Human Models for Planning Behavioral Interventions with Reinforcement Learning

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

Introduction

In many AI+human applications for 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 or passing an online course. Here, an AI agent plans *when* and *how* best to provide behavioral support in the form of a digital intervention (e.g. a suggested PT routine for the day).

Current reinforcement learning (RL) approaches have two major drawbacks when used to solve for the AI agent's intervention policy. First, default methods are too data-intensive for our online setting. For example, online algorithms in robotics require thousands of interactions to learn reasonable policies (e.g. in Yang et al. (2020)), but in frictionful tasks, we are limited to *tens to hundreds* of interactions per person (Trella et al. 2022). Second, existing planning methods solve for the AI agent's optimal policy by modeling the human as a black-box transition or value function. Unfortunately, in learning black-box representations of the human, we lose the ability to interpretably attribute human behavior to the model learned by the AI. I target interpretable and effective planning by the AI agent through the use of carefully crafted human models. To create and work with these human models, my work bridges across machine learning, behavioral science, and human-computer interaction (HCI).

Behavioral science provides us with formal theories and models of human decision-making in frictionful tasks. However, there is a gap in how to instantiate high-level constructs from behavioral science (such as temporal discounting in humans) into computational models (which describe the scale and functional forms of how temporal discounting changes over time). Machine learning offers paradigms that can elegantly encode the behavioral assumptions needed to form computational models. For example, temporal discounting from behavior science can be connected to the discount factor, γ, which is part of a Markov Decision Process (MDP). Such explicit computational models are powerful because they (1) provide the link between behavioral assumptions and the observed data; and (2) can be incorporated into the AI agent's planning. But, models that show promise in theory and simulation must be tested with real end-users, and user studies guided by HCI design principles can evaluate effectiveness. To fill these gaps, I aim to address the following questions:

Q1. What is the model? Reducing complex behavioral models to simpler ones that can be used for AI planning.

Q2. How to learn the human model? Updating the human model to individual-level data observed online.

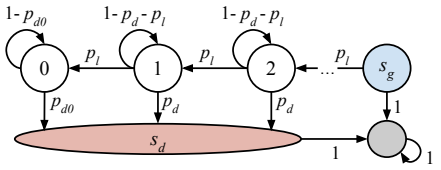
Q3. How to use the human model for intervention? Learning and testing intervention policies that work with real users.

Q1: What Is the Model?

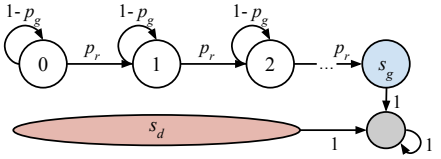
In Nofshin et al. (2024b), I define Behavior Model RL (BMRL), a framework in which an AI agent intervenes on a human modeled as a Markov Decision Process (MDP) with "maladapted" parameters. In maladapted human MDPs, humans act according to optimal policies that do not reach their stated goal (for example, the goal of rehabilitating an injury but taking actions such as skipping PT). One example of a maladapted MDP is having an extremely low discount rate, γ. This represents myopic decision-making, wherein an individual forgoes the long-term goal (being active) to avoid experiencing friction in the short-term (unpleasantness of exercising). In BMRL, the AI interventions improve γ. This model is inherently more interpretable because its parameters correspond to constructs from behavioral science (e.g. the human's level of discounting); when we infer these parameters from data, we learn something about the construct and how it relates to interventions.

The choice of human MDP is key to BMRL. My work considers how different assumptions about the human MDP impact the quality of the AI policy. In Nofshin et al. (2024b), I introduce the "chainworld", a simple human MDP that captures people's motivation as an increasing function of their proximity to the goal, also known as the goal-gradient hypothesis (Mutter and Kundisch 2014). Figure 1 visualizes the chainworld. Chainworld is an example of how modeling the human MDP allows the AI to achieve rapid personalization (see fig. 2).

BMRL can also uncover cross-disciplinary discrepancies in how we model humans. In psychology, humans are known as *hyperbolic discounters*; that is, they discount rewards that are t timesteps away using the function $1/1+kt$. However,



(a) When the human abstains from the behavior, they lose progress with probability p_ℓ or slip into the disengagement state s_d with probability p_d.



(b) When the human performs the behavior, they make progress toward the goal state s_g with probability p_g.

Figure 1: Graphical representation of the chainworld. Each state on the chain represents the progress toward the goal state, s_g.

in RL, we assume agents (including human ones) are *exponential discounters*– their discount functions are γ^t– because hyperbolic discounting is computationally intractable and must be approximated[1]. In Moore et al. (2025), we consider how the AI's assumptions about the human's discounting function affect the quality of its intervention policy. Despite the fact that humans are hyperbolic discounters, our simulations demonstrate that modeling people as exponential discounters lets us learn better AI policies online than if we were to use a hyperbolic approximation.

Q2: How Do We Learn the Human Model?

Once we define the human's MDP (or, for that matter, any model of the human's decision-making), we must infer the parameters of the model from behavioral data that is observed online. When the data is *human demonstration data in a single environment*, there is inherent non-identifiability in the human MDP parameters, as we show in (Ankile et al. 2023). For example, a human with myopic discounting vs. a human that perceives low rewards on the goal state will both behave according to goal-avoidant policies. When there is demonstration data from *multiple environments*, it is possible to combat non-identifiability by aggregating information from demonstrations across environments, but this becomes a difficult search problem over which environment to show the user, which we address in Nitschke et al. (2024). Finally, across all data sources, our inference over the human model parameters must consider the noisiness and scale of data that is available, which I have explored in Shin et al. (2023).

Q3: How To Use Human Models To Intervene?

Beyond algorithmic contributions, I have conducted user studies to test two components of my human models research. The first component is ongoing and tests whether the BMRL framework leads to measurably better behavioral outcomes. We are conducting this study in the language learning domain, which has the elements of a difficult behavior change task (long-term reward, difficult in the short-term), but is easy to recruit for and test. The second component is in explainable AI (XAI); not only should AI systems provide quality decisions, but they should also provide the explanations behind them. My XAI project asserts that different kinds of explanations are needed for different tasks. For example, when suggesting an exercise, the user might benefit from a concise explanation so that they can quickly check whether it aligns with their goals. On the other hand, when recommendations are being examined for subgroup bias, the developer might benefit from detailed explanations that reveal every part of the AI's decision. In Nofshin et al. (2024a) we consider whether human models can be used to simulate– and therefore anticipate– which properties are most appropriate for each task.

[1]Both γ and k are hyperparameters that control the severity of discounting

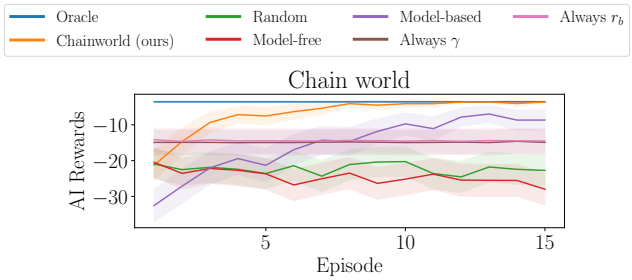


Figure 2: Using our chainworld model (orange), the AI reaches oracle-level performance (blue) quickest. Plot is AI rewards (y-axis) over multiple episodes (x-axis). Lines in upper-left corner mean the AI personalizes more quickly.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS1750358, Grant No. IIS-2107391, and the Graduate Research Fellowship Program Grant No. DGE 2140743. 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. The research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award number OD P41EB028242. Eura Nofshin's work on the project was supported by a gift fund from Benshi.ai.

References

Ankile, L. L.; Ham, B. S.; Mao, K.; Shin, E.; Swaroop, S.; Doshi-Velez, F.; and Pan, W. 2023. Discovering User Types: Mapping User Traits by Task-Specific Behaviors in Reinforcement Learning. arXiv:2307.08169.

Moore, I. M.; Nofshin, E.; Swaroop, S.; Murphy, S.; Doshi-Velez, F.; and Pan, W. 2025. When and Why Hyperbolic

Discounting Matters for Reinforcement Learning Interventions. In *Reinforcement Learning Conference*. Edmonton, Canada: Reinforcement Learning Journal.

Mutter, T.; and Kundisch, D. 2014. Behavioral mechanisms prompted by badges: The goal-gradient hypothesis.

Nitschke, P.; Ankile, L. L.; Shin, E.; Swaroop, S.; Doshi-Velez, F.; and Pan, W. 2024. AMBER: An Entropy Maximizing Environment Design Algorithm for Inverse Reinforcement Learning. In *Proceedings of the ICLM 2024 Models of Human Feedback for AI Alignment Workshop*. Workshop paper.

Nofshin, E.; Brown, E.; Lim, B.; Pan, W.; and Doshi-Velez, F. 2024a. A Sim2Real Approach for Identifying Task-Relevant Properties in Interpretable Machine Learning. *arXiv preprint arXiv:2406.00116*.

Nofshin, E.; Swaroop, S.; Pan, W.; Murphy, S.; and Doshi-Velez, F. 2024b. Reinforcement Learning Interventions on Boundedly Rational Human Agents in Frictionful Tasks. arXiv:2401.14923.

Shin, E.; Klasnja, P.; Murphy, S.; and Doshi-Velez, F. 2023. Online model selection by learning how compositional kernels evolve. *Transactions on Machine Learning Research*.

Trella, A. L.; Zhang, K. W.; Nahum-Shani, I.; Shetty, V.; Doshi-Velez, F.; and Murphy, S. A. 2022. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8): 255.

Yang, Y.; Caluwaerts, K.; Iscen, A.; Zhang, T.; Tan, J.; and Sindhwani, V. 2020. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning*, 1–10. PMLR, Virtual: PMLR.