural Information Processing Systems*, 31, 2018.

Brady Reynolds and Ryan Schiffbauer. Measuring state changes in human delay discounting: an experiential discounting task. *Behavioural processes*, 67(3):343–356, 2004.

Hanneke Scholten, Anouk Scheres, Erik De Water, Uta Graf, Isabela Granic, and Maartje Luijten. Behavioral trainings and manipulations to reduce delay discounting: A systematic review. *Psychonomic bulletin & review*, 26:1803–1849, 2019.

Matthias Schultheis, Constantin A Rothkopf, and Heinz Koeppl. Reinforcement learning with non-exponential discounting. *Advances in neural information processing systems*, 35:3649–3662, 2022.

Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International conference on machine learning*, pp. 5670–5679. PMLR, 2019.

Peter D Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409):2015–2020, 1998.

Peter D Sozou. Individual and social discounting in a viscous population. *Proceedings of the Royal Society B: Biological Sciences*, 276(1669):2955–2962, 2009.

Giles W Story, Ivo Vlaev, Ben Seymour, Ara Darzi, and Raymond J Dolan. Does temporal discounting explain unhealthy behavior? a systematic review and reinforcement learning perspective. *Frontiers in behavioral neuroscience*, 8:76, 2014.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. (arXiv:2407.17032), November 2024. DOI: 10.48550/arXiv.2407.17032. URL http://arxiv.org/abs/2407.17032. arXiv:2407.17032 [cs].

Jiayu Yao, Weiwei Pan, Finale Doshi-Velez, and Barbara E Engelhardt. Inverse reinforcement learning with multiple planning horizons. *Reinforcement Learning Journal*, 3:1138–1167, 2024.

Richard Yi, Heath Milhorn, Anahi Collado, Kate N. Tormohlen, and Jessica Bettis. Uncommitted commitment: Behavioral strategy to prevent preference reversals. *Perspectives on Behavior Science*, 43(1):105–114, Oct 2019. DOI: 10.1007/s40614-019-00229-8.

Guanghui Yu and Chien-Ju Ho. Environment design for biased decision makers. In *IJCAI*, pp. 592–598, 2022.

Guanghui Yu, Robert Kasumba, Chien-Ju Ho, and William Yeoh. On the utility of accounting for human beliefs about ai behavior in human-ai collaboration. *arXiv preprint arXiv:2406.06051*, 2024.

Mo Zhou, Yonatan Mintz, Yoshimi Fukuoka, Ken Goldberg, Elena Flowers, Philip Kaminsky, Alejandro Castillejo, and Anil Aswani. Personalizing mobile fitness apps using reinforcement learning. In *CEUR workshop proceedings*, volume 2068. NIH Public Access, 2018.

Supplementary Materials

The following content was not necessarily subject to peer review.

B Background details

B.1 Linking hyperbolic and exponential Q-values

Here, we elaborate on the result from Fedus et al. (2019) that hyperbolic Q-values can be approximated as an expectation over exponential Q-values:

$$Q_{\text{hyp}}(s,a;k) = \mathbb{E}_{\gamma\sim\text{Beta}(1/k,1)}\left[Q^*_{\text{exp}}(s,a;\gamma)\right]. \tag{19}$$

Starting with Eq. (1), we apply a change of variables $\gamma = \exp(-\lambda)$ which relates the survival probability γ with the hazard λ.

$$\begin{aligned}
&d_{\text{hyp}}(t;k) && && (20)\\
&= \int_{\lambda=0}^{\infty} \frac{1}{k}\exp(-\lambda(t+1/k))d\lambda && \text{From Eq. (1)} && (21)\\
&= \int_{\gamma=1}^{0} \frac{1}{k} \times -\gamma^{-1}\gamma^{t+1/k}d\gamma && && (22)\\
&= \int_{\gamma=0}^{1} \gamma^t \times \frac{1}{k}\gamma^{1/k-1}d\gamma && && (23)\\
&= \mathbb{E}_{\gamma\sim p(\gamma)}\left[\gamma^t\right] && p = \frac{1}{k}\gamma^{1/k-1} && (24)\\
&= \mathbb{E}_{\gamma\sim p(\gamma)}\left[d_{\text{exp}}(t;\gamma)\right]. && && (25)\\
& && && (26)
\end{aligned}$$

Note that the step from Eq. (21) to Eq. (22) follow from the change of variables, where $d\lambda = -\gamma^{-1}d\gamma$ and the respective bounds become $e^0 = 1$ and $e^{-\infty} = 0$.

Finally, the distribution over γ follows a Beta distribution. To see this, we relate $p(\gamma)$ to a uniform distribution by considering the CDF:

$$\begin{aligned}
F_\gamma(x) &= \int_0^x p(\gamma)d\gamma\\
&= \frac{1}{k}\int \gamma^{\frac{1}{k}-1}d\gamma\\
&= \frac{1}{k}\left(k\gamma^{\frac{1}{k}}\right)\Big|_{\gamma=0}^{x}\\
&= x^{\frac{1}{k}}.
\end{aligned}$$

This implies that $\gamma = U^k$, where $U \sim \text{Unif}(0,1)$. Equivalently, γ follows a beta distribution $\text{Beta}(1/k, 1)$.

Since Q-values are discounted sums of rewards, the above relationship holds for Q-values due to the linearity of expectations:

$$
\begin{aligned}
&Q_{\text{hyp}}(s, a; k) \\
&= \sum_{t=0}^{\infty} d_{\text{hyp}}(t; k) R_t \\
&= \sum_{t=0}^{\infty} \mathbb{E}_{\gamma} \left[d_{\text{exp}}(t; \gamma) R_t \right] \\
&= \mathbb{E}_{\gamma} \left[\sum_{t=0}^{\infty} d_{\text{exp}}(t; \gamma) R_t \right] \\
&= \mathbb{E}_{\gamma} \left[Q_{\text{exp}}(s, a; \gamma) \right].
\end{aligned}
$$

B.2 Behavior Model RL (BMRL)

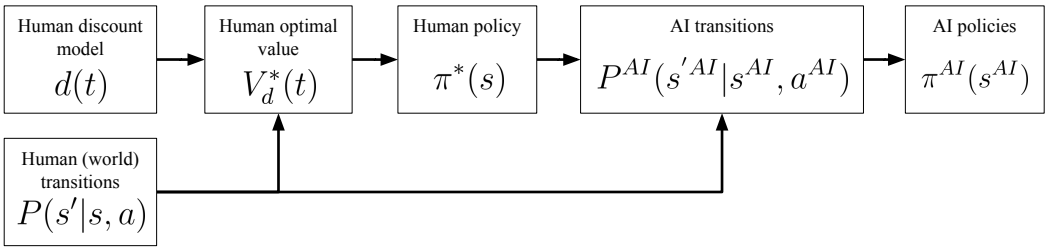

Figure 7: Overview of how the human's discount function affects the downstream AI policy in BMRL in Nofshin et al. (2024). The human agent's discount model and transitions affect the human agent's optimal value function, which in turns affects the human's optimal policy. The human's policy is completely encapsulated in the AI transitions, which in turn, affects the AI policy. Note that the human transitions appear twice; first to affect the optimal value, and then to affect the AI transitions.

B.3 Worlds Represented by Our Sampled Absorbing State MDPs

Cliff walking world The cliff walking world is a 2-D gridworld introduced in Sutton & Barto (2018) and appears throughout the literature, including by implementation in the popular Gymnasium library introduced by Towers et al. (2024). There is a start state, a goal state, and a set of "cliff" states that run along the bottom of the world. If the agent enters a cliff state, they transition back to the start state.

The goal state is well represented as an absorbing state. If the cliff is implemented as an absorbing state, then it corresponds to a distractor state, and the entire cliff world is an absorbing state MDP. If the cliff is implemented as a non-absorbing state (i.e., the agent gets sent back to the starting state if they enter a cliff state), then this is still an absorbing state MDP without any distractor states.

Chain world Nofshin et al. (2024) introduced the chain world, which captures a notion of a (human) user's progress toward some task. There is a disengagement state where once the user disengages, the user receives reward of 0 in perpetuity. There is also a goal state, and there are intermediary progress states. The goal state corresponds to the goal state in absorbing state MDPs. The disengagement states correspond to distractor states in absorbing state MDPs. Hence chain worlds are absorbing states.

Vegetarian cafe vs. donut chain world. Evans et al. (2016) introduces a world where agents face a tradeoff from going to nearby donut chain stores versus a further vegetarian cafe that is better for

their health. There is also a second path with a noodle shop. Both the donut chain stores and the noodle shop represent distractor absorbing states, while the further vegetarian cafe represents a goal absorbing state. Hence, this can be represented as an absorbing state MDP.

Path world Fedus et al. (2019) introduces a world of paths of varying lengths, where the agent faces a decision between the paths. We can represent the lengths of the paths as intermediate states, and the states at the end of each path are indeed absorbing states. These absorbing states vary in reward, and the largest is the goal absorbing state; the others are distractors. Hence, this world is well represented by our sampled absorbing state MDPs.

Precommitment and addiction Kurth-Nelson & Redish (2010) links hyperbolic discounting to notions of precommitment — which occurs when an agent takes a path that goes toward a single reward and excludes the possibility of the type of preference reversal seen in hyperbolic discounting — and addiction science and other manifestations of impulsivity in behavioral science.

The example given by Kurth-Nelson & Redish (2010) where an agent is able to commit to a larger world (their "Figure 1") is indeed well represented by an absorbing state MDP of the type we sample. There are two large rewards — which can be represented as goal absorbing states — and one small reward — which can be represented as a distractor absorbing state.

This is a particularly salient example because of the links to real-world applications in modeling behavior including in the study of addiction.

C Theoretical Analysis

C.1 Form of Value Functions for Absorbing State MDPs

Let L be the time to any absorbing state under policy π. Let R be the reward at that absorbing state. Both of these variables are *random* because of the randomness in the transitions P. In absorbing state MDPs, value functions will have the form:

$$
\begin{aligned}
& V^{\pi}(s) && (27)\\
&= \mathbb{E}_{\pi,P}\left[\sum_{t=1}^{\infty} d(t-1)R_t\right] && \text{Definition of value function}\\
&= \mathbb{E}_{L,R}\left[d(L-1)R + r_b\sum_{t=1}^{L} d(t-2)\right] && \text{MDP structure}\\
&= \mathbb{E}_{L,R}\left[d(L-1)R + r_b\sum_{t=0}^{L-2} d(t)\right] && \text{Shitfing sum}\\
&= \mathbb{E}_{L}\left[\mathbb{E}_{R}\left[d(L-1)R + r_b\sum_{t=0}^{L-2} d(t)\right]\right] && \text{Repeated expectations}\\
&= \mathbb{E}_{L}\left[d(L-1)\mathbb{E}_{R}\left[R\right] + r_b\sum_{t=0}^{L-2} d(t)\right]. && (28)
\end{aligned}
$$

If we assume the *transitions are deterministic*, then L and R are no longer random. Let $\ell^{(n)}$ be the length of the path from state s to absorbing state $s^{(n)}$ with reward $r^{(n)}$. Furthermore, if we assume *no burden* ($r_b = 0$), then only the absorbing state reward remains. The value of a policy $\pi^{(n)}$ that goes to absorbings state $s^{(n)}$ is:

$$
V_d^{(n)}(s) = d(\ell^{(n)} - 1)r^{(n)}. \tag{29}
$$

C.2 Proof that $\gamma_{\text{mhr}} < \gamma_{\text{safe}}$

$$\gamma_{\text{mhr}} < \gamma_{\text{safe}} \tag{30}$$
$$\implies \exp(-k) < \frac{1}{1+k} \tag{31}$$
$$\implies -k < \ln\left(\frac{1}{1+k}\right) \tag{32}$$
$$\implies k \geq \ln\left(\frac{1}{1+k}\right) \tag{33}$$
$$\implies k \geq \frac{1}{1+k} - 1, \qquad \text{(Note that } \ln(x) \leq x - 1\text{)} \tag{34}$$
$$\implies (k+1)^2 \geq 1 \tag{35}$$
$$\implies k^2 + 2k + 1 \geq 1 \tag{36}$$
$$\implies k^2 + 2k \geq 0. \tag{37}$$

The last line is always true, since $k > 0$.

C.3 Expanded details on false positive range

In AI interventions, false positives are when the AI intervenes despite the fact that the human agent would have reached the desired goal state without needing intervention. A scenario with a false positive requires three conditions to hold.

1. The (true) hyperbolic agent is already going to the big reward:

$$V_{\text{hyp}}^{(N)}(s;k) \geq V_{\text{hyp}}^{(n)}(s;k) \quad \text{for all } n \in \{1,\ldots,N\}$$

2. The exponential agent goes to the small reward:

$$V_{\text{exp}}^{(n)}(s;\gamma) \geq V_{\text{exp}}^{(N)}(s;\gamma) \quad \text{for any } n \in \{1,\ldots,N\}$$

3. The exponential agent under intervention goes to the big reward.

$$V_{\text{exp}}^{(N)}(s;\gamma+\delta_\gamma) \geq V_{\text{exp}}^{(n)}(s;\gamma+\delta_\gamma) \quad \text{for all } n \in \{1,\ldots,N\},$$

where $\delta_\gamma > 0$ refers to the increase in the exponential agent's discount factor.

Formalizing condition (1). Let i be the best option absorbing state (that is not the goal state), defined as:

$$i = \text{argmax}_{i=1,\ldots,N-1} V_{\text{hyp}}^{(i)}(s).$$

If the agent prefers the goal state, it means that the goal state is better than this alternate best-option:

$$V_{\text{hyp}}^{(N)}(s) \geq V_{\text{hyp}}^{(n)}(s) \quad \text{for all } n \in \{1,\ldots,N\} \tag{38}$$
$$\implies V_{\text{hyp}}^{(N)}(s) \geq V_{\text{hyp}}^{(i)}(s) \tag{39}$$
$$\implies \frac{1}{1+k\ell^{(N)}-k} r^{(N)} \geq \frac{1}{1+k\ell^{(i)}-k} r^{(i)} \tag{40}$$
$$\implies (1+k\ell^{(i)}-k) r^{(N)} \geq (1+k\ell^{(N)}-k) r^{(i)} \tag{41}$$
$$\implies r^{(N)} \geq \frac{1+k\ell^{(N)}-k}{1+k\ell^{(i)}-k} r^{(i)}. \tag{42}$$

Formalizing condition (2). Let j be the best option absorbing state (that is not the goal state) under exponential discounting, defined as

$$j = \text{argmax}_{j=1,\dots,N-1} V_{\text{exp}}^{(j)}(s).$$

If the agent prefers the distractor state, it means the best-option absorbing state is better than the goal state:

$$V_{\text{exp}}^{(n)}(s) \geq V_{\text{exp}}^{(N)}(s) \quad \text{for any } n \in \{1,\dots,N\} \tag{43}$$
$$\implies V_{\text{exp}}^{(j)}(s) \geq V_{\text{exp}}^{(N)}(s) \tag{44}$$
$$\implies \gamma^{\ell^{(j)}-1} r^{(j)} \geq \gamma^{\ell^{(N)}-1} r^{(N)} \tag{45}$$
$$\implies r^{(N)} \leq \gamma^{\ell^{(j)}-\ell^{(N)}} r^{(j)} \tag{46}$$

Formalizing condition (3). The condition becomes:

$$V_{\text{exp}}^{(N)}(s;\gamma+\delta_\gamma) \geq V_{\text{exp}}^{(n)}(s;\gamma+\delta_\gamma) \quad \text{for all } n \in \{1,\dots,N\} \tag{47}$$
$$\implies V_{\text{exp}}^{(N)}(s;\gamma+\delta_\gamma) \geq V_{\text{exp}}^{(j)}(s;\gamma+\delta_\gamma) \tag{48}$$
$$\implies (\gamma+\delta)^{\ell(N)} r^{(N)} \geq (\gamma+\delta)^{\ell(j)} r^{(j)} \tag{49}$$
$$\implies r^{(N)} \geq (\gamma+\delta)^{\ell^{(j)}-\ell^{(N)}} r^{(j)}. \tag{50}$$

Defining the false-positive range for γ. Since our choice of γ does not affect whether or not the hyperbolic agent prefers the goal state, we can ignore condition (1).

So, our AI will send a false positive if:

$$(\gamma+\delta)^{\ell^{(j)}-\ell^{(N)}} r^{(j)} \leq r^{(N)} \leq \gamma^{\ell^{(j)}-\ell^{(N)}} r^{(j)} \tag{51}$$
$$\implies (\gamma+\delta)^{\ell^{(j)}-\ell^{(N)}} \leq {}^{r^{(N)}}\!/_{r^{(j)}} \leq \gamma^{\ell^{(j)}-\ell^{(N)}} \tag{52}$$

This defines the range of values for γ under which a false positive might occur. We want to show that larger γ results in a smaller chance of false positives. This means that we want this range to be smaller the larger the γ.

D Empirical Experiment Details

D.1 Definition of the AI MDP

- **AI states.** The AI state $s^{AI} = [s, \alpha]$ is derived from the human agent's MDP. It includes the human's current state s and the human's action at the last timestep, α. At the beginning of an episode (when there is no concept of previous timestep for the human's action), α is initialized to 0.
- **AI actions.** The AI actions are binary; the AI either intervenes on the human agent's discount $a^{AI} = 1$ or does not intervene $a^{AI} = 0$.
- **AI rewards.** The AI receives a small negative reward for intervening ($a^{AI} = 1$), a large negative when the human agent enters a distractor state ($s = s^{(1)}, \dots, s^{(N-1)}$), and a large positive reward when the human agent enters the goal state ($s^{(N)}$).

$$R^{AI}(s^{AI}, a^{AI}) = \begin{cases} -0.2 & \text{if } a^{AI} = 1 \\ -1 & \text{if } s = s^{(1)}, \dots, s^{(N-1)} \\ 1 & \text{if } s = s^{(N)} \end{cases} \tag{53}$$

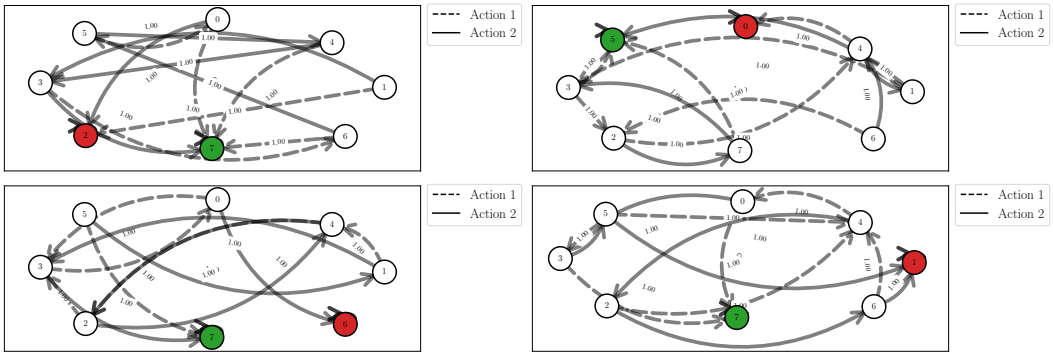

Figure 8: Examples of state diagrams for randomly sampled absorbing state MDPs.

- **AI transitions.** The AI transitions are determined by the gridworld in which the human agent operates, since they factorize into two distributions: $P^{AI}(s'^{AI}|s^{AI}, a^{AI}) = P(s'|s, \alpha')\pi(\alpha'|s, a^{AI})$, where $\pi(\alpha'|s, a^{AI})$ is the effect of the AI intervention on the human's action, and $P(s'|s, \alpha')$ is the effect of the human action on the next state.

  For gridworlds with stohastic transitions, we first sample a deterministic gridworld. We then add stochasticity of level ϵ. The original transition has probability $1 - \epsilon$. The probability of transitioning to the remaining N connections from state s is then ϵ/N.
- **AI discount.** We use an exponential discount function with $\gamma = 0.99$.

D.2 List of estimators

We are using the following integral estimation methods, which we refer to above as:

- **mcmc**: Monte Carlo estimation sampling from a target distribution and averaging.
- **quad**: Gaussian quadrature that approximates via selection of nodes and weights.
- **riemann**: A simple Riemann sum.
- **strat**: Stratified sampling done by sampling uniformly among the strata (divisions of the sampled support).
- **importance**: Importance sampling drawing from a proposal distribution and shifting to a target distribution.

D.3 Examples of randomly sampled absorbing state MDPs

Fig. 8 shows examples of state diagrams for randomly sampled absorbing state MDPs.

D.4 Experimental results in expanded settings

D.4.1 Stochastic transitions

In Fig. 9, noise $\eta \in [0, 1]$ represents the stochasticity of environment transitions. Formally, there is a $1 - \eta$ change of transitioning to state s' after taking action a in state s, and there is a η chance of transitioning to a random state that is not s'. The larger η, the more stochastic.

D.4.2 Optimal (deterministic) human policies

In Fig. 10, we show the impact of running a simulation in which the human agent follows an optimal, deterministic policy vs. a softmax policy. As expected, the softmax policy leads to noisier results.

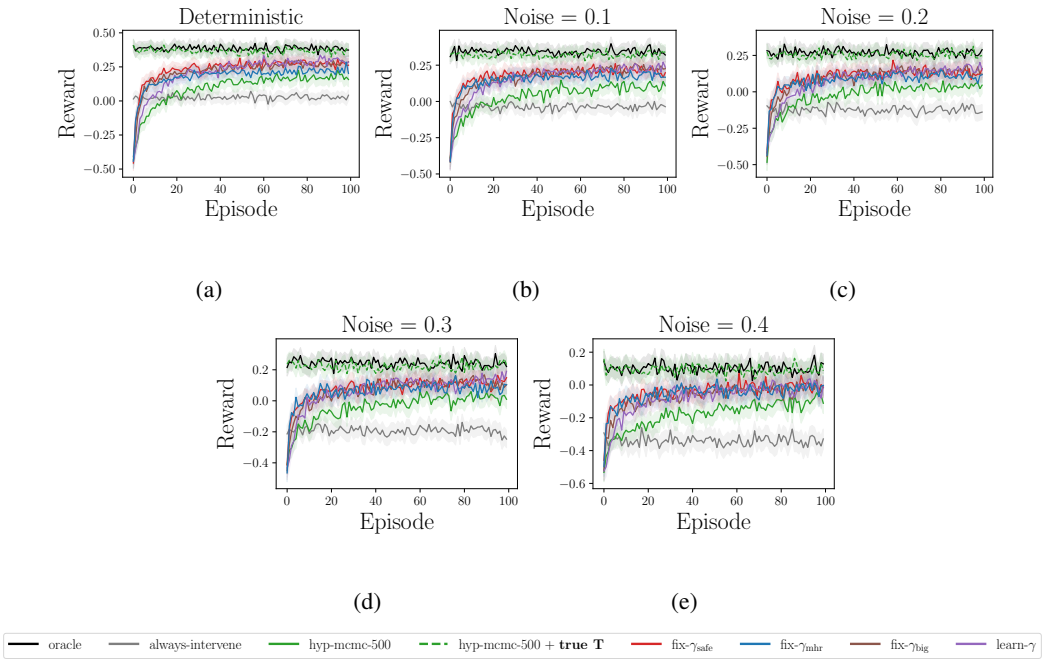

Figure 9: Cumulative reward of AI policy in sampled absorbing state MDPs with varying levels of environment stochasticity. The stochasticity does not affect the main trends; the exponential methods still outperform hyperbolic, and all policies outperform the naive always-intervene baseline.

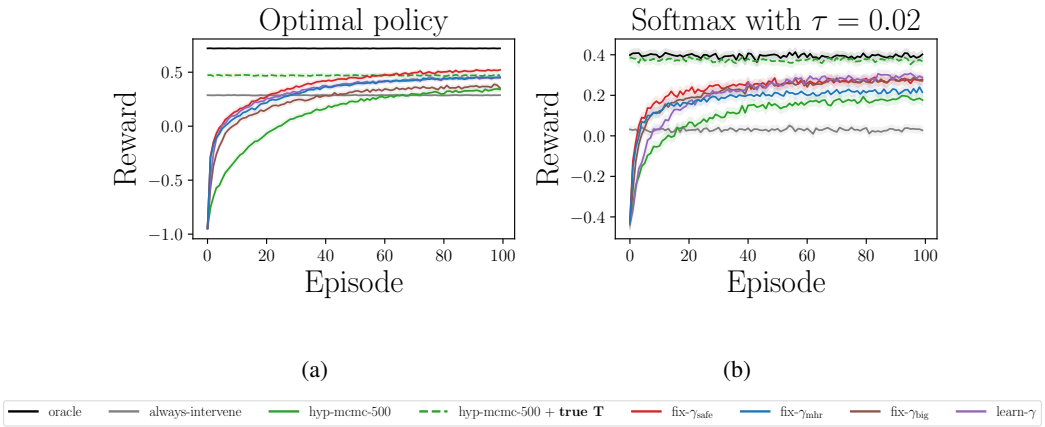

Figure 10: Cumulative reward of AI policy in sampled absorbing state MDPs with different action selection policies for the human agent. Most main trends remain the same, but the hyperbolic baseline with the true transitions does worse when the human is optimal (green, dotted line), due to small errors in the Q-values translating to errors in ranking actions.