RL with Temporal Representations Captures Reliable Phenotypes of Adaptive Persistence Behavior

Yixin Chen
Department of Psychological and Brain Sciences
Boston University
Boston, MA 02215
yxyxchen@bu.edu

Joseph T. McGuire
Department of Psychological and Brain Sciences
Boston University
Boston, MA 02215
jtmcg@bu.edu

Abstract

Parameters of reinforcement learning (RL) models fit to behavioral data have potential to serve as meaningful measures of latent individual differences. In such applications, RL parameters are usually estimated based on behavioral data from simple decision-making tasks in which the central decision is which of several discrete alternative actions to select. Previous studies have documented associations between these parameter estimates and cognitive and biological processes. However, in real life, many decisions require adaptation across contexts in both which action to choose and when to act. Only a limited amount of previous work has explored whether task-derived RL parameters can capture meaningful individual differences in this type of higher-dimensional behavioral output space. In this paper, we focus on voluntary persistence behavior—that is, deciding how long to continue waiting for an uncertain future prospect—a domain of temporally extended behavior that impacts the attainment of meaningful real-world outcomes and is thought to be altered in a range of mental health conditions. We developed an RL model with temporal representations and applied it to a behavioral paradigm for studying the experience-driven calibration of persistence, the *willingness-to-wait* task. Like human decision makers, our RL model was able to calibrate its level of persistence in a context-appropriate manner. More importantly, parameters of the model were able to capture multifacted individual differences in adaptive persistence behavior. Across independent testing sessions, these task-derived RL parameters were found to have moderate to high test-retest reliability, consistent with reflecting meaningful behavioral variation.

Keywords: reinforcement learning, computational phenotyping, computational psychiatry, individual differences, adaptive persistence, test-retest reliability

1 Introduction

Parameters of computational models fit to behavioral data have been increasingly used to evaluate individual differences (e.g., Patzelt, Hartley, & Gershman, 2018), an approach known as computational phenotyping. Task-derived model parameters have potential to offer deeper mechanistic insights compared with descriptive statistical summaries of behavior. It also has been proposed that such parameters can track latent factors affected in psychiatric conditions (Wiecki & Frank, 2010).

Previous studies have shown that reinforcement learning (RL) models can compress across-individual variance in raw task data into a small set of parameters. Evidence exists that these parameters have meaningful associations with cognitive processes (e.g., reward/punishment learning and optimism bias; Lefebvre et al., 2017), biological factors (e.g., dopamine function; Frank et al., 2007), and clinical phenotypes (e.g., major depression; Brown et al., 2021).

RL-based computational phenotyping has predominantly been used in combination with classic decision-making paradigms such as multi-arm bandit tasks. These paradigms present well-defined alternatives simultaneously on discrete trials, and essentially involve only one behavioral output per trial (i.e., which option to choose). In real life, however, important decisions occur in less-structured environments, in which decision makers encounter alternatives in an extended sequence. Such decisions involve two dimensions, which action to select and when to act. In recent years, an increasing number of RL models with temporal representations have been developed to tackle such decisions (e.g., Biedenkapp et al., 2021; Constantino & Daw, 2015; Niv et al., 2006; Moustafa et al., 2008). However, the majority focus on providing optimal solutions or replicating experimental effects, and the utility of RL models in capturing individual differences in complex decision environments is relatively understudied.

One example of two dimensional decisions in less-structured environments is deciding how long to keep waiting for an uncertain future prospect. Unconditional persistence is not always advantageous (Fawcett et al., 2012). Rather, different levels of persistence are favored in environments with different temporal statistics. Previous studies found that people can rapidly recalibrate their persistence behavior—becoming either more or less willing to tolerate delay—after a short period of direct experience with the temporal statistics of a new environment. Furthermore, substantial individual variation is apparent both in baseline levels of persistence and in the flexibility of recalibration across environments (McGuire & Kable, 2015). It is unclear, however, whether RL models can add to our understanding of these multifaceted individual differences.

In this paper, we present an RL model of adaptive persistence decisions and examine its ability to capture individual differences efficiently and reliably. The model was developed and tested using a behavioral paradigm for studying the experience-driven calibration of persistence, the *willingness-to-wait* task (McGuire & Kable, 2015, 2012; Lempert et al., 2018). In this task, participants continuously decide whether (and for how long) to continue waiting for a delayed and temporally uncertain reward. The alternative to waiting is to disengage from the current reward opportunity and move on to a new one. Akin to foraging paradigms (McNamara, 1982), participants' goal in the task is to maximize the *rate* of reward accrual over the course of a fixed time period by choosing between exploiting and abandoning individually encountered reward prospects. The probability distributions governing delay durations are manipulated across environments so that either higher or lower levels of persistence lead to better outcomes.

The outline of the remainder of the paper is as follows. Section 2 introduces an empirical data set (n = 60) that served as a test bed for model development. Section 3 introduces how we represent the willingness-to-wait task as a Markov decision process (MDP) and describes the structure of the model. Section 4 presents modeling results, including evaluations of test-retest reliability of parameter estimates in an independent large online data set.

2 Empirical data

The empirical data set consisted of a subset of the participants from a previously published study using the willingness-to-wait task (Lempert et al., 2018). 60 participants were randomly assigned to either a high-persistence (HP) environment or a limited-persistence (LP) environment. (Additional participants from the original study who underwent in acute stress manipulation were not included here.) Each participant completed three 7-min blocks of the task in the assigned environment. On each trial, a token was initially worth 0¢ and would mature to 10¢ after a random delay (see Figure 1a).

When the participant either sold the matured token or quit early, a new token was presented after a 2-s inter-trial interval. The HP group experienced delays drawn from a uniform distribution spanning 0-20 s. The LP group experienced delays drawn from a heavy-tailed distribution (Generalized Pareto distribution, $k = 3, \mu = 0, \sigma = 1.5$, truncated at 40 s; see Figure 1b).

The reward-maximizing behavioral strategy was defined as the giving-up time that maximized the average rate of return, denoted T^*. We denote the corresponding maximal average rate of return as ρ^*. In the HP environment, rewards continuously drew nearer as time elapsed such that the optimal strategy was always to wait ($\mathrm{T}^*_{\mathrm{HP}} = 20$ s, $\rho^*_{\mathrm{HP}} = 0.83$ ¢/

s). In the LP environment, rewards became sparser as time elapsed and the best strategy was to wait only until a certain threshold and quit if the reward had not been delivered ($T^*_{LP} = 2.22$ s, $\rho^*_{LP} = 0.93$ ¢/ s).

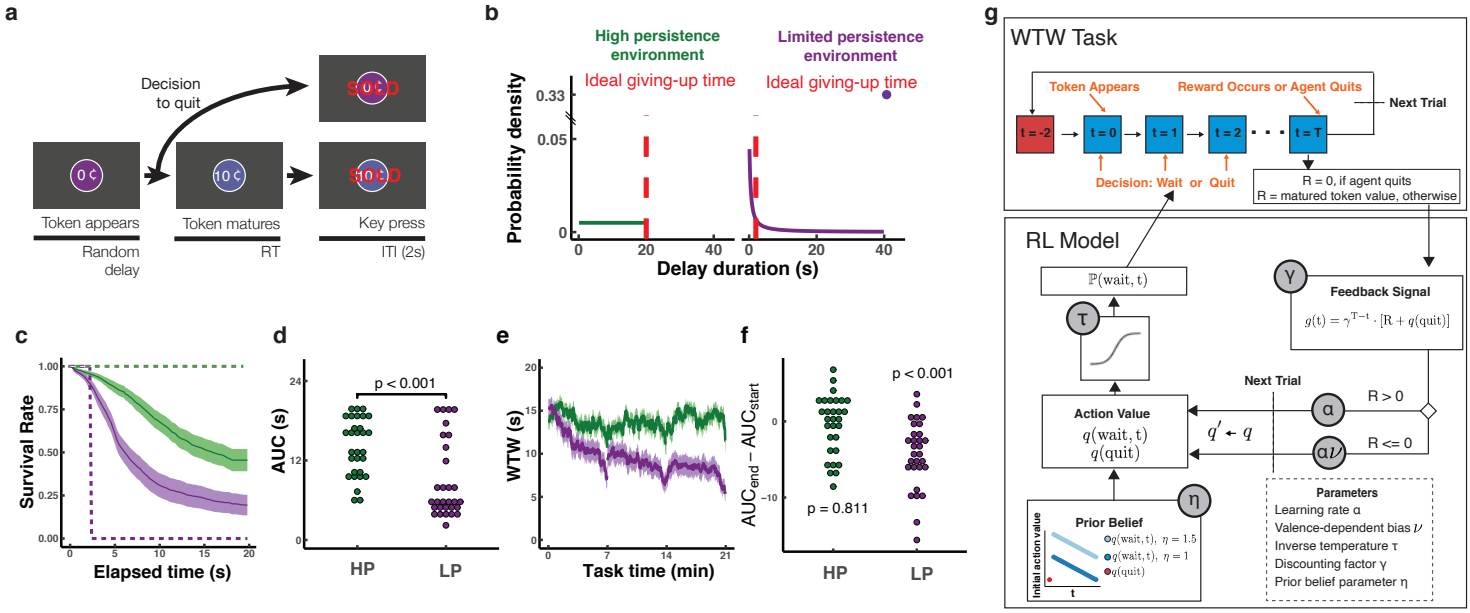

Figure 1: (a)-(b) Task design. (c)-(f) Empirical results. (g) Model Schematics.

To summarize each individual's overall level of behavioral persistence, we constructed a Kaplan-Meier survival curve, separately for each participant, that aggregated data across trials. The survival curve estimated the probability of the token "surviving" various lengths of time without the participant quitting, conditional on the reward not yet having been delivered (rewarded trials were treated as censored observations). The analysis was restricted to the 0 to 20 s range for which we had observations in both environments. The area under the survival curve (AUC) captured how much of the first 20 s the participant was willing to wait on average.

To characterize variation in behavior over time, we calculated a trialwise estimate of each individual's willingness to wait (WTW) throughout the experiment. During quit trials, this estimate was set equal to the observed waiting time. During other trials, the estimate was set equal to the longest time waited since the last quit trial. The estimates were capped at 20 s to make the two environments comparable and were resampled into 2-s bins before averaging across individuals. As a second measure of change over time, we estimated AUC using each individual's data in the first and the last half-block of the task (each with duration 3.5 minutes) and calculated the difference ($AUC_{end} - AUC_{start}$).

Figure 1c shows empirical survival curves averaged across participants in each environment and Figure 1d shows the associated AUC per individual. Figure 1e plots the local WTW estimate as a function of task time. Participants in the HP group were willing to persist longer (median AUC= 14.7 s) than those in the LP group (median AUC = 5.99 s, Wilcoxon rank-sum $p < 0.001$). Persistence level as measured by AUC did not change significantly from beginning to end in the HP group (median $AUC_{end} - AUC_{start}$ = 0.297 s, Wilcoxon signed-rank $p = 0.811$), whereas it significantly decreased in the LP group (median $AUC_{end} - AUC_{start}$ = -3.86 s, signed-rank p < 0.001, Figure 1f).

Substantial individual differences accompanied the group-level effects described above. Figures 1d and 1f show significant variability within each group in overall AUC and start-vs-end change in AUC. These two measures were relatively independent in both environments (HP: Spearman $\rho = 0.418$, $p = 0.021$; LP: Spearman $\rho = 0.176$, $p = 0.350$).

3 RL model

Figure 1g (top) shows the willingness-to-wait task as an MDP. Time within a trial is divided into as a sequence of discrete, evenly spaced time steps, each with a duration of 1 s (e.g., $t = 0, 1, 2, ...$). In each trial, the agent starts at the initial time step $t = -2$. A token will appear after a 2-s ITI, at $t = 0$. In each following time step, the agent needs to choose whether to keep waiting or quit. If the agent chooses to wait and the token remains unmatured, it will proceed to the next time step. Alternatively, if the agent quits or the token matures in this time step, it will return to $t = -2$ to start a new trial. For each trial, we record the trial-wise payoff as R and the time when the agent exits the trial as T. If the agent voluntarily quits, $R = 0$. Alternatively, if it waits until the token matures, $R = 10$.

Figure 1g (bottom) shows the structure of our RL model. The value of continuing to wait varies with time elapsed since the token's onset, and is denoted $q(\text{wait}, \text{t})$. The value of quitting, $q(\text{quit})$, is constant for all temporal states. The agent makes a wait-or-quit choice at each 1-s time step by comparing the values of both options, $\text{P}(\text{wait}, \text{t}) = \frac{1}{1+e^{-\tau[q(\text{wait},\text{t})-q(\text{quit})]}}$, where τ is the inverse temperature parameter that controls action selection noise.

Adopting a batch learning approach, the value estimates for all time steps t encountered in the current trial are updated simultaneously upon trial completion. The q values are updated toward an update target $g(\text{t}) = \gamma^{\text{T}-\text{t}} \cdot [\text{R} + q(\text{quit})]$, where γ is the temporal discounting parameter. R represents the current trial's reward outcome and $q(\text{quit})$ stands for the expected discounted reward from future trials.

Trials with positive and non-positive payoffs are updated separately as follows:

$$q'(\text{quit}) = \begin{cases} q(\text{quit}) + \alpha \cdot [g(\text{t} = -2) - q(\text{quit})], & \text{if R} > 0 \\ q(\text{quit}) + \alpha \cdot \nu[g(\text{t} = -2) - q(\text{quit})], & \text{if R} \leq 0 \end{cases}$$

$$q'(\text{wait}, \text{t}) = \begin{cases} q(\text{wait}, \text{t}) + \alpha \cdot [g(\text{t}) - q(\text{wait}, \text{t})], & \text{if R} > 0 \\ q(\text{wait}, \text{t}) + \alpha \cdot \nu[g(\text{t}) - q(\text{wait}, \text{t})], & \text{if R} \leq 0 \end{cases}$$

where the learning rate parameter α controls the step size, the valence-dependent bias parameter ν controls differential learning from positive and non-positive rewards, and $\text{t} = -2$ is the initial time step in a trial aligned with the onset of the 2-s inter-trial interval.

We define the initial value of $q(\text{quit})$ as $\frac{(\rho^*_{\text{HP}}+\rho^*_{\text{LP}})\cdot 0.5}{0.15}$. The numerator is the average optimal rate of return across the HP and LP environments, and the denominator is a scaling constant. In this way, we guarantee the initial value is in an approximate range of the ground truth and the agent is agnostic about in which environment in which it starts.

Our scheme for initializing $q(\text{wait}, \text{t})$ is based on the assumption that, in the absence of experience, there will be some upper limit on an individual's willingness to wait. It is also consistent with the observation that, in contexts with open-ended delays, the expected remaining delay can increase linearly as a function of time elapsed (Barabási, 2005). To achieve this, we choose linearly decreasing initial values as a function of elapsed time modulated by a free parameter η, $q(\text{wait}, \text{t}) = -0.1\text{t} + \eta + q(\text{quit})$. A larger value of η indicates a higher initial propensity to wait (Figure 1g bottom, Prior Belief).

4 Modeling results

Our RL model was able to reproduce the effect of temporal environment observed in the empirical data. We simulated two virtual participants playing the task for 2 hours, one in the HP environment while the other in the LP environment. The parameters were hand-selected. Figure 2a shows, for each simulated participant, the AUC values in the first 16 minutes (split into four intervals of equal size) and the last 4 minutes of the 2 hours. The RL model learned to wait longer in the HP environment than in the LP environment and achieved nearly optimal performance asymptotically. Figure 2b plots the learned relative value of waiting compared to quitting, $q(\text{wait}, t) - q(\text{quit})$, at different time points throughout the simulation. In the HP environment, the asymptotic relative value of waiting was always greater than zero, consistent with the optimal strategy of always continuing to wait. In the LP environment, it plummeted within the first several seconds, consistent with the optimal strategy of waiting up to a certain threshold.

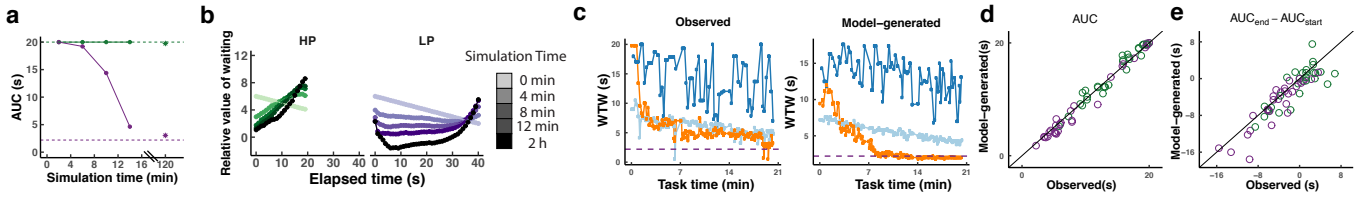

Figure 2: All model-generated results were averaged across 10 simulations. (a)-(b) Simulating group-level effects of temporal environment. (c)-(e) Capturing individual differences.

Model parameters were estimated for each individual using Bayesian methods implemented in Stan with uniform priors. Our RL model efficiently captured multiple aspects of behavioral variation. Figure 2c plots the observed trialwise WTW estimates of three example participants in the LP group alongside model-generated behavior using individually fit parameters (averaged across 10 simulations). The RL model faithfully replicated WTW at the beginning and the end of the task, the degree of adjustment over time, and the level of random variation across trials. Across all participants, our RL model explained a significant portion of the variance in AUC (95.73% in HP and 98.22% in LP, Figure 2d) and in $\text{AUC}_{\text{end}} - \text{AUC}_{\text{start}}$ (HP = 68.05% , LP = 82.52%, Figure 2e).

Mixed results have previously been reported for test-retest reliability of task-derived RL parameters (Weidinger et al., 2019; Brown et al., 2021), casting doubt on whether they capture meaningful individual differences or capture predictive/explanatory factors relevant to mental health. To evaluates the reliability of the parameters in our RL model, we conducted a test-retest study on Amazon Mechanical Turk. Participants complete two sessions of the willingness-to-wait task spaced approximately three weeks apart. Each session consisted of 10 minutes in the LP environment followed by 10 minutes in the HP environment. Individual-level parameter estimates from Session 1 were significantly correlated with those from Session 2 (Figure 3, Spearsman ρ shown in red). The results suggest task-derived RL parameters can capture stable, trait-like individual variation in persistence-related decision-making behavior.

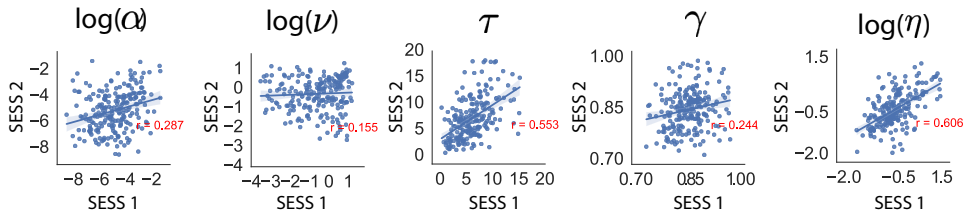

Figure 3: Test-retest reliability of RL parameters.

References

Barabási, A. L. (2005, may). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211. Retrieved from www.nature.com/nature doi: 10.1038/nature03459

Biedenkapp, A., Rajan, R., Hutter, F., & Lindauer, M. (2021). TempoRL: Learning When to Act. Retrieved from http://arxiv.org/abs/2106.05262

Brown, V. M., Zhu, L., Solway, A., Wang, J. M., McCurry, K. L., King-Casas, B., & Chiu, P. H. (2021). Reinforcement Learning Disruptions in Individuals with Depression and Sensitivity to Symptom Change following Cognitive Behavioral Therapy. *JAMA Psychiatry*, *78*(10), 1113–1122. doi: 10.1001/jamapsychiatry.2021.1844

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. , *15*(4), 837–853. doi: 10.3758/s13415-015-0350-y

Fawcett, T. W., McNamara, J. M., & Houston, A. I. (2012). When is it adaptive to be patient? A general framework for evaluating delayed rewards. *Behavioural Processes*, *89*(2), 128–136. doi: 10.1016/j.beproc.2011.08.015

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(41), 16311–16316. doi: 10.1073/pnas.0706111104

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 67. doi: 10.1038/s41562-017-0067

Lempert, K. M., McGuire, J. T., Hazeltine, D. B., Phelps, E. A., & Kable, J. W. (2018, feb). The effects of acute stress on the calibration of persistence. *Neurobiology of Stress*, *8*, 1–9. Retrieved from https://doi.org/10.1016/j.ynstr.2017.11.001 doi: 10.1016/j.ynstr.2017.11.001

McGuire, J. T., & Kable, J. W. (2012). Decision makers calibrate behavioral persistence on the basis of time-interval experience.
doi: 10.1016/j.cognition.2012.03.008

McGuire, J. T., & Kable, J. W. (2015). Medial prefrontal cortical activity reflects dynamic re-evaluation during voluntary persistence. *Nature Neuroscience*, *18*(5), 760–766. doi: 10.1038/nn.3994

McNamara, J. (1982, apr). Optimal patch use in a stochastic environment. *Theoretical Population Biology*, *21*(2), 269–288. doi: 10.1016/0040-5809(82)90018-1

Moustafa, A. A., Cohen, M. X., Sherman, S. J., & Frank, M. J. (2008). A role for dopamine in temporal decision making and reward maximization in parkinsonism. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *28*(47), 12294–12304. doi: 10.1523/jneurosci.3116-08.2008

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2006). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. doi: 10.1007/s00213-006-0502-4

Patzelt, E. H., Hartley, C. A., & Gershman, S. J. (2018). Computational Phenotyping: Using Models to Understand Individual Differences in Personality, Development, and Mental Illness. *Personality Neuroscience*, *1*. Retrieved from https://doi.org/10.1017/pen.2018.14 doi: 10.1017/pen.2018.14

Weidinger, L., Gradassi, A., Molleman, L., & van den Bos, W. (2019). Test-retest reliability of canonical reinforcement learning models.. doi: 10.32470/ccn.2019.1053-0

Wiecki, T. V., & Frank, M. J. (2010). *Neurocomputational models of motor and cognitive deficits in Parkinson's disease* (Vol. 183) (No. C). Elsevier B.V. Retrieved from http://dx.doi.org/10.1016/S0079-6123(10)83014-6 doi: 10.1016/S0079-6123(10)83014-6