Reinforcement Learning Interventions on Boundedly Rational Human Agents in Frictionful Tasks

Eura Nofshin
Harvard University
Cambridge, USA
eurashin@g.harvard.edu

Siddharth Swaroop
Harvard University
Cambridge, USA
siddharth@seas.harvard.edu

Weiwei Pan
Harvard University
Cambridge, USA
weiweipan@g.harvard.edu

Susan Murphy
Harvard University
Cambridge, USA
samurphy@fas.harvard.edu

Finale Doshi-Velez
Harvard University
Cambridge, USA
finale@seas.harvard.edu

ABSTRACT

Many important behavior changes are *frictionful*; they require individuals to expend effort over a long period with little immediate gratification. Here, an artificial intelligence (AI) agent can provide personalized interventions to help individuals stick to their goals. In these settings, the AI agent must personalize *rapidly* (before the individual disengages) and *interpretably*, to help us understand the behavioral interventions. In this paper, we introduce Behavior Model Reinforcement Learning (BMRL), a framework in which an AI agent intervenes on the parameters of a Markov Decision Process (MDP) belonging to a *boundedly rational human agent*. Our formulation of the human decision-maker as a planning agent allows us to attribute undesirable human policies (ones that do not lead to the goal) to their maladapted MDP parameters, such as an extremely low discount factor. Furthermore, we propose a class of tractable human models that captures fundamental behaviors in frictionful tasks. Introducing a notion of *MDP equivalence* specific to BMRL, we theoretically and empirically show that AI planning with our human models can lead to helpful policies on a wide range of more complex, ground-truth humans.

KEYWORDS

Reinforcement learning; Personalization; Agent-based modeling of humans; Bounded rationality

ACM Reference Format:
Eura Nofshin, Siddharth Swaroop, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. 2024. Reinforcement Learning Interventions on Boundedly Rational Human Agents in Frictionful Tasks. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 10 pages.

1 INTRODUCTION

In many AI+human applications of behavior change, AI agents assist the human in performing *frictionful* tasks, where making progress toward the human's goal requires sustained effort over time with little immediate gratification. Examples include physical therapy (PT) programs, adherence to scheduled medication, or passing an online course. Two key challenges for AI agents in these settings are rapid personalization [26, 35, 43] and learning interpretable policies for intervention [41, 44]. In frictionful tasks, since effort exerted by the human does not reap immediate benefits, the AI agent must learn a personalized intervention policy for each human in a small number of interactions, or risk disengagement. These policies must also be interpretable to experts in behavioral science so that they can discover which interventions work for which individuals, and investigate why.

Grounded in behavioral literature that treats humans as sequential decision-makers (e.g. [24, 33, 37, 38, 50]), we model the human as an agent planning under a "maladapted" Markov Decision Process (MDP). In maladapted human MDPs, the optimal policy does not reach the human's stated goal; for example, in physical therapy (PT), the goal may be a rehabilitated shoulder and the maladapted MDP parameter may be an extremely low discount rate, γ. This results in myopic decision-making, wherein an individual forgoes the long-term goal (rehabilitated shoulder) to avoid experiencing friction in the short-term (unpleasantness of PT). The AI agent helps the individual achieve their long-term goal by changing the maladapted human MDP (and thereby the optimal policy).

While there is existing reinforcement learning (RL) literature for optimizing interventions on human utility functions (i.e. reward) in maladapted MDPs [19, 46, 50], interventions on γ have not been optimized from an RL perspective. On the other hand, in behavioral science, humans have been observed to use a problematically low γ [34] and scientists have developed interventions to change a human's γ (e.g. [17]). However, no work optimizes for *when* and with what mechanisms to intervene on the parameters of the human's maladapted MDP.

In this paper, we introduce a *flexible* and *behaviorally interpretable* framework called "Behavior-Model RL" (BMRL). In BMRL, the human is modeled as an RL agent, whose actions are *behaviors*, such as performing or skipping PT; the AI agent provides personalized assistance by delivering *interventions* on the human's maladapted MDP parameters. By linking the behaviors of our human agents to their MDP parameters, BMRL allows us to *interpret* the mechanism behind the human's maladapted decision-making. Our framework is also more *flexible* than existing ones since we allow the AI agent's actions to include operations on any part of

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

the human MDP (such as γ). By solving for the AI agent's optimal policy, we learn the best set of interventions to change the human agent's behavior and to help the human reach their goal.

Unfortunately, current RL approaches have two major drawbacks when used to solve for the optimal AI agent policy in BMRL. First, most planning methods are too data-intensive for our setting, in which personalization occurs online. For example, online algorithms in robotics require thousands of interactions to learn reasonable policies (e.g. in [39, 40, 45]), but in frictionful tasks, we are limited to *tens to hundreds* of interactions [41]. Second, existing planning methods model the human as a black-box transition or value function. Unfortunately, in learning black-box representations of the human agent, we lose the ability to interpretably attribute human behavior to their MDP parameters.

In this paper, we propose a tractable planning method for the AI agent in our BMRL framework. Our method provides the AI agent with a useful inductive bias, in the form of a human model that captures important behavioral patterns in frictionful tasks. Specifically, we identify a small, behaviorally grounded model of the human that the AI agent can leverage to rapidly personalize interventions, including previously under-explored interventions on γ. Then, we introduce the concept of "AI equivalence" to identify a class of more complex human models for which AI policies learned in our simple human model can be lifted with no loss of performance. In our empirical analysis, we test whether AI planning with our small model is robust to complex human models that are not covered by our equivalence result. Throughout all of this, our small model preserves scientific interpretability– in fact, it has an analytical solution for the human behavior policy– which allows experts to inspect and learn from the AI policies.

2 RELATED WORKS

Computational modeling of human behaviors. Behavioral scientists have developed and verified several computational models of *dynamic* human decision-making. Unlike *static* models, such as Social Cognitive Theory [2], *dynamic* models of decision-making apply to interactive human-AI settings, since they capture person-level variation and changes over time, as in Zhang et al. [47]. Scientists developed these models to explain *offline data* from frictionful settings such as health (e.g. [18, 43, 48]), energy [20], and experience sampling [14] or to capture broader behaviors such as risk [16] and adherence [27]. However, these models involve too many latent variables– corresponding to internal human processes– to facilitate rapid AI learning from *online data*. In contrast, we propose a minimal, behaviorally grounded model, one whose set of latent parameters is small and structured enough that our AI can learn.

Computational modeling of human agent deficiencies. RL is frequently used to model the complex mechanisms underlying human behavior, from the firing of dopaminergic neurons in the brain (e.g. in [24, 33]) to frictionful tasks such as mindful eating [37], weight loss [1], and smoking cessation [38]. Although these works use RL to model humans, the models themselves are not used to enrich planning for an AI agent. One exception is inverse reinforcement learning, in which the AI agent infers the human agent's rewards (e.g. [3, 49]), transitions [30], discount factor [11], or entire MDP [7, 13, 32], but does not *intervene* on the parts of the human MDP

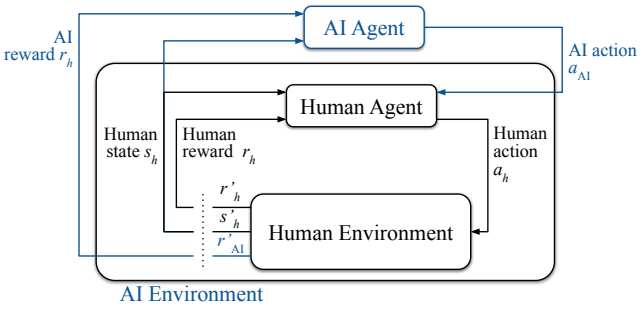


Figure 1: Overview of BMRL. The human agent interacts with the environment as in standard RL. The AI agent's actions affect the human agent. The human agent + environment form the AI environment.

that are maladapted. When the AI agent does intervene, the changes are limited to the human's reward [19, 36, 46, 50] or states [6, 31]. Our BMRL framework is flexible enough to incorporate AI interventions on multiple parts of the human MDP, including the discount factor or transitions.

Equivalence of (human) MDP models. In RL, there are notions of equivalence that can reduce larger human MDPs to smaller, more manageable ones. Equivalence, as defined in bisimulations [10], homomorphism [28], and approximate homomorphisms (e.g. [29, 42]), requires that one human MDP strictly preserves the transition and reward functions of another, given a mapping between the state and action spaces. State abstraction methods, which equate the optimal value function between two human-level MDPs, are less strict [15]. However, these equivalences are still *stricter than necessary* in our setting, where we only care that the human MDPs are similar enough that the AI agent policy will not differ. Furthermore, the simpler MDPs recovered by these methods are not guaranteed to be behaviorally valid or interpretable. In our approach, we define two human MDPs as equivalent if they lead to the same *AI optimal policies*, and we use this definition to build up to more complex human MDPs from a behaviorally interpretable one.

3 THE BEHAVIOR MODEL RL (BMRL) FRAMEWORK FOR AI INTERVENTIONS

We define a formal framework, called BMRL, in which an AI agent learns to intervene on a human agent's maladapted MDP parameters (overview in fig. 1).

3.1 Assumptions on human agent

In BMRL, human agents perform optimal planning on (subconscious) knowledge of their MDP,

$$\mathcal{M}_h = \langle \mathcal{S}_h, \mathcal{A}_h, T_h, R_h, \gamma_h, s_g, s_d \rangle, \tag{1}$$

where $s_g, s_d \in \mathcal{S}_h$ are absorbing goal (e.g. a rehabilitated shoulder) and disengagement states (e.g. quitting PT).

Though in general, it is possible for the human's perception of the states $\mathcal{S}_h$, actions $\mathcal{A}_h$ and transitions T_h to be maladapted, in this paper we assume that the human's perception matches the true environment. On the other hand, we allow the human's rewards

R_h and discount γ_h to vary by perception. For example, one may skip PT because of a tendency to ignore long-term rewards (low γ_h) while another may skip PT because they find the workout to be extremely unpleasant (bad R_h).

We assume that at any point the human subconsciously "knows" their own MDP, solves for the optimal policy, and uses it to select actions. In future work, BMRL can extend to sub-optimal human planning. Despite being optimal, our human agents are still boundedly rational because their MDP is maladapted. That is, under certain values of γ_h, R_h, even an optimal human policy will *never* lead to the goal state (e.g. if the path to the goal reward is laced with extremely negative rewards). The existence of maladapted MDPs in humans is shown in behavior science, where myopic discounting has been linked to excessive alcohol intake [34] or miscalibrated rewards have been linked to unhealthy eating [37]. Despite subconscious knowledge of their own MDP, our human agents are still boundedly rational because (1) they may not be *conscious* of their deficiencies and unable to target them; (2) even if aware, they may still struggle to change their deficiencies. In both cases, behavioral interventions (delivered by the AI agent) can help.

3.2 AI agent

Our AI agent encourages the human agent toward the goal by intervening on the human's decision-making parameters, such as γ_h. To do so, the AI agent plans according to an MDP,

$$\mathcal{M}_{AI} = \langle \mathcal{S}_{AI}, \mathcal{A}_{AI}, T_{AI}, R_{AI}, \gamma_{AI} \rangle, \tag{2}$$

with known rewards R_{AI} and unknown transitions T_{AI}.

Upon observing state $s_{AI} = [s_h, a_h]$, which consists of the human's current state and *previous* action, the AI agent must decide whether to intervene on the human's discount ($a_{AI} = a_\gamma$), reward ($a_{AI} = a_R$), or to do nothing ($a_{AI} = 0$). In practice, a discounting intervention a_γ could be "episodic future thinking," where individuals imagine future events as if they are presently occurring [4]; this could executed as a guided activity in-app. A common intervention on reward a_R is to offer extrinsic rewards, such as badges [8]. Domain experts would determine how the interventions are executed, e.g. if the burden intervention should be a badge, motivational message, or cash.

To encourage policies that quickly lead to the goal state, the AI agent receives a positive reward when the human reaches the goal state, a negative reward when the human disengages, and a negative reward for the "cost" of intervening. The AI's transitions factorize into two probability distributions, $T_{AI}(s_{AI}, a_{AI}, s'_{AI}) = P(s'_h|s_h, a_h)P(a'_h|s_h, a_{AI}) = T_h(s_h, a_h, s'_h)\pi_h(a'_h|s_h, a_{AI})$. The first distribution refers to the human-level transitions T_h. The second distribution is over human actions; it is the human policy that results from the AI's intervention on the human's MDP. Importantly, we assume that the effect of AI actions on the human MDP is *temporary*. For example, if the AI agent increases the human's discount factor γ_h to γ'_h in the current time step, the human's discounting will have reverted to γ_h at the next time step.

In table 1, we provide a comparison on what the AI and human agents separately know and observe. Note that all of the AI agent's unknown parameters pertain to the human MDP $\mathcal{M}_h$ and are contained in the AI's transitions T_{AI}. Instead of explicitly learning $\mathcal{M}_h$ to form T_{AI}, we could directly estimate T_{AI} or Q^*_{AI} using standard model-based or model-free techniques. However, by learning $\mathcal{M}_h$, we take advantage of the known structure of the problem; the better the AI's model of $\mathcal{M}_h$, the better the inductive bias for forming T_{AI} (and therefore π^*_{AI}).

	Human agent	AI agent
Knows...	$\mathcal{S}_h, \mathcal{A}_h, T_h, R_h, \gamma_h$	$\mathcal{S}_{AI}, \mathcal{A}_{AI}, R_{AI}$
Does not know...	–	T_{AI} (includes T_h, R_h, γ_h)
Observes...	$\mathcal{S}_h, \mathcal{A}_h, \mathcal{A}_{AI}$	$\mathcal{S}_h, \mathcal{A}_h, \mathcal{A}_{AI}$

Table 1: Overview of what is known, unknown, and observable to the human and AI agent. the AI agent does not know (and must infer) the human agent's MDP (R_h, γ_h) and the true environmental transitions (T_h).

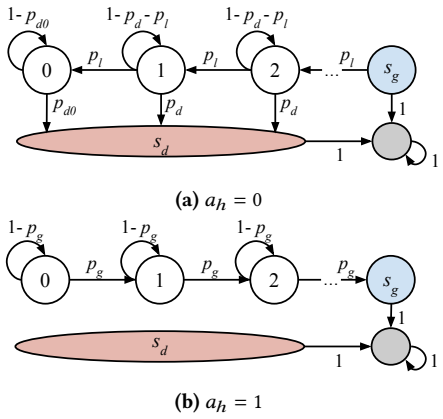


Figure 2: Graphical representation of the chainworld.

4 RAPID PERSONALIZATION IN BMRL VIA A SIMPLE HUMAN MODEL

4.1 Chainworlds: a simple human model that captures progress-based decision-making

In this section, we define *chainworlds*, a class of simple human MDPs that the AI agent will use as a stand-in model for the *true* human decision-making process. Chainworlds are based on the observation that many frictionful tasks contain a notion of human progress toward a goal; for example, in PT, the progress toward a rehabilitated shoulder may be summarized by the current strength of the joint. We summarize these "progress-based" settings with a "progress-only" class of human MDPs, shown in fig. 2, which we call *chainworlds* and denote $\mathscr{M}_{\text{chain}}$.

Each element of $\mathscr{M}_{\text{chain}}$ is as follows:

- *States* $s_h \in \{s_0, s_1, \ldots, s_N = s_g, s_d\}$. The N states are 1-D, discrete, and represent progress toward the goal. The goal state at the end of the chain, $s_N = s_g$ means that the human has rehabilitated their shoulder. The disengagement state s_d means that the human has disengaged from PT.
- *Actions* $a_h \in \{0, 1\}$. The human decides to perform ($a_h = 1$) or not perform ($a_h = 0$) the goal-directed behavior. In the future, this could be extended to categorical actions. That said, many important applications have binary actions, such as "exercise or

not" in PT, "smoke or not" in smoking cessation, and "adhered or not" in medication adherence.

- *Rewards.* The human's utility function is the reward,

$$R_h(s, a, s') = \begin{cases} r_b, & a = 1 \\ r_\ell, & s' < s \\ r_g, & s = s_g \\ r_d, & s = s_d. \end{cases} \tag{3}$$

  Goal behaviors, such as doing PT, incur a cost representing burden $r_b < 0$. Similarly, losing progress incurs $r_\ell < 0$. The goal and disengagement states have positive utility, $r_g > 0$ and $r_d > 0$.
- *Transitions.* The human knows that there is p_g probability that they will move toward the goal as a result of the behavior, p_ℓ probability that they will lose progress from abstaining, and p_d probability that they will disengage from abstaining. These probabilities are fixed across states, except for the first state s_0, which has a separate probability of disengagement $p_{d0} \geq p_d$.
- *Discount.* The human exponentially discounts future rewards via $\gamma_h \in [0, 1)$. We leave other behaviorally relevant forms of discounting, such as hyperbolic discounting [9], as future work.
- *Effect of AI interventions.* When $a_{AI} = a_\gamma$ the human's discount γ_h *increases* by $\Delta_\gamma > 0$, and when $a_{AI} = a_b$ the human's burden $r_b < 0$ *decreases* by Δ_b. We clip $\gamma_h + \Delta_\gamma$ to be between 0 and 1.

Each individual is an instance of the chainworld, $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$, with parameters $\theta = \{r_b, r_\ell, r_g, r_d, p_g, p_\ell, p_d, p_{d0}, \gamma_h, \Delta_\gamma, \Delta_b\}$. For example, some people tend to prioritize short-term rewards (with a low γ_h) while others prioritize long-term rewards (with a high γ_h). The parameters θ must be inferred by the AI.

Closed-form Solutions for Human Policies in Chainworlds. Chainworlds are inspectable to behavioral experts because there is an analytical solution for the optimal value function (all derivations in appendix A [25]). For a chainworld MDP $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$, the optimal value function maximizes between the value of a policy that always pursues the goal, $\pi_g(s_n) = 1$, and the value of a policy that always chooses to disengage, $\pi_d(s_n) = 0$, where s_n for $n \in 0, \ldots, N$ refers to the n-th state on the chain. The value of goal pursuit is,

$$V_\theta^{\pi_g}(s_n) = r_g \left(\frac{\gamma p_g}{z}\right)^{N-n} + r_b \left(\frac{1 - (\gamma p_g / z)^{N-n}}{1 - \gamma}\right), \tag{4}$$

where $z = 1 - \gamma(1 - p_g)$. The value of goal pursuit, $V_\theta^{\pi_g}(s_n)$, trades off between the long-term utility of the goal (the r_g term) and the burden one accumulates to get there (the r_b term). The value of disengagement is,

$$V_\theta^{\pi_d}(s_n) = r_d \left(\frac{\gamma\, p_{d0}}{v}\right) \left(\frac{p_\ell\, \gamma}{u}\right)^n + (\gamma\, p_d\, r_d + p_\ell\, r_\ell) \left(\frac{1 - (\gamma p_\ell / u)^n}{1 - \gamma(1 - p_d)}\right), \tag{5}$$

where $v = 1 - \gamma(1 - p_{d0})$ and $u = 1 - \gamma(1 - p_d - p_\ell)$. The first term in the equation (with r_d), represents the value of disengagement from state 0, after having lost all prior progress. The second term represents the value of disengagement after state 0, which factors in the cost of disengagement r_d and of losing progress r_ℓ.

These equations allow us to hypothesize about the diverse space of AI actions that will encourage the human towards the goal, such as actions to increase the human's level of motivation (increasing r_g) or that highlight the consequences of quitting (decreasing r_d).

4.2 Different humans yield different AI policies

At this point, we have fully specified an AI MDP as defined in section 3.2, in which the human MDP is a chainworld $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$. Solving this AI MDP will yield an optimal AI policy, which is the best intervention plan for a given human with parameters θ. Importantly fig. 3 demonstrates that personalization is necessary because humans with different θ require different optimal AI policies.

s_d	s_0	s_1	s_2	s_3	s_4	s_g
0	0	0	0	0	a_γ	0

(a) Highly myopic human ($\gamma = 0.1$) with high burden ($r_b = -2$).

s_d	s_0	s_1	s_2	s_3	s_4	s_g
0	a_b	a_b	a_γ/a_b	a_γ/a_b	0	0

(b) Highly myopic human ($\gamma = 0.1$) with low burden ($r_b = -0.3$).

Figure 3: Example of different optimal AI policies for two humans with different chainworld parameters. Each square is a chainworld state. An a_b means AI should select action to reduce r_b, while a_γ means AI should select action to increase γ. Red solid and blue dotted lines show start and end of intervention window.

5 THEORETICAL ANALYSIS: WHEN IS CHAINWORLD GOOD ENOUGH?

In this section, we define an *equivalence class* of more complex human MDPs for which an AI agent that plans with the chainworld can still learn the optimal policy.

Definition 5.1 (AI equivalence of human MDPs). We consider two human MDPs $\mathcal{M}_h \equiv \widehat{\mathcal{M}}_h$ under state mapping $f : \mathcal{S}_h \to \widehat{\mathcal{S}}_h$ and action mapping $g_s : \mathcal{A}_h \to \widehat{\mathcal{A}}_h$ if the corresponding optimal AI policies are equal, so that $\pi^*_{AI}\left([s_h, a_h]\right) = \widehat{\pi}^*_{AI}\left([f(s_h), g_{s_h}(a_h)]\right)$ for all $[s_h, a_h] \in \mathcal{S}_{AI}$.

The state mapping f and (state-specific) action mapping g_s let us map from the state and action space of one MDP to the other. In terms of the chainworld, our definition states that if the optimal AI action in the chainworld MDP is the same as the optimal AI action in the true MDP for all states (after applying the mappings), then the two are equivalent.

Our equivalence in definition 5.1 is not as strict as the homomorphisms equivalence. Unlike homomorphisms, we *do not* seek human MDPs that have the same rewards and transitions as chainworld. In fact, we do not even seek MDPs that result in the same optimal human policy as chainworld. Instead, we only care that the two human MDPs are similar enough to result in the same *optimal AI policy*. As a result, we get the largest set of human MDPs that admits simple planning of optimal interventions by the AI agent.

5.1 Optimal AI policies for chainworld MDPs

Under definition 5.1, the class of MDPs that is equivalent to chainworlds is determined by the space of AI policies that chainworlds can express. In this section, we show that all chainworld MDPs $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$ result in AI optimal policies that follow a "three-window format," which we refer to as $\bar{\Pi}$. Throughout this section, we describe the AI policy in terms of the chainworld states, s_n, where n refers to n-th state on the chain; even though the *previous* human actions are technically part of the AI state, they do not affect the *best current* action in the AI's optimal policy.

A "three-window" AI policy consists of: window 1 (no intervention is effective enough to make human perform the behavior), window 2 (intervention window), and window 3 (human performs behavior without intervention). Two examples are in fig. 3. The size of these windows varies and may even be 0. For example, if the interventions have no effect ($\Delta_\gamma = 0, \Delta_b = 0$) then the intervention window will be size 0. The three windows are a consequence of how the AI's action affects the human's optimal policy; when the AI agent intervenes on the human, it changes the human's MDP parameters, which in turn, might change the human's optimal policy.

To succinctly describe the human's optimal policy, we introduce "human thresholds" t in definition 5.2; when the human is in a state past the threshold, their optimal policy is to pursue the goal. A human with a smaller threshold t will pursue the goal from farther away. An effective AI action moves the threshold t to a state *preceding* the human's current state, so that the human chooses to move.

Definition 5.2 (Human threshold). For a chainworld $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$, define $t \in \{0, \dots, N-1\}$ as the threshold where $\pi^*_\theta(s_n) = 0$ for $n \le t$ and $\pi^*_\theta(s_n) = 1$ for $n > t$.

Even if the AI agent *can* intervene to prompt the human toward the goal, whether or not the optimal AI *does* intervene depends on the configuration of the AI rewards. If intervening has negligible cost, then the AI agent will intervene as soon as it is able. On the other hand, if there is a high cost, then the AI agent will wait until the human is closer to the goal, to minimize the total number of interventions needed. We define AI threshold t_{AI} below, as the point at which the reward of reaching the goal outweighs the cost of interventions required to reach it:

Definition 5.3 (AI threshold). For a human chainworld $\mathcal{M}_\theta \in \mathscr{M}_{\text{chain}}$ and AI MDP $\mathcal{M}_{AI}$, define AI threshold $t_{AI} \in \{0, \dots, N-1\}$ as the chainworld state in which the value of the goal is greater than the value of disengagement. For states s_n where $n > t_{AI}$, the AI values are $V^{\pi_g}_{AI}(s_n) > V^{\pi_d}_{AI}(s_n)$, and for states where $n \le t_{AI}$, the AI values are $V^{\pi_g}_{AI}(s_n) \le V^{\pi_d}_{AI}$.

The human and AI thresholds define the intervention windows for the AI policy in theorem 5.4.

Theorem 5.4 (Chainworld AI policies). *Suppose we are given:*

- *An AI MDP $\mathcal{M}_{AI} = \langle \mathcal{S}_{AI}, \mathcal{A}_{AI}, T_{AI}, R_{AI}, \gamma_{AI} \rangle$, where the actions are to do nothing ($a_{AI} = 0$), intervene on the discount ($a_{AI} = a_\gamma$), or to intervene on burden ($a_{AI} = a_b$)*
- *A human MDP $\mathcal{M}_\theta \in \mathscr{M}_{chain}$, which results in human thresholds t^0_h, t^γ_h, and t^b_h under AI actions 0, a_γ, and a_b, respectively*

Let $t^{min}_h = \min\left\{t^0_h, t^\gamma_h, t^b_h\right\}$ denote the earliest human threshold as a result of any AI action. *Let t_{AI} denote the AI intervention threshold, as in definition 5.3. Then, the optimal AI policy is,*

$$\pi^*_{AI}(s_n) = \begin{cases} 0, & n \le t^{min}_h \\ 0, & t^{min}_h < n \le t_{AI} \\ a_\gamma, & \max\{t_{AI}, t^\gamma_h\} < n \le t^0_h \\ a_b, & \max\{t_{AI}, t^b_h\} < n \le t^0_h \\ 0, & n > t^0_h, \end{cases} \tag{6}$$

and π^*_{AI} belongs to the three-window policy class, $\bar{\Pi}$.

The proof is in appendix B.1 [25]. Note that if both a_b and a_γ are valid options in the intervention window (when $t_{AI} < n \le t^0_h$), then the AI agent will prefer the less expensive intervention. Theorem 5.4 shows that every chainworld results in an optimal AI policy belonging to $\bar{\Pi}$. Theorem 5.5 shows the reverse; for any human MDP whose corresponding AI policy is $\pi_{AI} \in \bar{\Pi}$, there exists a chainworld MDP whose AI policy is also π_{AI}.

Theorem 5.5 (Chainworld equivalence class). *If human MDP $\mathcal{M}_h$ has corresponding AI policy $\pi_{AI} \in \bar{\Pi}$, then $\exists\theta$ for $\mathcal{M}_\theta \in \mathscr{M}_{chain}$ such that $\mathcal{M}_\theta \equiv \mathcal{M}_h$.*

Proof in appendix B.2 [25]. Theorem 5.5 means that *any human MDP* that results in a three-window AI policy—that is, consists of three regions: impossible to help, can be helped by the AI, and does not need help— belongs to the chainworld equivalence class. In section 6, we will show that the AI agent can plan interventions using chainworld as a substitute for another human MDP in the same class, without any loss in performance.

5.2 Realistic human models that are equivalent to chainworld

Ultimately, we care that the chainworld equivalence class contains *realistic* models of humans that align with the behavioral literature. In this section, we provide examples of human MDPs that capture a meaningful behavior not covered by chainworlds, yet whose optimal AI policy is still in the equivalence class $\bar{\Pi}$.

Monotonic chainworlds. In monotonic chainworlds, the closer one gets to the goal, the higher the relative value of pursuing it.

Definition 5.6 (Monotonic chainworlds). For a monotonic chainworld $\mathcal{M}$, the value of goal-pursuit increases closer to the goal: $V^{\pi_g}(s_n) - V^{\pi_d}(s_n) \le V^{\pi_g}(s_{n+1}) - V^{\pi_d}(s_{n+1})$ for all states $n = 1, \dots, N-1$.

For example, consider chainworlds in which the probability of disengagement p_d decreases the closer the agent is to the goal (the human is less likely the quit the closer they are to recovery). Monotonic chainworlds relate to the goal-gradient hypothesis, which states that motivation to reach a goal increases with proximity [23]. In appendix C.1 [25], we prove that all monotonic chainworlds are AI equivalent to our chainworld.

Progress worlds. Progress worlds, while potentially multi-dimensional, have a one-dimensional notion of progress.

Definition 5.7 (Progress worlds). Suppose $\mathcal{M}$ is a D dimensional, path-connected graph with an absorbing goal state s_g, an absorbing disengagement state s_d, and actions that allow movement between states on the graph. Let $d(s, s')$ denote the shortest graph distance from s to s'. $\mathcal{M}$ is a progress world if $d(s, s_d) = d(s', s_d)$ and $d(s, s_g) = d(s', s_g)$ for all pairs of $s, s' \in \mathcal{S}$.

In our PT example, "progress" may depend on a combination of metrics such as joint strength, the ability to perform daily tasks, and so on. We show in appendix C.4 [25] that worlds in which states can be mapped to a one-dimensional distance are equivalent to our chainworlds. This type of equivalence is simple yet useful, as it lets us reduce high-dimensional worlds to a single dimension of interest. Definition 5.7 restricts us to graphs in which all shortest paths between the disengagement and goal state are of the same length. Intuitively, this means that a single chainworld can represent all paths (and therefore, the entire world). Though not all graphs are progress worlds, in our empirical experiments, we test the chainworld AI's robustness to graphs that break this definition.

Multi-chain worlds. In multi-chain worlds, there is a principle dimension that corresponds to progress toward the goal (as in our simple chainworld) but there may be several additional dimensions associated with different ways of dropping out.

Definition 5.8 (Multi-chain worlds). A multi-chain world $\mathcal{M}$ consists of C chains, each of length N_c. The first chain, $c = 0$, is the *goal chain*; when the human reaches the end of this chain, they have reached the goal. The remaining chains, $s_1, \ldots, s_{C-1}$, are disengagement chains; when the human reaches the end of *any* of these chains, they disengage. When $a = 1$, the human moves along the goal chain with probability p_0 while staying still in the disengagement chains. When $a = 0$, the human stays still in the goal chain and (independently) moves along each of the c disengagement chains with probability p_c.

In our PT example, the principle chain might still correspond to the overall strength of the joint as a measure of progress toward recovery. Additional chains, corresponding to the level of motivation, level of pain, etc., may all represent mechanisms that cause disengagement. This form of multi-chain reflects how disengagement is described in the behavioral literature (e.g. [21, 22]). In appendix C.5.1 [25], we show equivalence to multi-chain worlds whose disengagement chains are of length 2, which corresponds to real-world situations in which one of many factors can abruptly trigger disengagement at any point (e.g. the PT patient is re-injured).

Negative effect worlds. These are chainworlds in which the AI intervention has the opposite intended effect on the human.

Definition 5.9 (Negative effect worlds). A negative effect world $\mathcal{M}$ is defined exactly as the chainworld, except that $\Delta_\gamma < 0$ (AI intervention on discount γ_h *decreases* it) or $\Delta_b > 0$ (AI intervention on burden r_b *increases* it).

The efficacy of a behavioral intervention is known to vary by individual (e.g. [5]). In appendix C.2 [25], we prove that negative effect worlds result in AI policies that correspond to chainworlds where the intervention has *no effect* (i.e. $\Delta_\gamma = 0$ and $\Delta_b = 0$).

6 EMPIRICAL ANALYSIS: TESTING ROBUSTNESS OF CHAINWORLD

We test how AI planning using chainworld benefits performance, especially as we remove our assumptions and make the *true* human model dissimilar to chainworld.

6.1 Setup

All experiments are over 200 trials of 15 episodes each, and each trial corresponds to a human whose MDP parameters θ are sampled. Not all settings of θ correspond to individuals that can reach their goal—for example, consider a human whose burden is so high that no AI intervention can make them act. Here, we report results for the subset of sampled humans that can reach the goal under the *oracle AI policy*. Doing so preserves the relative ordering of method performances and reduces noise; in appendix fig 11 [25] we give an example of results that include individuals who never reach the goal.

Baselines. Our baselines are ways to learn the AI policy online. Using data $\mathcal{D}_{AI} = \{(s_{AI}, a_{AI}, s'_{AI}, r_{AI})\}$, the ***model-free*** approach directly estimates Q^*_{AI} via Q-learning. The ***model-based*** method estimates T_{AI} using the observed transitions and then solves for π^*_{AI} with certainty equivalence. Both approaches bypass the need for explicitly solving for a human policy. The ***always γ*** and ***always B*** are "no personalization," in which the AI policy is to always intervene on γ and B, respectively. Our method, ***chainworld***, estimates the parameters θ from $\mathcal{D}_{AI}$.

6.2 Results under *no model misspecification*

In perfect conditions, the AI agent can use chainworld to reach oracle-level performance in the fewest episodes. When the true human matches our inductive bias, i.e. both are chainworlds, we achieve the fastest personalization in fig. 4. In contrast, model-free requires hundreds of episodes before it learns policies that are better than random (which we demonstrate in fig. 10 of the appendix [25]).

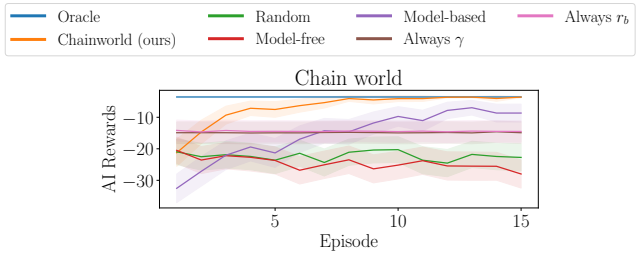
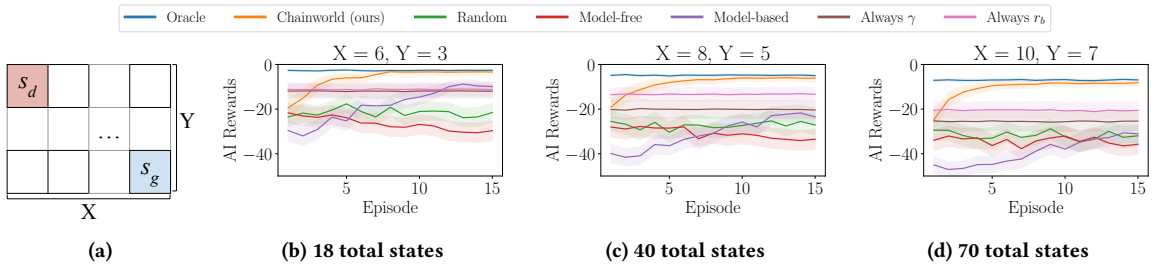


Figure 4: When the true human model is a chainworld, our method rapidly personalizes. Plot is AI rewards (y-axis) over multiple episodes (x-axis). Lines in upper-left personalize quicker.

Our method's performance scales to high-dimensional human models equivalent to the chainworld. In the prior theoretical section, we provided examples of human MDPs that reduce to the chainworld. The gridworld in fig. 5a is one such world since it is a type of distance world. In fig. 5, our method still personalizes the fastest in increasingly large state spaces, because the



(a) (b) 18 total states (c) 40 total states (d) 70 total states

Figure 5: Chainworld scales to large gridworlds. Example gridworld on left. Going right, the grid's width (X) and height (Y) increases.

Assumption	Equiv?	Low misspecification		High misspecification	
		Chainworld (ours)	Top baseline	Chainworld (ours)	Top baseline
Noise in burden r_b	No	**−14.47 ± 3.63**	**−14.43 ± 3.63**	**−35.96 ± 3.36**	**−33.43 ± 3.4**
Noise in utility of goal r_g	No	**−5.53 ± 1.71**	−14.76 ± 3.38	**−6.9 ± 2.22**	−14.66 ± 3.34
Noise in utility of progress loss r_ℓ	No	**−5.97 ± 1.94**	−14.78 ± 3.39	**−11.01 ± 3.29**	**−15.43 ± 3.54**
Noise in utility of disen. r_d	No	**−8.08 ± 2.58**	−15.18 ± 3.44	**−13.38 ± 3.54**	**−14.63 ± 3.41**
Noise in prob. of disen. p_d	No	**−5.03 ± 1.46**	−14.78 ± 3.39	**−6.41 ± 2.45**	**−12.13 ± 4.05**
Noise in prob. of disen. at state 0, p_{d0}	No	**−5.8 ± 1.86**	−14.81 ± 3.4	**−5.83 ± 1.86**	−14.36 ± 3.3
Noise in prob. of losing progress p_ℓ	No	**−5.05 ± 1.51**	−14.78 ± 3.39	**−5.19 ± 1.81**	−13.38 ± 4.13
Noise in prob. of making progress p_g	No	**−5.82 ± 1.77**	−15.24 ± 3.49	**−19.38 ± 4.34**	**−17.85 ± 3.72**
Noise in discount γ_h	No	**−7.75 ± 2.42**	−15.83 ± 3.56	**−20.7 ± 4.03**	**−21.19 ± 3.93**
Params. fixed across states	Yes	—	—		
Mapping many dimensions to chainworld	Yes	—	—		
Wrong distance to goal in mapping	No	**−21.18 ± 3.84**	**−15.62 ± 3.15**	−35.8 ± 3.8	**−24.52 ± 3.3**
Wrong distance to disengagement in mapping	No	**−10.11 ± 2.44**	**−15.62 ± 2.44**	**−27.27 ± 3.86**	**−24.52 ± 3.3**
Diseng. from multiple factors	Yes	—	—		
Human selects actions non-optimally	No	**−7.23 ± 2.27**	−16.01 ± 4.01	**−24.27 ± 3.85**	**−23.39 ± 3.68**
AI intervention has negative effect	Yes				

Table 2: Reward earned by the AI in episode six. Each row is an assumption violated by the environment. Chainworld is better than or within 95% confidence interval of the top-performing baseline (out of five total baselines) in all but one setting. Conditions marked with "yes" in the "Equivalence?" column were shown in section 5.2 to preserve theoretical equivalence under misspecification.

number of chainworld parameters is invariant to the size of the gridworld. On the other hand, model-based degrades; it is *worse* than the personalization-free baselines and the same as random baselines, even after 15 episodes. This is because the transition matrix that model-based must estimate scales with the size of the gridworld. Model-free approaches are even more inefficient in the 2-D setting than in the 1-D chainworld.

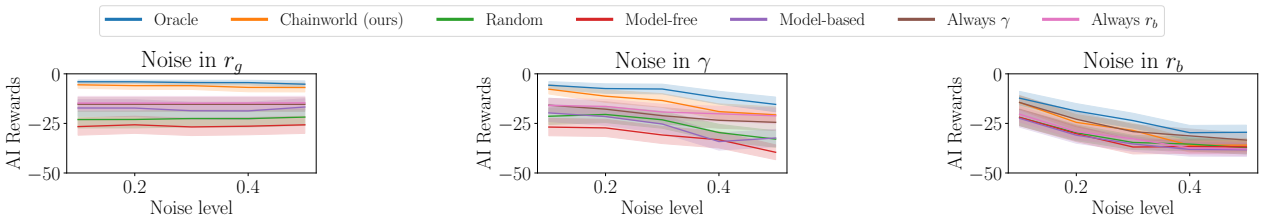
6.3 Robustness results under *model misspecification*

In true frictionful settings, the AI agent will encounter humans that are more sophisticated than the chainworld. Our remaining experiments in table 2 test if AI performance is robust to misspecification when we remove our assumptions about humans. In section 5.2, we theoretically showed that a subset of these assumptions can be removed without affecting the AI. The remaining assumptions we test empirically, and we show our method is more robust to increasing levels of misspecification than baselines. The definition of "low" vs. "high" misspecification is specific to the experiment.

Experiment on noise in chainworld parameters. In this experiment, we test AI performance when the true human model is a chainworld whose parameters vary each timestep due to noise. This mimics situations in which unobservable factors, such as mood, affect parameters, such as burden r_b. We vary each parameter in isolation. Our comparison must account for the domains of different parameters, since $\gamma_h \in (0, 1)$ while rewards such as $r_b \in \mathbb{R}$. At each timestep, the parameter of interest x is sampled uniformly from $x \sim \text{Uniform}(\bar{x} - \epsilon c, \bar{x} + \epsilon c)$, where $\bar{x}$ is the mean parameter value for that individual and the noise level is determined by the parameter range c and the error level $\epsilon \in [0, 1]$. We set parameter range $c = 5$ for reward parameters and to $c = 1$ for transition parameters and γ_h. We define low misspecification as is $\epsilon = 0.1$ and high misspecification as $\epsilon = 0.5$.

Experiment on action selection. Instead of optimally, humans select actions according to a softmax policy, $\pi_h(a|s) \propto \exp\{Q_h(s, a)/\epsilon\}$, where ϵ is the level of noise. We define low misspecification as $\epsilon = 0.05$ and high misspecification as $\epsilon = 0.2$.

Experiment on misspecified mapping. This experiment tests robustness to differences in *model structure*. The true human is no



(a) Chainworld robust to low and high mis. **(b) Chainworld is robust to low mis.** **(c) Environment challenges all methods.**

Figure 6: Examples of robustness experiments. Chainworld is robust to all levels of misspecification fig. 6a, robust to low levels of misspecification with maintainence at high levels fig. 6b, and all methods, including oracle, struggle to perform well in fig. 6c. Details and plots for all environments in appendix D.1 and E.3 [25], respectively.

longer a chainworld, but a gridworld as in fig. 5a. However, the gridworld in this experiment is no longer equivalent to our chainworld because the goal state s_g is not in the lower-right corner at $[X, 0]$. In fact, the equivalence degrades as ϵ increases for $[X, \epsilon]$. We set the grid dimensions as $X = 8, Y = 5$ and define low misspecification as $\epsilon = 1$ and high misspecification as $\epsilon = 4$.

We are robust to low levels of misspecification. In table 2, our method outperforms baselines in *9 out of 12 robustness experiments* under low levels of misspecification. With high misspecification, when our method is not the best, it falls within two standard errors of the next-best method in all but one condition.

Some humans are difficult to intervene on overall, even for the oracle. All methods, including the oracle, earn fewer rewards when the burden parameter r_b is noisy (see fig. 6c). This indicates that it is particularly important to model r_b well in frictionful tasks. For example, we may ensure that features predictive of burden, such as mood, are part of the AI's state space, so that we can estimate r_b.

To reduce (non-equivalent) human models to the chainworld, it is important that we capture distance to goal well. Since chainworld is one-dimensional, it can only represent worlds whose multi-dimensional states can be mapped to one dimension. When such a mapping is not possible, we must choose between capturing progress toward goal (e.g. how far does the patient feel from shoulder recovery?) or distance from disengagement (e.g. how close to giving up does the patient feel?). Under the "wrong distance to goal / disengagement mapping" condition in table 2, we show that capturing progress toward goal matters more. This implies that chainworlds can still be applied to settings where we cannot model all factors that lead to disengagement, so long as we have an accurate way of measuring the human's progress to the goal.

7 CONCLUSION AND FUTURE WORK

In this paper, we introduced Behavior Model Reinforcement Learning (BMRL), a framework for AI agents to intervene on human agents performing frictionful tasks. We proposed a simple model of the human agent– the chainworld– that the AI agent can use to rapidly personalize. Using a novel definition of equivalence between human models in BMRL, we defined a theoretical class of human MDPs that chainworld can generalize to and showed that this class contains behaviorally meaningful models of humans.

Our chainworlds are not psychologically verified human models; in future work, we will formally test the modeling assumptions with user studies. To apply BMRL in the real world, we must also consider the ethics of AI intervention. Mainly, we must ensure the AI does not manipulate the human. BMRL should only be used for people who already have a long-term goal, and the AI must not change that goal. Subgroup fairness should also be considered during learning and personalization.

Although we aimed to be comprehensive in testing chainworld's robustness, there were limitations to our approach. First, we did not evaluate how multiple misspecifications may compound to affect AI performance. Second, our analyses assumed that the mapping from the true MDP to the chainworld is given. In some applications this is reasonable; in PT, a domain expert is likely to know which factors contribute to a patient's perception of "progress" (the mapping from a distance world to a chainworld). In other cases, one will need to learn this mapping in conjunction with the chainworld parameters.

We made several simplifying assumptions on the human + AI interactions. We avoided a POMDP formulation by assuming that there are no delayed effects of the AI's actions on the human MDP. However, habituation (reduced effectiveness of repeated interventions) is a well-studied phenomenon in digital interventions (e.g. [12]). Furthermore, we avoided multi-agent RL by assuming that the human is *not learning*, and instead, is solving an (implicitly) known MDP at each time step. We did not consider suboptimality of the human agent's planning, such as (small) fixed-horizon planning. Finally (and excitingly), BMRL is adaptable to more diverse AI interventions. Our paper focused exclusively on interventions to the human's discount and reward. In many applications, the human's perception of state, actions, and transitions may also be impaired. Similarly, behavioral interventions on perceptions of state, actions, and transitions exist and could be incorporated into our framework.

8 ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391 and the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under 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. ES's work was supported by a gift fund from Benshi.ai and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE2140743.

REFERENCES

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

[2] Albert Bandura. 1999. Social cognitive theory: An agentic perspective. *Asian journal of social psychology* 2, 1 (1999), 21–41.

[3] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*. PMLR, California USA, 783–792.

[4] Jeremiah Michael Brown and Jeffrey Scott Stein. 2022. Putting prospection into practice: Methodological considerations in the use of episodic future thinking to reduce delay discounting and maladaptive health behaviors. *Frontiers in Public Health* 10 (2022), 1020171.

[5] Christopher J Bryan, Elizabeth Tipton, and David S Yeager. 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour* 5, 8 (2021), 980–989.

[6] Kaiqi Chen, Jeffrey Fong, and Harold Soh. 2022. Mirror: Differentiable deep social projection for assistive human-robot communication. In *Robotics: Science and Systems*. Robotics: Science and Systems, New York USA.

[7] Owain Evans, Andreas Stuhlmüller, and Noah Goodman. 2016. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30. AAAI, Arizona USA.

[8] Joseph Fanfarelli, Stephanie Vie, and Rudy McDaniel. 2015. Understanding digital badges through feedback, reward, and narrative: a multidisciplinary approach to building better badges in social environments. *Communication Design Quarterly Review* 3, 3 (2015), 56–60.

[9] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. 2019. Hyperbolic Discounting and Learning over Multiple Horizons. arXiv:1902.06865 [stat.ML]

[10] Robert Givan, Thomas Dean, and Matthew Greig. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147, 1-2 (2003), 163–223.

[11] Babatunde H Giwa and Chi-Guhn Lee. 2021. Estimation of Discount Factor in a Model-Based Inverse Reinforcement Learning Framework. https://hdl.handle.net/1807/125220

[12] Lisa Gotzian. 2023. Modeling the decreasing intervention effect in digital health: a computational model to predict the response for a walking intervention. https://doi.org/10.31219/osf.io/6v7d5

[13] Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. 2021. Inverse decision modeling: Learning interpretable representations of behavior. In *International Conference on Machine Learning*. PMLR, PMLR, Virtual, 4755–4771.

[14] Alireza Khanshan, Pieter Van Gorp, and Panos Markopoulos. 2023. Simulating Participant Behavior in Experience Sampling Method Research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 7 pages. https://doi.org/10.1145/3544549.3585586

[15] Lihong Li, Thomas J Walsh, and Michael L Littman. 2006. Towards a unified theory of state abstraction for MDPs.

[16] Quanying Liu, Haiyan Wu, and Anqi Liu. 2019. Modeling and Interpreting Real-world Human Risk Decision Making with Inverse Reinforcement Learning. arXiv:1906.05803 [cs.LG]

[17] Eran Magen, Carol S Dweck, and James J Gross. 2008. The hidden-zero effect: Representing a single choice as an extended sequence reduces impulsive choice. *Psychological Science* 19, 7 (2008), 648–649.

[18] Cesar A Martin, Daniel E Rivera, Eric B Hekler, William T Riley, Matthew P Buman, Marc A Adams, and Alicia B Magann. 2018. Development of a control-oriented model of social cognitive theory for optimized mHealth behavioral interventions. *IEEE Transactions on Control Systems Technology* 28, 2 (2018), 331–346.

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

[20] Nataliya Mogles, Julian Padget, Elizabeth Gabe-Thomas, Ian Walker, and Jee-Hang Lee. 2018. A computational model for designing energy behaviour change interventions. *User Modeling and User-Adapted Interaction* 28 (2018), 1–34.

[21] Irena Moroshko, Leah Brennan, and Paul O'Brien. 2011. Predictors of dropout in weight loss interventions: a systematic review of the literature. *Obesity reviews* 12, 11 (2011), 912–934.

[22] Isaac Moshe, Yannik Terhorst, Sarah Paganini, Sandra Schlicker, Laura Pulkki-Råback, Harald Baumeister, Lasse B Sander, and David Daniel Ebert. 2022. Predictors of dropout in a digital intervention for the prevention and treatment of depression in patients with chronic back pain: secondary analysis of two randomized controlled trials. *Journal of Medical Internet Research* 24, 8 (2022), e38261.

[23] Tobias Mutter and Dennis Kundisch. 2014. Behavioral mechanisms prompted by badges: The goal-gradient hypothesis. In *ICIS 2014 Proceedings*, Vol. 12. ICIS, New Zealand.

[24] Yael Niv. 2009. Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53, 3 (2009), 139–154.

[25] Eura Nofshin, Siddharth Swaroop, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. 2024. Reinforcement Learning Interventions on Boundedly Rational Human Agents in Frictionful Tasks. arXiv:2401.14923 [cs.AI]

[26] Joonyoung Park and Uichin Lee. 2023. Understanding Disengagement in Just-in-Time Mobile Health Interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–27.

[27] Peter Pirolli. 2016. A computational cognitive model of self-efficacy and daily adherence in mHealth. *Translational behavioral medicine* 6, 4 (2016), 496–508.

[28] Balaraman Ravindran and Andrew G Barto. 2002. Model minimization in hierarchical reinforcement learning. In *Abstraction, Reformulation, and Approximation: 5th International Symposium, SARA*, Vol. 2371. Springer, Springer, Berlin, Heidelberg, Canada, 196–211.

[29] Balaraman Ravindran and Andrew G Barto. 2004. Approximate homomorphisms: A framework for non-exact minimization in Markov decision processes.

[30] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. 2018. Where do you think you're going? inferring beliefs about dynamics from behavior. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 1461–1472.

[31] Siddharth Reddy, Sergey Levine, and Anca Dragan. 2021. Assisted perception: optimizing observations to communicate state. In *Conference on Robot Learning*. PMLR, PMLR, London UK, 748–764.

[32] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. 2019. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*. PMLR, PMLR, California, USA, 5670–5679.

[33] Hanan Shteingart and Yonatan Loewenstein. 2014. Reinforcement learning and human behavior. *Current opinion in neurobiology* 25 (2014), 93–98.

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

[35] Seyed Amin Tabatabaei, Mark Hoogendoorn, and Aart van Halteren. 2018. Narrowing reinforcement learning: Overcoming the cold start problem for personalized health interventions. In *PRIMA 2018: Principles and Practice of Multi-Agent Systems: 21st International Conference*. Springer, Springer, Tokyo Japan, 312–327.

[36] Aaquib Tabrez, Shivendra Agrawal, and Bradley Hayes. 2019. Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Korea, 249–257. https://doi.org/10.1109/HRI.2019.8673104

[37] Véronique A Taylor, Isabelle Moseley, Shufang Sun, Ryan Smith, Alexandra Roy, Vera U Ludwig, and Judson A Brewer. 2021. Awareness drives changes in reward value which predict eating behavior change: Probing reinforcement learning using experience sampling from mobile mindfulness training for maladaptive eating. *Journal of behavioral addictions* 10, 3 (2021), 482–497.

[38] Véronique A Taylor, Isabelle Moseley, Shufang Sun, Ryan Smith, Alexandra Roy, Vera U Ludwig, and Judson A Brewer. 2021. Awareness drives changes in reward value which predict eating behavior change: Probing reinforcement learning using experience sampling from mobile mindfulness training for maladaptive eating. *Journal of behavioral addictions* 10, 3 (2021), 482–497.

[39] Jonas Tebbe, Lukas Krauch, Yapeng Gao, and Andreas Zell. 2021. Sample-efficient reinforcement learning in robotic table tennis. In *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, IEEE, China, 4171–4178.

[40] Mohammad Thabet, Massimiliano Patacchiola, and Angelo Cangelosi. 2019. Sample-efficient deep reinforcement learning with imaginary rollouts for human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, IEEE, Macau, 5079–5085.

[41] Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. 2022. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms* 15, 8 (2022), 255.

[42] Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. 2020. Plannable Approximations to MDP Homomorphisms: Equivariance under Actions. arXiv:2002.11963 [cs.LG]

[43] Shihan Wang, Chao Zhang, Ben Kröse, and Herke van Hoof. 2021. Optimizing adaptive notifications in mobile health interventions systems: reinforcement learning from a data-driven behavioral simulator. *Journal of medical systems* 45 (2021), 1–8.

[44] Xuhai Xu. 2022. Towards Future Health and Well-being: Bridging Behavior Modeling and Intervention. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, USA, 1–5.

[45] Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Tingnan Zhang, Jie Tan, and Vikas Sindhwani. 2020. Data efficient reinforcement learning for legged robots. In

Conference on Robot Learning. PMLR, PMLR, Virtual, 1–10.

[46] Guanghui Yu and Chien-Ju Ho. 2022. Environment Design for Biased Decision Makers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, Austria, 592–598.

[47] Chao Zhang, Joaquin Vanschoren, Arlette van Wissen, Daniël Lakens, Boris de Ruyter, and Wijnand A IJsselsteijn. 2022. Theory-based habit modeling for enhancing behavior prediction in behavior change support systems. *User Modeling and User-Adapted Interaction* 32, 3 (2022), 389–415.

[48] Chao Zhang, Shihan Wang, Henk Aarts, and Mehdi Dastani. 2021. Using Cognitive Models to Train Warm Start Reinforcement Learning Agents for Human-Computer Interactions. arXiv:2103.06160 [cs.AI]

[49] Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. 2020. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems* 33 (2020), 19238–19250.

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