ABSTRACT

Title of Thesis: INFORMATION UNCERTAINTY INFLUENCES LEARNING STRATEGY FROM SEQUENTIALLY DELAYED REWARDS

Thesis Directed By: Caroline Charpenier, Assistant Professor, University of Maryland

The problem of temporal credit assignment has long been posed as a nontrivial obstacle to identifying signal from data. However, human solutions in complex environments, involving repeated and intervening decisions, as well as uncertainty in reward timing, remain elusive. To this end, our task manipulated uncertainty via the amount of information given in their feedback stage. Using computational modeling, two learning strategies were developed that differentiated participants' updates of sequentially delayed rewards: eligibility trace whereby previously selected actions are updated as a function of the temporal sequence - and tabular update - whereby additional feedback information is used to only update systematically-related rather than randomly related past actions. In both models, values were discounted over time with an exponential decay. We hypothesized that higher uncertainty would be associated with (i) a switch from tabular to eligibility strategy and (ii) higher rates of discounting. Participants' data (N = 142) confirmed our first hypothesis, additionally revealing an effect of the starting condition. However, our discounting hypothesis had only weak evidence of an effect and remains an open question for future studies. We explore potential explanations for these effects and possibilities of future directions, models, and designs.

INFORMATION UNCERTAINTY INFLUENCES LEARNING STRATEGY FROM SEQUENTIALLY DELAYED REWARDS

By

Sean Richard Maulhardt

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2023

Advisory Committee:
Professor Caroline Charpentier, Chair
Michael R Dougherty
Matthew R Roesch

Introduction

The credit assignment (CA) problem poses an obstacle for various computational systems, wherein its solution allows the system to learn from environmental states. CA involves determining the causal contributions from various states to an outcome. Potential contributing states can be thought of as any assortment of variables, from microbiological enzymes to ingredients for a dinner. For instance, let's consider the scenario of receiving a compliment at dinner. To attribute credit accurately, one must recall past behaviors that contributed to the overall pleasantness of the meal. While a pinpointing comment about the specific ingredients that enhanced the dining experience can offer valuable insights, a general compliment on the entire dinner introduces additional uncertainty. Questions may arise, such as "Was the entire dinner enjoyable, or just certain parts of it?" or "Which actions should be replicated in future meals?" To mitigate the challenges of CA, some individuals may opt for prepackaged dinners, which provide a predetermined recipe minimizing the need for complex evaluation of uncertain prospects. However, this approach comes at the cost of reduced flexibility and the potential loss of pleasantness from a unique and personalized recipe. Ultimately, people credit based on various state features, including time constraints, available resources, information given, task uncertainty, and culinary expertise. Examining certain experimental solutions offers simplifying assumptions at the expense of real-world validity. In this study, our experimental design taps into real-world decision-making processes that revolve around reward uncertainty, emphasizing a temporal credit assignment problem. This approach provides insights into participants' strategies for overcoming such obstacles in real-world environments.

In Reinforcement Learning (RL), agents need to evaluate the state features responsible for producing specific outcomes, as they seek to approach better-than-expected outcomes and

avoid outcomes that were less-than-expected (Sutton & Barto, 2018). Over time, an agent can learn the associations between different state features, but the environmental contingencies will significantly influence the efficiency of learning relevant state features. Outcome values help us derive which features one should weigh when maximizing reward from their environment. Therefore, an agent seeking reward from a changing environment may be able to modulate the efficiency of the system to adapt to various environmental obstacles. The parameters that make up the system also provide a concise manner of explanation and can be used in comparison between similar computational systems in different degrees of information uncertainty. Various methods in RL have been proposed to handle delayed sequential rewards, leading to a diverse set of predictions, that depend on the free parameters employed (Singh & Sutton, 1996; Watkins & Dayan, 1992). Due to the variety of potential predictions, matching model and environment remains a common theme in overcoming learning problems. These different algorithms are seen as internal computational systems that take in external information as input and transform this into a policy-based action for the agent. To handle changes in uncertainty, agents must make computational adjustments to their model, such as through modulating free parameters or in different attempts to assign credit.

In terms of environmental complexity, agents frequently encounter real world events that happen over multiple stages and consist of chaining stages and their outcomes that occur over long-time horizons. CA becomes more challenging in situations with longer time lags, as the distance of the time horizon increases, thereby increasing the noise and effort to identify the precise source of outcomes. Furthermore, in situations with long time lags, the presence of intervening stages can introduce additional uncertainty and disrupt learning (Kearns & Singh, 2000). Amidst the competing state features, the process of dispersing a single credit signal to

multiple stages can lead to spreading credit too thin, or the inverse, updating multiple signals to a single stage could lead to overweighting (Minsky, 1961). Depending on the RL method employed, one can decrease the number of sampled trials needed to discriminate between randomly- and systematically-related stages. However, human implementations in such real-world contexts remain elusive when considering the degree of model efficiency.

Human Solutions to CA

The relationship between machines and human computational solutions can be described as symbiotic, with each side offering unique contributions. Human solutions can serve as valuable directives for machine implementations, pushing the boundaries of efficiency and generalizability. However, deviations from normative algorithms can also shed light on the nuances and limitations of human cognition (Kahneman & Tversky, 1979; Niv, 2009). In complex environments, humans possess the ability to navigate obstacles with limited data, surpassing the capabilities of machines. Human cognition can handle challenges such as sparse rewards, partially observable states, and long-term consequences, even with limited experience (Daw et al., 2011; Gershman & Daw, 2017). On the other hand, machines possess resource advantages that are beyond human capacity, particularly in handling large amounts of information and making predictions in areas like forecasting, classification, or regression (Alzubi et al., 2018).

While machine models can effectively handle variance when given large degrees of information, humans must assess the veracity of information before dedicating cognitive resources to overcoming uncertainty. Striking the right balance between incorporating information and cost-efficiency is a shared challenge for both humans and machines. Similar to machines, managing cognitive resources becomes more challenging when faced with increased

environmental uncertainty. Humans possess an innate drive for rewards, and thus, learning to navigate complex sequential environments becomes crucial. However, as the complexity of these environments increases, our understanding of effective strategies to navigate uncertainty remains limited (Gershman & Daw, 2017). RL has offered a normative framework for addressing the intricacies of gaining reward in sequential environments, presenting an avenue for comparison with human behavior in overcoming long-time horizon contingencies (Daw et al., 2011; Gershman et al., 2014; Moran et al., 2019; Walsh & Anderson, 2011). RL computational systems entail mechanisms that were taken from animal and human research. These systems combine mechanisms of earlier work in psychology theory pertaining to conditioning, learning, and action selection.

Early Work and RL

One of the earliest and most robust psychological findings revolved around learning the relationship between events. Pavlovian conditioning is the process by which an agent pairs a neutral stimulus with an unconditioned stimulus to produce a conditioned response. Furthermore, the agent can learn the contingencies of the fixed environment, but cannot exert influence over this relationship. Rescorla and Wagner (1972) famously produced a learning rule that was able to quantify the mechanisms involved, thereby explaining a qualitative theory through a quantitative system. The key feature of the learning model involves elements of surprise and expectation, where a neutral and unconditioned pairing provokes surprise if there is a deviation from the agent's expectation. This discovery, now termed a prediction error, has been a key mechanism for RL models as learning can only occur under elements of surprise (Schultz et al., 1997). However, the original model could not account for a sequential order of events, thus limiting the model to only immediate pairings.

The temporal difference (TD) learning model was able to overcome such limitations, which extends the Rescorla-Wagner rule with the expected future reward of each state (Sutton & Barto, 2018). Thus, an agent would simulate the value of the current state and the values of potential future transitions. Both models can be adjusted to incorporate a bidirectional effect between the agent's actions and the environment's feedback. Such a cohesion is known as operant conditioning, the process that shapes an agent's behavior through accruing feedback by interacting with the environment. The feedback provides necessary information so that actions can be optimized over time and facilitate the process of maximizing reward from the environment. One modification to these RL models allow us to glimpse into the valuation of state-action pairs, such as in Q-learning or the SARSA algorithm (Sutton & Barto, 2018; C. J. Watkins & Dayan, 1992). Thus, actions could be tracked based on the agent's current state and the chaining of these states would allow flexible adaptation to changing rewards. Learning would occur in separate systems, one that tracked the chaining between states and another that recorded valuation of state-action pairs. Tolman (1948) famously coined the term cognitive map, which described the internal model that animals would use to learn the spatial relationships in the environment regardless of reinforcement. Although more computationally extensive, the animals could efficiently shift navigation to new reward placements if they expended the resources necessary to learn the environmental contingencies.

The Features of an Environment

Such is the case in any environment, one can create abstractions of the state features like representing distance in terms of feet or miles or represent time delay in seconds or days. Abstractions simplify the environment so that the agent is not overloaded with an abundance of features to consider. Relationships between states can be represented in RL as a transition matrix,

a probabilistic matrix for transitioning to new states, creating abstractions of the environment based on the attributed features of the state space. Research to date has shown how learning the contingencies in a transition matrix, then leads to delayed feedback being properly credited to the necessary state (Gläscher et al., 2010; Lehmann et al., 2019; Moran et al., 2019; Walsh & Anderson, 2011). However, a multistep environment with fixed transitions where the uncertainty is represented in a partially observable reward function has been rarely investigated. Solutions that consider delayed feedback often entail embedding a TD algorithm with an exponentially decaying eligibility trace but can either make use of an explicit transition matrix or not. If the temporal sequence of events contains intervening events that can disrupt the temporal continuity, then an agent may erroneously credit feedback if they do not know the structure of transitions.

Abstracting the environment involves considering various features. Both Pavlovian and operant conditioning can be seen as CA problems, which often yield immediate solutions, where the abstraction of feature space is related to the stimuli that occur alongside feedback. However, ambiguity in causal attribution or association learning arises under longer temporal windows, conditional probability, type of response, proximity distance, and when intervening events come into play (Shanks et al., 1989; Wasserman et al., 1983). To simplify these complex environments, experimental methodologies can be employed, reducing the possibilities for abstracting feature space. Nevertheless, examining the interplay between multiple features, such as those of temporal continuity and disruption, can uncover intriguing findings. For instance, a delayed Pavlovian conditioning task may yield clearer causality attribution when intervening intervals are absent.

The unpredictable nature of intervening intervals and their outcomes can lead to confusion when attempting to both retain and inhibit updates for the stimuli along similar

abstractions. Surprisingly, little attention has been given in the human literature to the impact of intervening events on retention, such as in the N-back task (Kane et al., 2007). Separate tasks that consider the interaction between delayed reward and cognitive mechanisms have shown how excessive cognitive demand can increase preference for immediacy of reward (Aranovich et al., 2016; Szuhany et al., 2018). If participants over deploy cognitive resources, they may be incentivized to favor immediate reward, but individual differences in computational capacity appear to influence these effects. However, a unified task that gives the participants the agency to use additional information at cost of more cognitive resources could provide further insights into the tradeoff between value and time in a realistic decision environment (Solway et al., 2017).

Temporal Paradigms

The relationship between computational modeling and delayed reward has a rich literature and starts with one of the most influential constructs, temporal discounting. Temporal discounting describes value as a decreasing hyperbolic function of time (Bickel et al., 2011; Kirby, 1997; McKerchar et al., 2009). A single free parameter governs the rate of discounting, which increases under excessive cognitive stress (Aranovich et al., 2016), higher ratings of addiction (MacKillop et al., 2011), decreased working memory (Bickel et al., 2011), inability to simulation future states (Ballance et al., 2022; Daniel et al., 2016), and has been proposed as a transdiagnostic marker for impulsivity (Bickel et al., 2012). While most tasks of this nature focus on description, experiential tasks have been found to influence temporal discounting estimates, as the additional complexity introduces unique variance to the supposed concept of impulsivity (Horan et al., 2017; Kvam et al., 2021; Reynolds et al., 2006; Solway et al., 2017). The description-experience gap describes the differing effects of constructs when participants are forced to sample information rather than given explicit information. Furthermore, experiential

measures of temporal discounting appear to happen within short time horizons (McClure et al., 2007; Seinstra et al., 2018) and have neural representations in different substrates compared to immediate reward (Kable & Glimcher, 2007, 2010; McClure et al., 2004).

One experimental paradigm, intertemporal choice, has positioned individuals to tradeoff between value and time through the medium of an explicit reward and an experiential delay (McClure et al., 2004, 2007; McKerchar et al., 2009). Experiencing delay can create concrete costs, such as boredom or discomfort; and opportunity costs, such as losing other sources of reinforcements (Paglieri, 2013). For example, an individual waiting for cooking time might feel boredom or lose other opportunities that could occur outside of the cooking time. However, individuals often engage in independent opportunities of reinforcement, which might represent a more realistic paradigm of trading off between value and time. It is not clear how individuals handle waiting times when other independent intervening events are inserted between an event and the reinforcer. Forgoing value may lead to a more cognitive-efficient strategy; and inversely, eschewing cognitive efficiencies may lead to better long run payoffs. In other words, balancing opportunity costs requires an algorithmic tradeoff between resources and optimality, where prioritizing resources may lead to biased predictions about future rewards in an effort to minimize cognitive effort (Kim et al., 2021; C. J. C. H. Watkins, 1989). How individuals differently adapt their learning strategy may lead to accepting inefficiencies to maximize value. In line with this reasoning, complexity is not always a merited strategy and could lead to a reduction in accuracy when additional information is irrelevant (Glaze et al., 2018). Thus, information is vital for appropriating the correct amount of complexity to an environmental problem.

Experimental Task

Tanaka et al. (2009) presented a task that incorporates temporal CA and the inherent dilemma of the dinner example. In a repeat decision task, immediate and delayed feedback was conjoined, that is, feedback was the summed reward from the current and three-trials back choice. This might be thought of as akin to receiving an entire dinner compliment, which provokes a partially observable reward function, rather than one for each separate course. Their analysis showed that participants eschewed any type of structural approach and favored a simple learning solution of augmenting a TD model with an eligibility trace. The eligibility trace assigns credit based on the temporal sequence of previously experienced states to both systematically and randomly related signals. Over time, randomly-related signals will become more infrequent to systematic signals and result in eventually identifying the true relationship. Against this model, here, we developed a tabulation method that is designed to partition the task based on the time-horizon. This tabulation model only credits systematically-related states, uses a value function that augments time as a dimension, and utilizes this new value function to generate separate prediction errors. The eligibility trace has shown promise in a variety of tasks, but can become a suboptimal approach when considering an experimental task that contains randomly related events in the time horizon (Daw et al., 2011; Gläscher et al., 2010; Lehmann et al., 2019; Walsh & Anderson, 2011). Furthermore, the eligibility trace solution might have been warranted due to Tanaka et al. (2009)'s partially observable feedback presentation and constant stimulus-outcome association. Indeed, the agent would have to learn to dissociate, but over time would fully observe the reward function. Often the case, cues help participants maximize reward over long time horizons rather than only focusing on options that are immediately reinforcing (Gureckis & Love, 2009; Walsh & Anderson, 2014). However, in the absence of cues, participants need to rely on other avenues of information. One such possibility is the reward signal itself (Dayan,

2009).

RPEs appears to have neural signatures in phasic dopamine spikes in the basal ganglia (Kobayashi & Schultz, 2008). Furthermore, immediate and delayed reward signals appear to be dissociated in the striatum and orbitofrontal cortex (OFC), respectfully (Ballard & Knutson, 2009; McClure et al., 2004). While neural signatures regarding immediate reward are only represented in the striatum, immediate and delayed signals are both represented in the OFC. These findings seem to indicate that delayed reward signals are separable from those of immediate and could in theory lead to opposing valuations based on an environment's contingencies. Information can help an agent decipher value in a temporal CA task, but participants must have a willingness to employ greater cognitive resources to reduce the uncertainty of estimates.

The different solutions might be seen under the typical dichotomy between RL algorithms, Model-Free (MF) versus Model-Based (MB) RL. While MF RL is about direct learning from sampling the environment without building a cognitive map, MB RL on the other hand, is about building a cognitive map that will aid in flexibly adapting to the dynamics of the environment. We refrain from utilizing these terms, as our task does not allow a clean distinction between models but does have some relevant features (Daw et al., 2011; Moran et al., 2019). First, MF RL can be advantageous when the underlying state space and structural model is not immediately observable. Second, MB RL is more taxing to implement due to the additional layer of the structure. Third, MF and MB RL will often arrive at similar decisions, such as in spatial navigation, but the flexibility in switching on the fly to new reward states is only afforded to MB RL. The individual dilemma of determining the correct weight mixture of these two strategies can be difficult; indeed, incorporating additional information can afford larger rewards at the cost of computation and efficiency (Hastie et al., 2009; Niv et al., 2015; C. J. C. H. Watkins, 1989).

This work aims to address the problem of CA in sequentially delayed rewards and to characterize the strategies that an agent might implement under different degrees of uncertainty. Environmental obstacles, such as through intervening events, can have direct costs on temporal contiguity and lead to inefficiencies of the proposed computational solution. Many computational models have mechanisms in place for handling uncertainty, such as through modulating free parameters or implementing stochastic policies. One proposed solution is to use an undirected and automated process, such as that of a prepackaged dinner, complimented with one that uses a directed and systematic solution, which could be thought of as a home recipe. Determining the correct mixture weight is difficult and environmental information, such as in available information and degree of uncertainty, could dictate to an agent when to weigh one solution over another. We employ a competing computational model, that combines two strategies that implement CA under different degrees of efficiency. We experimentally manipulate the degree of information uncertainty through giving participants two forms of reward feedback.

Predictions

With a few key modifications, our study has adapted Tanaka's task to incorporate a conjoined, like in their original study, and disjointed feedback, which separates the influence of immediate and delayed reward. We also aim to build a hybrid model that captures both an eligibility trace and tabulation-based solution. This proposal's questions are as follows:

"Will Participants Increase Computational Resources When Presented with Additional Information That Reduces Uncertainty?"

Due to the dominance of the eligibility trace in past research, we seek to understand if additional information will change the agent's CA strategy to a solution that can partition the

environment into task-relevant time horizon. Previously, the task's structure may have prompted the usage of an eligibility trace due to the uncertainty in disjointing the feedback. We hypothesize that the weight for tabular model will be higher than the eligibility model for the disjoint condition but decreased for the joint condition.

"Is Temporal Discounting Different in a Repeat Learning Task with Intervening Events?"

We seek to understand how temporal discounting rates modulate during learning. However, due to not having a reliable measure of discounting, we will have to look at rates of choosing stimuli with immediate versus delayed rewards for the two experimental conditions. We hypothesize that the disjoint condition will contain a greater frequency for choosing delayed than in the conjoined condition.

"Can CA Strategy Use Predict Self-Reported Alcohol Use and Impulsivity?"

Ultimately, we hope to use our computational findings to gain a more sophisticated understanding of transdiagnostic measures in real-world experiments. Many experiments have separated mechanisms, such as inhibition, working memory capacity, or decision-making quality, but a unified task oftentimes is unfeasible due to limitations on interpretability. We predict that our parameters of interest will be able to predict scores on self-reported measurements of mental health outcomes.

Methods

Participants

163 participants were recruited from Prolific (an online database service) for an hour and a half long study over two sessions. Prolific inclusion criteria included: fluency in English, ages over 18, and no color blindness. Sessions were broken apart by two days to one week, but participants were allowed to complete the second session within that flexible interval. They were

compensated a total of $30 for participating with a bonus dependent on their proportion of selecting the higher valued stimuli. Due to the possibility of external aid, participants were given instructions to not use any additional help and given an end-questionnaire asking if they had used external aid. Although all participants were paid, 13 participants were dropped for not completing the second stage, six were removed for admitting to using external aids, and two were dropped for duplicate stages. The resulting sample contains 142 total subjects (81 males, 60 females, 1 prefer not to say). Ages ranged from 18 to 63 (M_{age} = 26, SD_{age} = 6.57) with employment statuses (50 full-time, 34 unemployed, 23 other, 22 part-time, 6 full-time nonpaid workers, and 7 missing). Most participants (n = 90) were from Europe (33 Portugal, 30 Poland, 6 Italy, 5 Hungary, 5 United Kingdom, 4 Greece, 4 Spain, and 3 other). The other subjects were predominately spread across North America (n = 22) and South Africa (n = 22); and lastly, a total of eight other subjects in the Middle East and Asia. Although all participants were fluent in English, first languages were predominately Portuguese (n = 34), Polish (n = 31), Spanish (n = 25), English (n = 10), Other (n = 22), and missing (n = 20).

Materials

The CA task was adapted from (Tanaka et al., 2009), and built with PsychoPy3 (Pierce et al., 2022). Two questionnaires were also given, the Alcohol Use Disorder Identification Test (AUDIT) and Barratt Impulsivity Scale (BIS-11) (Barratt, 1983; Bohn et al., 1995). The AUDIT contains 10-items about the consumption of alcohol, such as “How often do you have a drink containing alcohol?”. The BIS-11 is a 4-point Likert scale with subcategories for attention, cognitive instability, motor, perseverance, self-control, and cognitive complexity. At the end of all materials, an exit survey was administered with questions assessing if they used external aids, the difficulty of the task ($M_{difficulty}$ = 60.56, $SD_{difficulty}$ = 25.64), and two open-questions on

noticing anything particular or the way they had learned the values.

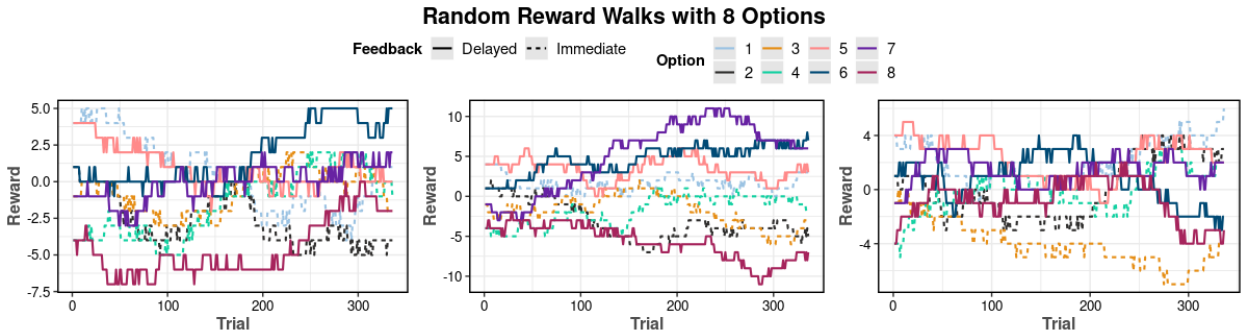
CA Task.

Subjects were given explicit instructions, leading questions, and diagrams of the task structure. The leading questions were aimed at helping the participants understand and therefore were intentionally thought-provoking. In sum, participants answered four instruction questions to help solidify their understanding ($M_{correct}$ = 56% $SD_{correct}$ = 19%). Upon answering one of the four instruction questions incorrectly, further instruction was provided. No subjects were excluded based on instruction questions. At the end of the instructions, the final question asked if they felt they understood the task. Among the subjects, 87% reported feeling comfortable and 13% feeling slight to moderate confusion. The specific object reward and whether they had a delay had to be learned throughout the task. Participants were told in the instructions that delayed objects always had a fixed delay of two trials ahead.

On each trial, participants were instructed to use the mouse to click one of the two objects displayed. They had a total of 15 seconds to make a choice, otherwise no selection was made. There was a total of 8 objects (4 associated with immediate and 4 with delayed feedback) with 336 trials presenting every unique pair of the 8 objects 12 repeated times. Sixteen images were randomly assigned to one of the unique objects across the two sessions, resulting in sixteen different object stimuli. Upon selection, a green box surrounded the selected object, and the participant went to the feedback stage. The feedback was dependent on both the reward (starting rewards: -4, -1, 1, 4) and a fixed delay (0, 2). The reward changed over time with three fixed gaussian random walks, $N(0, .25)$, which was then later rounded to a nearest integer (see Figure 1). The random walk conditions were chosen due to the limitations of the online software and an error on the deployment of those conditions (originally 4 conditions). Depending on choice from

previous trials, participants could receive none, immediate, delayed, or a conjoined immediate and delayed reward.

Figure 1. Three example fixed random walk reward value patterns for 8 stimuli (color) (color) across trial time. Each stimulus was associated with either an immediate (solid lines) or a delayed (dashed lines) reward, and with one of the following starting reward value: 4, 1, -1, -4, then slowly drifted throughout the task based on the random walks. Note that the participant would get a random sequence of unique pairs (28 total) in one of the random walk conditions.



Conditions

Subjects started in either the disjoint (n = 74) or conjoined (n = 70) condition, which modulated the feedback on the conjoined rewards. The conjoined condition gave participants both rewards conjoined together, such as receiving a delayed '2' reward from two trials and an immediate '5' reward from the current trial, which would then be displayed as 7. On the other hand, the disjoint condition gave participants a dissociable reward. The feedback displayed two boxes titled 'immediate reward' and 'delayed reward'; and consequently, did not sum the reward together (see Figure 2).

Procedure

Each participant filled out a consent form that described the type of task they would receive compensation for. After each participant completed the CA task, they took the AUDIT,

BIS-11, and exit survey. Time taken, in minutes, for the CA task (Disjoint: $M_{end} = 25.27$, $SD_{end} = 8.77$, Conjoined: $M_{end} = 25.32$, $SD_{end} = 10.07$) and surveys ($M_{end} = 7.25$, $SD_{end} = 3.85$) were reasonable. Response rate for the CA task was very high ($M_{response} = 99.5\%$). Per trial time was calculated after removing non-answered trials, for an average trial time of 1.59 seconds ($SD_{trialtime} = 1.37$). After completing both sessions, participants were given their bonus. Those who did not complete the second session were paid for the first session but did not receive the bonus payment.

Analysis

Additional variables were produced from the experimental properties imbued in the 8 options. Reward difference calculated the difference in total reward between two options at the current trial. Current valence type was a categorical variable that depicted the valence between both options with “positive”, “mixed”, and “negative” conditions; for example, if one option had a gain and the other a loss then it would be branded as “mixed” regardless of immediate or delayed contingencies. Similarly, current choice type, another categorical variable, reflected if both, neither, or a blended delayed contingency was given on the current trial with “immediate”, “delayed”, and “mixed”, respectfully. Finally, two variables were created for the chosen reward one and two trials back for delayed rewards only. This was to help determine if the delayed option influenced current choice. All continuous variables were standardized ($\mu = 0$, $\sigma = 1$), and categorical predictors were set to dummy codes.

Logistic Regressions

Analysis will start with a more agnostic approach, using a cross-classified logistic regression to characterize the drivers of behavior, such as random walk reward, temporal delay, and condition. All trial data is used on a cross-classified logistic regression with an intercept

variance for both subjects and unique combination pair, along with a slope variance for trial number. To achieve convergence, a derivative-free optimizer that allows bounding through quadratic approximation (bobyqa) will be employed. The data will be subset to reflect only unequal reward options, so that our predicted variable, optimality, can be computed on trial level data. Optimality is defined as trial-level choice of the greater cumulative reward. Our predictors will include condition, stage, reward difference, log-transformed trial time, chosen reward one and two trials back, current valence type, current choice type, and trial number. For interpretation, coefficients will be exponentiated to produce odds ratios, and can be used as a means of quantifying effects sizes, given that the variables are on similar scales.

Next, we will be looking into the signals of delay learning through simulating choice rules from mapping our different reinforcement learning models to behavioral signals in the estimates of our regression output. The *Time-Forward Logistic Regression* considers:

$$\text{Data} \subseteq \text{choice type} = \text{'mixed'}\ (\text{IvD} \mid \text{DvI})$$

$$\text{Chosen\_Delay} \sim \text{Condition}*(I_t + I_{t+1} + I_{t+2} + D_t + D_{t+1} + \mathbf{D_{t+2}}) \quad (1)$$

$$+ (I_t + I_{t+1} + I_{t+2} + D_t + D_{t+1} + D_{t+2} | \text{sub})$$

Delayed (D) reward trials (t+2) would be related to choosing the delayed option when against an immediate in a mixed choice type, which would then correspond to a coefficient in each of the models, eligibility and tabular. We fit the models to account for the full structure of the task in terms of time-led reward but removed the condition interaction for minimal variance contribution. The R packages 'lme4' and 'nlme' package will be used to model all fixed and random effects. The package 'afex' allows for parameterized bootstrapping estimates, resulting in more stable significance tests. Afterwards, we will consider this in relation to different reinforcement learning strategies.

Model-Based

After, we will test a set of multilevel models to dissociate the comparison effects of the two competing RL strategies, explained below, one with less complexity (eligibility trace) and the other with more (tabular). Optimizing these parameters on participant data produces five fitted parameters, weighted model strategy at decision time (β-eligibility and -tabular), model decay of credit trace for past actions (λ-eligibility and -tabular), and a shared learning rate for immediate actions (α). Analysis will be two-fold, in one we will aggregate the effects of the parameters across the two stages and the other using the individual-level decision trials to decipher participants approach to the task. A multilevel regression with a random intercept for each subject level parameter will be modeled with the fixed effects of condition and stage for each parameter. Covariates of age, sex, random walk condition, and employment status will be used as fixed effects in all mixed models (see equation 7). Each parameter will test a full model with an interaction, between condition and stage, against a partial model without an interaction. A Chi-Square (χ^2) will determine if the interaction fit better explains the data and determines if the effects should be investigated further. If the interaction model is significant, then conditional effects will be probed across condition and stage. Interclass correlations (ICC) were calculated on each parameter to quantify the dissociation of variance, whether between- or within-subjects. Furthermore, due to the paired nature of the data, compound symmetry will be used to identify negative correlations between the participant pairs and can be interpreted as a standardized ICC.

$$Parameter\ condition * stage + covariates + (1|combination) + (1 + trials|sub) \quad (2)$$

Reinforcement Learning Models

Eligibility Trace

The eligibility trace model use the Rescorla-Wagner learning rule (δ) to calculate the

difference between reward and expected value (Rescorla & Wagner, 1972).

$$\delta_t = r_t(a) - v_t(a) \tag{3}$$

At the current time point (t), this calculates the difference between the actual reward (r) and estimated value (v). Actions (a) are then updated with a replacing eligibility trace, such that the unchosen actions are discounted.

$$et_t(a_i) = \begin{cases} 1 & if a_i = a_t \\ \lambda_{et} et_{t-1}(a_i) & if a_i \neq a_t \end{cases} \tag{4}$$

The replacing eligibility trace (et) updates all options (i) using a free decay rate (λ_{et}), bounded between 0 and 1, for each action. The current selection is updated with a replacing eligibility trace (1), so that the current selection has no decay (Singh & Sutton, 1996). The value function is then updated for each action.

$$v_{t+1}(a_i) \leftarrow v_t(a_i) + \alpha \delta_t et_t(a_i) \tag{5}$$

The learning rate (α) is a free parameter that determines the magnitude of the update from the RPE and eligibility trace. Thus, the temporal sequence is highly meaningful for the valuation of past actions.

Tabular

The tabular based method has an explicit representation of the temporal sequence (Sutton & Barto, 2018).

$$\delta_t(d) = r_t(a) - Q_t(d, a) \tag{6}$$

Two RPEs are calculated, one for the immediate reward and the other for two-trials previous choice based on the represented delay (d). The Q-learning function considers both the delay and the action chosen.

$$Q_{t+1}(d, a) \leftarrow Q_t(d, a) + \alpha \lambda_{tab}(d = 2) \delta_t(d) \tag{7}$$

Both RPEs are used to update the immediate choice and the choice two trials ago. The learning

rate (α) is shared between both models, as differences were minimal, and interpretation would be reflected in the other parameters. However, two-different free decay rates (lambdas) λ_{tab} and λ_{et} were used in the two separate models. Note that only λ_{tab} is used when we consider the delayed feedback (d = 2), whereas, immediate is not discounted at all. Because the instructions were explicit about not crediting the action chosen one trial ago, the tabular model skips updates for the one trial delay. On the other hand, the eligibility trace will bleed credit into the randomly related objects dependent on the temporal sequence in which they were experienced.

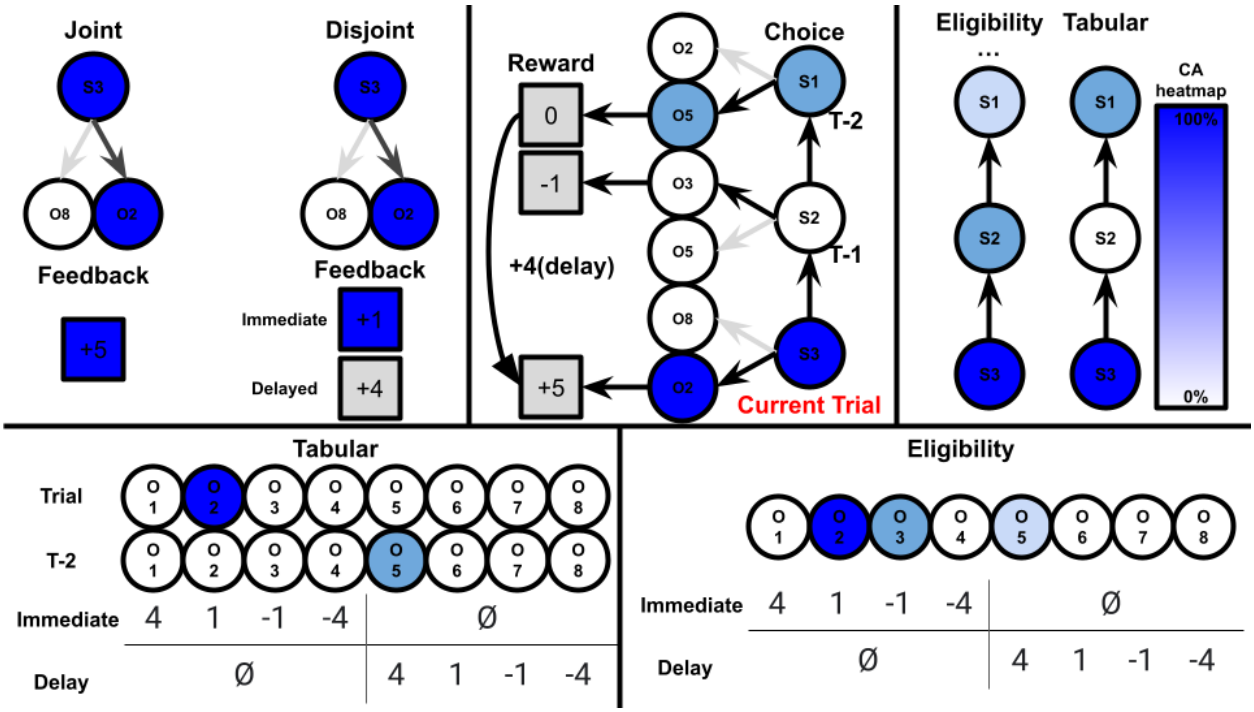
Both models, eligibility (et) and tabular (tab), are placed into a SoftMax function. The hybrid model infuses these two strategies at decision time through two free parameters which reflect a hybrid model that weighs the strategy contribution at decision time, referred to as strategy weight (β). These betas are used to weigh the value function calculated for each of the separate models.

$$\pi_t(a) = \frac{exp\,[\beta_{et} v_t(a) + \beta_{tab} \sum Q_t(d,a)]}{\sum_{i=0}^{n} exp\,[\beta_{et} v_t(a_i) + \beta_{tab} \sum Q_t(d,a_i)]} \quad (8)$$

When considered alone, the hybrid model can be reduced to either the eligibility or tabular through a strategy weight of zero.

Figure 2. Top-left, shows the difference in feedback presentation depending on the condition the participant is in, and the outcome used to generate the prediction error for immediate feedback. The triple period indicates that credit can trace farther than the three states. Top-middle, shows two-alternative forced choice trials and the participants selection (darker arrow) in current trial (red), previous, and trial-minus-two. Colors correspond to the tabular model updating the immediate choice (+1) and trial-minus-two (+4), as each generates a prediction error to update the value function. Top-right, shows the temporal sequence of assigning credit (blue heatmap) based on the free parameters of lambda, such that tabular skips

assigning credit to the previous state (S2). Bottom-left, shows the value function for the tabular model, which contains a matrix of objects as columns and delay as rows. Bottom-right, shows the value function of the eligibility trace, which uses a single prediction error to update. Additionally, the initial outcome of the 8 options is given in the bottom rows.

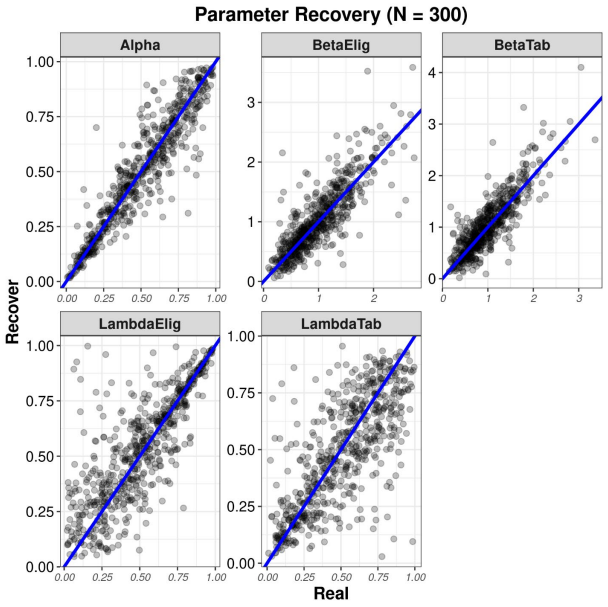


Hybrid Parameter Recovery

The model fitting used the 'deoptim' package in R, which reflects a differential evolutionary algorithm for global optimization (Mullen et al., 2011). Global optimization is used to avoid parameters being trapped in local minimums. Random numbers were chosen using different parameters in a beta and gamma distribution and then used to make choices on our task. Choice data from the simulation was then used to recover the input parameters. The simulation and recovered parameters were reasonable. Correlations between real and recovered parameters for disjoint: learning rate (r = .9), decision weight-tabular (r = .91), decision weight-eligibility (r = .93), decay rate-tabular (r = .79), decay rate-eligibility (r = .86). Conjoined: learning rate (r

= .96), decision weight-tabular (*r* = .83), decision weight-eligibility (*r* = .76), decay rate-tabular (*r* = .68), decay rate-eligibility (*r* = .75). Figure 3 depicts the real and recovered parameter values collapsing across the two conditions.

Figure 3. Parameter recovery (N = 300) for the five parameters collapsed across the two experimental conditions. Elig = Eligibility Model, Tab = Tabular Model.



Models Choice Predictions

We sought to understand how the models might differ regarding the current experimental task. To achieve this, a simulation was set up so that model generations would experience the same type of random trial sequences a human participant would experience. Both eligibility and tabular were built to exhibit behavior only for that model along with a hybrid model that implemented both. The shared learning rate used a randomly generated beta distribution, rbeta(1.5, 1.5). Two different decision weights were generated from the same gamma distribution, rgamma(1, 1). Finally, a set of five fixed decay rates, .3, .5, .75, .85, .9, were used in generation of the simulated choice data. In summary, tabular appears to be more efficient in updating delayed options and does not have a much noise in the valuation of all options.

Figure 4. Value Comparisons. Tracks the value calculations for the models, eligibility and tabular, along with the actual objective rewards for the 8 options. Titles display the value as a function of decay rate (lambda) for each style. Subtitles indicate if the option was immediate (Delay0) or delayed (Delay2) and contained the initial starting value. Eligibility appears to fluctuate more in the disjoint condition, where the tabular model dissociates signal into two values. Models appear to differ most on those of small sample size, such as in delayed rewards as compared immediate gains. Initially values appear to have a heavy bias in favoring whatever the starting value was. However, all two-forced trial pairs (28 total) were of random sequence, but their reward magnitude followed one of three predetermined reward walk sequences. Models appear to make similar predictions on value, but delayed values have more divergence between models.

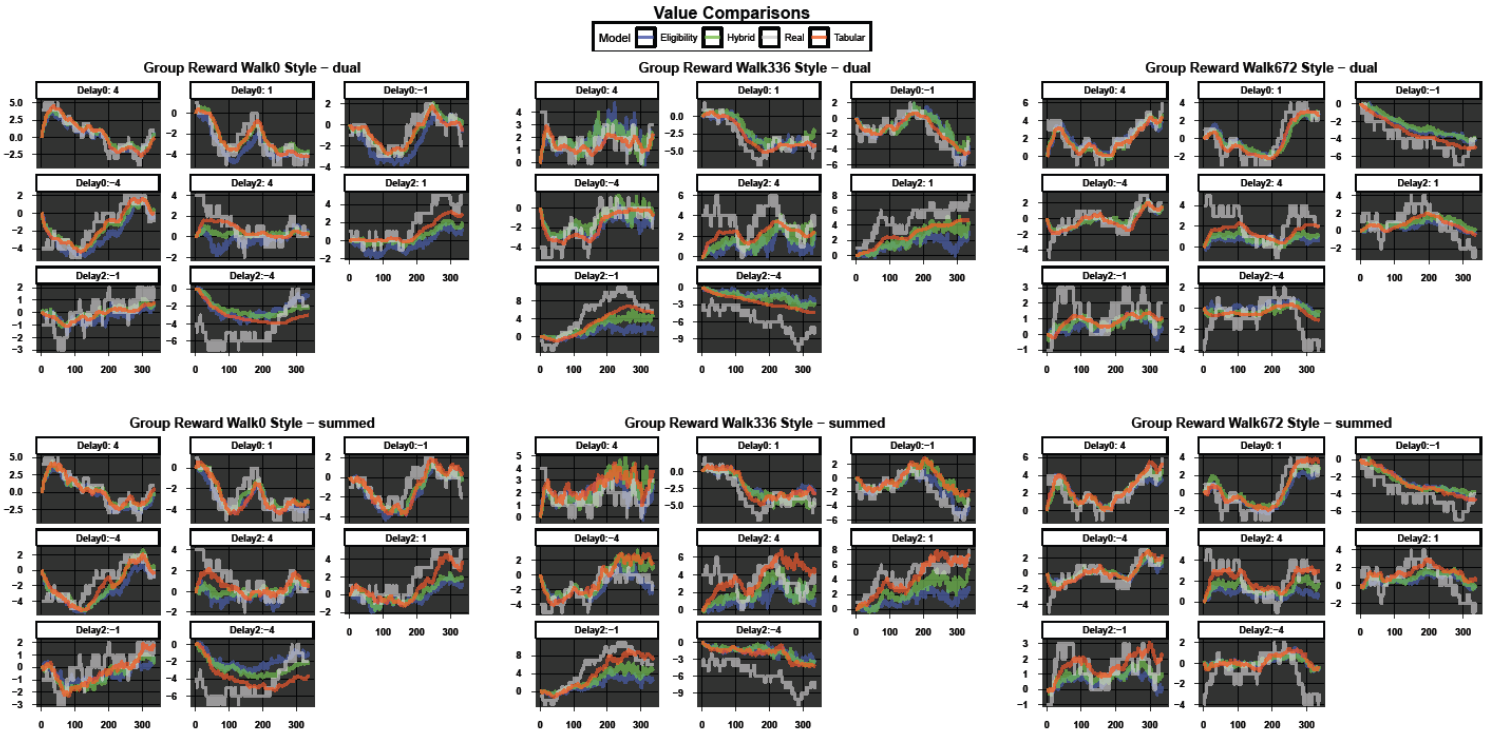
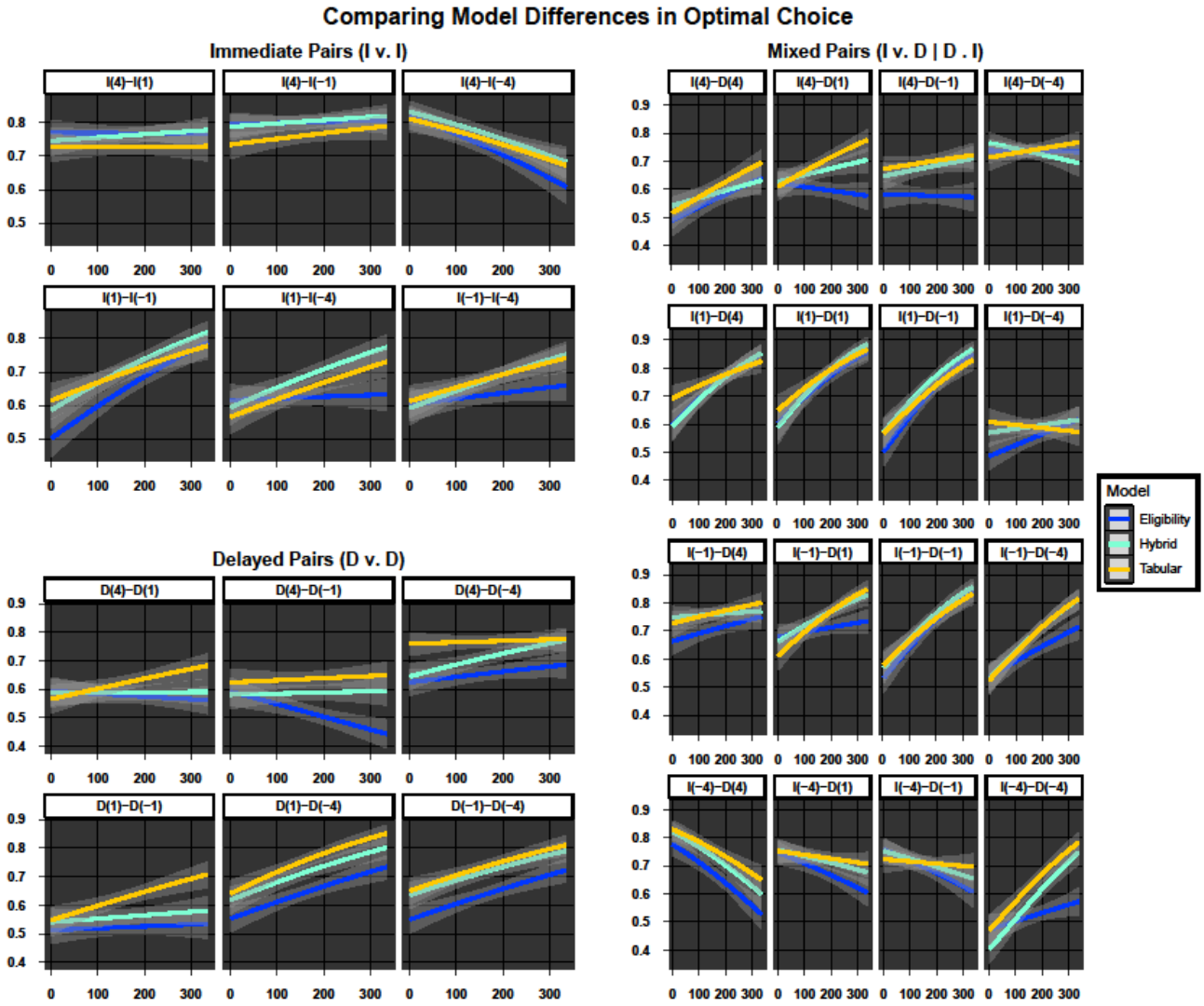


Figure 5. Comparing Model Differences in Optimal Choice. Subsets the dataset (N =

4500 simulated agents) for trials that involve a value difference and a dichotomous predicted optimality which indicates when the agent selects the higher valued object. Data trends show the model accuracy over time with ascending accuracy of eligibility, hybrid, and tabular, where largest differences appear in trials involving less sampling, such as those with delay or negative valence.



Finally, we will employ the use of exploratory statistics to inform interpretation and guide our later predictions about the phenomenon at hand. Such is the case in our study, which

will aggregate trials across subject-level data. T-tests and correlations will be used to see if there is still an effect of condition and stage without the dissociation of variance. Regressions will be used on test trial time, self-reports (AUDIT and BIS-11), and these will not dissociate variance based on subject. Note that we hypothesized that the tabular weight would be higher in disjoint than the eligibility weight and that the decay parameter would be increased in the conjoined condition.

Results

Descriptive

CA task: Proportion of selecting optimality when the reward difference is not equal to 0, Disjoint (M = .68, SD = .11) and Conjoined (M = .64, SD = .09). Proportion of selecting delay when between immediate and delay option, Disjoint (M = .58, SD = .1) and Conjoined (M = .55, SD = .09). Proportion of selecting optimal delay when between immediate and delay option, Disjoint (M = .43, SD = .11) and Conjoined (M = .4, SD = .09).

Self-reports: AUDIT scores (M = 4.3, SD = 3.96), BIS scores (M = 91.33, SD = 14.23) with subscales, attention (M = 9.91, SD = 2.47), cognitive stability (M = 6.29, SD = 1.67), motor (M = 31.58, SD = 5), perseverance (M = 7.08, SD = 1.49), self-control (M = 25.72, SD = 5.22), and cognitive complexity (M = 10.75, SD = 2.37). Correlation between AUDIT and BIS was not significant $r = .07$, $p = .4$.

Reinforcement learning parameters for disjoint condition: alpha (M = .32, SD = .28), beta-eligibility (M = .46, SD = .38), beta-tabular (M = .72, SD = .67), lambda-eligibility (M = .62, SD = .29), and lambda-tabular (M = .54, SD = .26). Reinforcement learning parameters for conjoined condition: alpha (M = .24, SD = .28), beta-eligibility (M = .53, SD = .46), beta-tabular (M = .52, SD = .47), lambda-eligibility (M = .61, SD = .32), and lambda-tabular (M = .46, SD

= .27).

Model-Agnostic

The cross-classified logistic regression considers trial-by-trial data optimal reward differences (see Table 1). Wu et al. (2012) outlines variance components of dichotomous models which calculate the ICC by

$$\hat{\rho} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (\pi^2/3)} \quad (9)$$

where the residual variance is fixed at $\pi^2/3$ and $\hat{\sigma}^2$ is the estimated intercept variance. Furthermore, the fixed effects account for 28.5% of the total variance of both fixed and random effects. The intercept represents participants average odds of picking optimality in the first stage of the disjoint condition, along with positive valence and immediate choice types. Participants on average had the highest odds to select optimality in the disjoint condition of the first stage for mixed valence immediate choice types. The interaction was significant, indicating different levels of optimality pertaining to the different combinations of condition and stage. Specifically, moving from disjoint-to-conjoined in the first stage was associated with a significant decrease in odds, OR = .65, z = -6.4, p < .001, 95% CI [.57, .74]. Moving from conjoined-to-disjoint in the second stage was associated with a non-significant effect, p = .56. Moving from the first-to-second stage in the disjoint condition was associated with a significant decrease in odds, OR = .74, z = -4.46, p < .001, 95% CI [.65, .85]. Moving from the second-to-first stage in the conjoined condition was associated with a significant decrease in odds, OR = .84, z = -2.51, p = .01, 95% CI [.74, .96]. The results dictate that trial types of higher reward differences were associated with greater chance to choose the optimal option. Longer choice times were indicative of less likelihood in choosing an optimal option. Additionally, choice types that were delayed seemed more difficult to choose optimally as compared to mixed and then immediate. Valence

was also illustrative of optimal choice in descending order of mixed, negative, and positive. Finally, the reward received one and two trials ago appeared to have a slight effect on current choice.

Table 1

Cross-Classified Logistic Mixed Effects Regression Predicting Optimal Choice

		Chosen Optimal	
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	2.78	2.29 – 3.37	**<0.001**
RewardStyle [conjoined]	0.65	0.57 – 0.74	**<0.001**
Experiment Stage [2]	0.74	0.65 – 0.85	**<0.001**
Rewards Difference	1.27	1.24 – 1.30	**<0.001**
Log(Trial Time)	0.89	0.88 – 0.91	**<0.001**
Current ValenceType [mixed]	1.31	1.25 – 1.36	**<0.001**
Current ValenceType [negative]	1.12	1.06 – 1.19	**<0.001**
Current ChoiceType [mixed]	0.79	0.65 – 0.96	**0.016**
Current ChoiceType [delayed]	0.63	0.50 – 0.79	**<0.001**
Trial	1.03	1.00 – 1.06	0.091
*RewardStyle [conjoined] * Experiment Stage [2]*	1.60	1.24 – 2.07	**<0.001**
Random Effects			
σ^2	3.29		
$\tau_{00\ ProlificID}$	0.18		
$\tau_{00\ Current.Combination}$	0.04		
$\tau_{11\ ProlificID.Trial\ s}$	0.03		
$\rho_{01\ ProlificID}$	0.48		
ICC	0.07		
$N_{Current.Combination}$	28		
$N_{ProlificID}$	142		
Observations	85978		
Marginal R^2 / Conditional R^2	0.045 / 0.113		

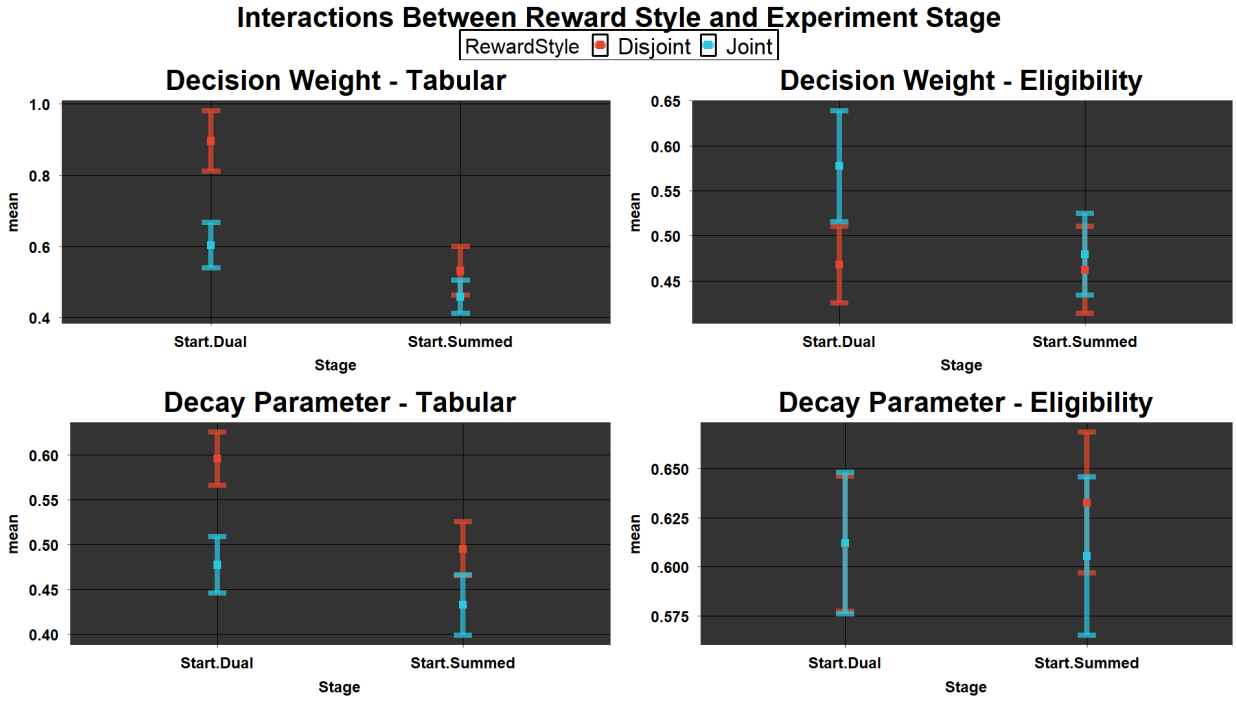
Time-Forward Logistic Regression

Condition had weak effect sizes, but each model appears to be predicted by the correct reward trial type index. Eligibility 95% CI OR (I_t = [.96,.99] I_{t+1}= [.95, .98] I_{t+2} = [.97, 1] D_t =

[.98, 1] D_{t+1} = [.99, 1.01] D_{t+2} = [1.22, 1.25]), Tabular 95% CI OR (I_t = [1,1.09] I_{t+1}= [.97, 1] I_{t+2} = [.96, .99] D_t = [.98, 1] D_{t+1} = [.98, 1.01] D_{t+2} = [1.35, 1.39]), and participant's data 95% CI OR (I_t = [1.29,1.33] I_{t+1}= [.93, .97] I_{t+2} = [.95, .98] D_t = [.99, 1] D_{t+1} = [.99, 1.01] D_{t+2} = [1.17, 1.2]). To note, effect sizes show that participants appear to be using information about the immediate reward to select the delay option, along with some weak effects in immediate one trial back.

Reinforcement Learning Model Parameters

Figure 5. RL parameters means and 95% confidence intervals partitioned across condition and stage. Stage represents if the participant started in the disjoint condition (Start.Disjoint) or ended in the disjoint condition (Start.Conjoined).



Each RL parameter is fit into a mixed effects model with a random intercept for each subject to control for within-subject effects. For β-tabular, an unconditional model allowed an estimate of the ICC, accounting for 27.08% of the total variation between subjects. The best fitting model consisted of an interaction with covariates, showing a significant change in deviance for the

interaction, $\chi^2(1) = 8.15$, $p = .004$. The addition of the covariates also showed a significant change in deviance compared to a non-covariate interaction model, $\chi^2(11) = 29.86$, $p = .002$. Collapsing across the covariate variables' groups showed that each covariate added a significant amount of variance, except random walk, $p = .43$ and sex, $p = .055$. Analysis continued using all covariates regardless of significant effects. Furthermore, both conditional and interaction effects of condition and stage were significant across all competing models. The interaction was significant, $\beta = .41$, $t(130.7) = 2.77$, $p = .006$, 95% CI [.13, .7]. The conditional effect of condition from disjoint -to-sum in the first stage was significant, $\beta = -.39$, $t(250.78) = -4.05$, $p < .001$, 95% CI [-.57, -.2]. The effect of condition from sum-to- disjoint in the second stage was not significant, $\beta = -.03$, $t(249.82) = -.31$, $p = .76$, 95% CI [-.21, .15]. The effect of stage from 1-to-2 in the disjoint condition was significant, $\beta = -.32$, $t(250.06) = -3.35$, $p < .001$, 95% CI [-.5, -.14]. Finally, the effect of stage from 2-to-1 in the conjoined condition was not significant, $\beta = -.1$, $t(249.93) = -1.03$, $p = .3$, 95% CI [-.28, .08]. Collapsing the groups in condition, $F(139.23) = 9.31$, $p = .003$ produced a significant effect, but the collapsed stage, $F(138.35) = 3.57$, $p = .06$, did not reach a significant finding.

For β-eligibility, an unconditional model allowed an estimate of the ICC, accounting for 16.36% of the total variation between subjects. The best fitting model consisted of no-interaction with a single covariate, showing no significant change in deviance for the interaction, $\chi^2(1) = 1.51$, $p = .22$. The addition of the covariates also showed a significant change in deviance compared to a non-covariate model, $\chi^2(10) = 26.95$, $p = .005$. Aggregating the covariate variables showed that only the covariate of random walk added a significant amount of variance, $p < .001$. Furthermore, condition and stage effects were not significant across the competing models. The effect of condition was not significant, $p = .29$, nor was stage, $p = .33$.

For λ-tabular, the fit was based on a generalized least squares model with compound symmetry, as the within-subject correlation was negative, ρ = -.04. The best fitting model consisted of an interaction with no covariates, showing a significant change in likelihoods for the interaction, $\lambda_{LR} = 5.58$, $p = .02$. The addition of the covariates showed no significant change in likelihoods compared to a non-covariate model, $\lambda_{LR} = 6.36$, $p = .85$. Aggregating the covariate variables showed that none of the covariates added a significant amount of variance, all $p > .35$. Both conditional and interactional effects were significant across all competing models. The interaction was significant, $\beta = .15$, $t = 2.37$, $p = .02$, 95% CI [.025, .27]. The conditional effect of condition from disjoint -to-sum in the first stage was significant, $\beta = -.16$, $t = -3.69$, $p < .001$, 95% CI [-.25, -.077]. The effect of condition from sum-to- disjoint in the second stage was not significant, $\beta = .02$, $t = .4$, $p = .69$, 95% CI [-.07, .1]. The effect of stage from 1-to-2 in the disjoint condition was significant, $\beta = -.1$, $t = -2.27$, $p = .024$, 95% CI [-.19, -.01]. Finally, the effect of stage from 2-to-1 in the conjoined condition was not significant, $\beta = -.05$, $t = -1.01$, $p = .31$, 95% CI [-.13, .04]. Collapsing across both condition, $F(1) = 8.29$, $p = .004$ and stage, $F(1) = .76$, $p = .38$, shows only the effect of condition on λ-tabular.

For λ-eligibility, an unconditional model allowed an estimate of the ICC, accounting for 10.96% of the total variation between subjects. The best fitting model consisted of a non-interaction with no covariates, showing no significant change in deviance for the interaction, $\chi^2(1) = .04$, $p = .85$. The addition of the covariates also showed no significant change in deviance compared to a non-covariate model, $\chi^2(11) = 9.91$, $p = .54$. Aggregating the covariate variables showed that none of the covariates added a significant amount of variance, all $p > .3$. The main effect of condition was not significant, $p = .7$. The main effect of stage was also not significant, $p = .69$.

For α, an unconditional model allowed an estimate of the ICC, accounting for 6.03% of the total variation between subjects. The best fitting model consisted of a non-interaction with a single covariate of random walk, showing no significant change in deviance for the interaction, $\chi^2(1) = .47$, $p = .49$. The addition of random walk showed a significant change in deviance compared to the non-covariate model, $\chi^2(1) = 8.85$, $p = .002$. The main effect of condition was significant, $\beta = -.07$, $t(140.22) = -2.28$, $p = .03$, 95% CI [-.13, -.008]. The main effect of stage was not significant, $p = .68$.

Exploratory Analysis

Correlations between variables were inspected as well. Table 1 depicts the Pearson correlation coefficients and their associated p-values among the model parameters. Noteworthy, the two lambda values appear to have a very weak association, which gives further justification against a single lambda parameter. BIS and AUDIT did not have evidence of an association, $r(140) = .07$, $p > .05$. Higher values of optimality were positively associated with model specific parameters, where beta-tabular ($r = .69$), lambda-tabular ($r = .36$), beta-eligibility ($r = .18$), lambda-eligibility ($r = .16$), and alpha ($r = -.07$, $p > .05$). Higher averages of time taken between trials appeared to be correlated with higher levels of optimality, $r(282) = .23$, $p < .01$, higher rates of beta-tabular, $r(282) = .27$, $p < .01$, higher rates of lambda-eligibility, $r(282) = .14$, $p = .02$, and lower rates of BIS, $r(282) = -.12$, $p < .05$.

Paired t-tests for condition regarding reinforcement learning parameters were as follows. Alpha, $t(141) = 2.41$, $p = .02$, 95% CI [.01, .14], $d = .2$. beta-eligibility, $t(141) = -1.42$, $p = .16$, 95% CI [-.16, .03], $d = -.11$. beta-tabular, $t(141) = 3.21$, $p = .002$, 95% CI [.07, .3], $d = .27$. lambda-eligibility, $t(141) = .37$, $p = .71$, 95% CI [-.05, .08], $d = .03$, and lambda-tabular, $t(141) = .09$, $p = .005$, 95% CI [.03, .15], $d = .24$. These t-tests indicate an increase in tabular

parameters when moving from conjoined to disjoint condition.

Paired t-tests for experiment stage regarding reinforcement learning parameters were as follows. Alpha, $t(141) = -.26$, $p = .8$, 95% CI [-.07, .06], $d = -.02$, beta-eligibility, $t(141) = -1.06$, $p = .29$, 95% CI [-.14, .04], $d = -.09$, beta-tabular, $t(141) = 1.97$, $p = .05$, 95% CI [0, .24], $d = .17$, lambda-eligibility, $t(141) = -.38$, $p = .7$, 95% CI [-.08, .05], $d = -.03$, and lambda-tabular $t(141) = .97$, $p = .33$, 95% CI [-.03, .1], $d = .08$. Beta-tabular appears to be the only parameter effected from the stage, but each parameter might contain an interaction with the first stage displaying higher effect sizes.

Exploratory linear regressions were also run on optimal chosen, time taken, AUDIT, and BIS as predicted variables. Data was subset to only consist of trials where there was a reward difference between the two options. Overall model with all reinforcement learning parameters, interaction between style and stage, and the random walk conditions as a covariates showed a significant model fit, $R^2 = .75$, $F(9, 274) = 91.52$, $p < .001$. Significant model predictors were beta-tabular, $b = .13$, $SE = .005$, $t(274) = 22.64$, $p < .001$, 95% CI [.12, .14], $\omega_p^2 = .69$. Beta-eligibility, $b = .09$, $SE = .008$, $t(274) = 10.35$, $p < .001$, 95% CI [.07, .12], $\omega_p^2 = .29$. Lambda-tabular, $b = .09$, $SE = .01$, $t(274) = 7.53$, 95% CI [.07, .12], $\omega_p^2 = .18$. The model shows a quick sanity check for the two separate models' ability to select the higher valued option.

Log-transformed time taken was also used in all reinforcement learning parameters, interaction between style and stage, and random walk conditions as a covariate. The overall model was significant, $R^2 = .12$, $F(9, 274) = 4.23$, $p < .001$. The only significant predictor was beta-tabular, $b = .22$, $SE = .05$, $t(274) = 4.71$, $p < .001$, 95% CI [.13, .31], $\omega_p^2 = .07$. Along with one marginal predictor alpha, $b = -.18$, $SE = .1$, $t(274) = -1.88$, $p = .06$, 95% CI [-.38, .008], $\omega_p^2 = .008$. However, due to the mouse-click nature of the experiment, trial-time taken should be

interpreted with caution. Overall AUDIT model was not significant, $F(9, 274) = .38$, $p = .94$. Along with BIS model showing no overall significant effect, $F(9, 264) = .7$, $p = .71$.

Discussion

The classic dichotomy between automaticity or undirected and systematic or directed modes of processing poses certain boundaries when considering available information (Wason & Evans, 1974). The task from Tanaka et al. (2009) allowed us to pose a form of this dichotomy in a repeated decisions task with intervening events that would stress a temporal CA problem. To this, our competing RL framework combined two CA strategies, eligibility trace and tabulation into a single model. The eligibility trace model which was held to be a less efficient version was the dominant model in the original task. However, in our new task design, tabulation was able to transform temporal information into less uncertainty and participants appeared to only increase their decision weight when starting in the disjoint condition. This might be due to the daunting nature of the task in the conjoint condition, which typically has more computational demand and difficulty. The disjoint condition on the other hand had more individuals using a higher tabulation weight at decision time. Behavioral data appears to map correspondingly with our models, but participants appear to be using additional information from the yoked immediate choice to select the delayed choice. Additionally, it isn't exactly clear where the model differences would be most overt, as our value and pair comparisons generally show different degrees of efficiency rather than fundamentally different modalities of behavior.

With different modalities of behavior, Moran et al. (2019) was able to design a CA task that leveraged a sequential design with structural ambiguity to provide clean dissociations between model-free and model-based CA. Their design incorporates the structural CA problem, as their source of uncertainty arises from the influence of previously learned paired relationships.

Some pairs share a common object, and only upon revealing the unique object can the state uncertainty regarding the structural assignment be resolved. In contrast, Tanaka et al.'s (2009) design incorporates the temporal CA problem, as their source of uncertainty is found in the timing of rewards. Determining whether a reward might be immediate or delayed challenges the participant to track temporal information to reduce uncertainty. Certainty, there is a large degree of overlap between temporal and structural CA, but we view this distinction as meaningful (Agogino & Tumer, 2004). Another point of difference regarding uncertainty is when the ambiguity is presented, as for Moran the ambiguity is presented at the start of each trial. With the Tanaka design, the ambiguity is presented in the temporal contiguity of choice pairs, as the difficulty of learning is a joint relationship between the random sequence and participant choice. Thus, Tanaka's design cannot make clean distinctions between a model-based and model-free system, as multiple MF and MB processes could be involved during task. One could solve the problem with MB prospection on predicting the reward two-trials ahead or engage MF retrospection to bleed credit to past state-action pairs. Regarding this, Tanaka's task is difficult to solve prospectively as intervening test items continue to disrupt working memory and displace the temporal contiguity between feedback and choice. Uncertainty is also represented in the reward function regarding the separation of feedback, where both sources of reward would be summed together. To note, we informed our participants of a fixed delay timing, but not the specific objects that carried the informed fixed contingency. How participants are solving this task remains a conundrum, where our two-condition system - one with disjoint reward and one conjoined - was developed under the idea that one might increase propensity to engage a prospective system given disjoint feedback.

Additional reward information led participants to weigh up a model that incorporated

temporal information to reduce uncertainty despite increased computational demands. The implementation of the eligibility trace model seems to exhibit little dependence on specific experimental contingencies, indicating its potential as a strong baseline across diverse environments, and highlighting its suitability for generalization (Lehmann et al., 2019; Moran et al., 2019; Tanaka et al., 2009; Walsh & Anderson, 2011). The eligibility trace model assigns credit based on the temporal sequence of events; and thereby eschewing the need to build a structural model of the temporal sequence. The current experimental design, with dissociated reward, should allow participants to have full knowledge of the current contingencies and maximize the rate of making optimal decisions. Due to the challenging nature of the task, participants were only able to leverage this additional information when starting in a disjoint condition but in either condition when starting in conjoint. Humans have an ability to engage resources to overcome obstacles in a learning problem, given they have necessary information. However, people can forgo effortful processes and fall victim to the bias of the random sequence in front of them or just choose to not even engage. Immediacy of update might appear as a nontaxing mechanism, but the inhibitory response to withhold credit for other states, such as those of delay, may be more taxing as to having to account for other factors.

The last generation of RL research has focused on differences between model-based and model-free learners (Daw et al., 2011). Recent research has made significant progress in distinguishing between model-free and model-based approaches to CA, revealing a dynamic interplay between retrospective and prospective processes (Deserno et al., 2021; Moran et al., 2019, 2021; Shahar et al., 2019, 2021). This emerging body of work highlights how information updating the current model can lead to retrospectively reassigning credit based on revealed structural information. Thus, the cognitive map that is stored in long-term memory can be

leveraged to correctly update previous cached action values of the model-free system. In the current task, working memory will be the expended resource to learn the relationship pairs as distractor items as distractor items are also test items that must be further incorporated to update relationships. Thus, our task is not purely encode-and-recall, but participants must strategically incorporate reward feedback in the manner they see fit.

The current strategies are restricted to mechanisms involved in RL, such as those of eligibility traces, which trace credit to distant actions. However, taking models from other paradigms, such as those of memory and temporal contiguity might shed light into the diversity of strategies involved (Shanks et al., 1989). These models rely on episodic memory which is often encoded into a buffer system that can incorporate several chunks based on individual differences. Another common model type involves temporal discounting which is used to value different stimuli dependent on everyone's rate of discounting. However, these model types have predominantly been applied to hypothetical future rewards rather than learning value in an experiential format (Bickel et al., 2011; Horan et al., 2017). Collins and Frank (2012) hybrid working memory and RL model, value modifications involving n-back models (Harbison et al., 2011), or a novel implementation of the successor representation (Gershman et al., 2012) could be potential avenues for further discovery of the involved processes.

Extensively, the backpropagating eligibility trace is a solution to solving learning problems of long time horizons (Sutton & Barto, 2018). Stemming from the discounting parameter, the less discounting the more complexity a person is using, as backpropagation of credit is tied to the policy that was instated (Chelu et al., 2020; Kearns & Singh, 2000). The trace of eligibility exponentially decays relative to the distance from experience, but future hypothetical rewards are often hyperbolically discounting (Mazur, 2013). Such a question might

be posed in the descriptive-experiential gap or between the way that retrospective-prospective rewards are handled.

In summary, credit assignment is a nontrivial problem for both humans and machines. The mechanisms implemented in human solutions might provide clarity in novel machine approaches and machine solutions might provide boundaries for human investigation. Despite the complex and demanding nature of the task, participants were able to overcome shifting rewards in a delayed repeated decision task with intervening events. As such, we manipulated the manner of feedback to incentivize participants to increase effortful processes in hope that our strategies might have differential effects. To this, we found evidence of increased weight of a decision strategy that was more efficient conditional on the randomized order participants experienced the task. Thus, evidence would suggest that participants may choose to utilize this information or forgo it dependent on unknown circumstances. We hope that further investigations might look towards new avenues in this experimental design to further understand how credit assignment might be implemented.

Bibliography

Agogino, A. K., & Tumer, K. (2004). *Unifying temporal and structural credit assignment problems*. Autonomous Agents and Multi-Agent Systems Conference.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Aranovich, G. J., McClure, S. M., Fryer, S., & Mathalon, D. H. (2016). The effect of cognitive challenge on delay discounting. *NeuroImage*, *124*, 733–739. https://doi.org/10.1016/j.neuroimage.2015.09.027

Ballance, B. C., Tuen, Y. J., Petrucci, A. S., Orwig, W., Safi, O. K., Madan, C. R., & Palombo, D. J. (2022). Imagining emotional events benefits future-oriented decisions. *Quarterly Journal of Experimental Psychology*, *75*(12), 2332–2348.

Ballard, K., & Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. *NeuroImage*, *45*(1), 143–150. https://doi.org/10.1016/j.neuroimage.2008.11.004

Barratt, E. S. (1983). Impulsivity: Cognitive, behavioral, and psychophysiological correlates. *Biological Bases of Sensation Seeking, Impulsivity, and Anxiety*.

Bickel, W. K., Jarmolowicz, D. P., Mueller, E. T., Koffarnus, M. N., & Gatchalian, K. M. (2012). Excessive discounting of delayed reinforcers as a trans-disease process contributing to addiction and other disease-related vulnerabilities: Emerging evidence. *Pharmacology & Therapeutics*, *134*(3), 287–297. https://doi.org/10.1016/j.pharmthera.2012.02.004

Bickel, W. K., Yi, R., Landes, R. D., Hill, P. F., & Baxter, C. (2011). Remember the Future: Working Memory Training Decreases Delay Discounting Among Stimulant Addicts.

Biological Psychiatry, *69*(3), 260–265. https://doi.org/10.1016/j.biopsych.2010.08.017

Bohn, M. J., Babor, T. F., & Kranzler, H. R. (1995). The Alcohol Use Disorders Identification Test (AUDIT): Validation of a screening instrument for use in medical settings. *Journal of Studies on Alcohol*, *56*(4), 423–432.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Daniel, T. O., Sawyer, A., Dong, Y., Bickel, W. K., & Epstein, L. H. (2016). Remembering versus imagining: When does episodic retrospection and episodic prospection aid decision making? *Journal of Applied Research in Memory and Cognition*, *5*(3), 352–358. https://doi.org/10.1016/j.jarmac.2016.06.005

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Dayan, P. (2009). Prospective and retrospective temporal difference learning. *Network: Computation in Neural Systems*, *20*(1), 32–46. https://doi.org/10.1080/09548980902759086

Deserno, L., Moran, R., Michely, J., Lee, Y., Dayan, P., & Dolan, R. J. (2021). Dopamine enhances model-free credit assignment through boosting of retrospective model-based inference. *eLife*, *10*, e67778. https://doi.org/10.7554/eLife.67778

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*(1),

101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182–194. https://doi.org/10.1037/a0030844

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The Successor Representation and Temporal Context. *Neural Computation*, *24*(6), 1553–1568. https://doi.org/10.1162/NECO_a_00282

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V., & Gold, J. I. (2018). A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, *2*(3), 213–224. https://doi.org/10.1038/s41562-018-0297-4

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293–313. https://doi.org/10.1016/j.cognition.2009.03.013

Harbison, J., Atkins, S. M., & Dougherty, M. R. (2011). *N-back training task performance: Analysis and model*. *33*(33).

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Horan, W. P., Johnson, M. W., & Green, M. F. (2017). Altered experiential, but not hypothetical, delay discounting in schizophrenia. *Journal of Abnormal Psychology*, *126*(3), 301–311.

https://doi.org/10.1037/abn0000249

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. https://doi.org/10.1038/nn2007

Kable, J. W., & Glimcher, P. W. (2010). An "As Soon As Possible" Effect in Human Intertemporal Decision Making: Behavioral Evidence and Neural Mechanisms. *Journal of Neurophysiology*, *103*(5), 2513–2531. https://doi.org/10.1152/jn.00177.2009

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99–127). World Scientific.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Kearns, M. J., & Singh, S. (2000). *Bias-Variance Error Bounds for Temporal Difference Updates.* 142–147.

Kim, D., Jeong, J., & Lee, S. W. (2021). Prefrontal solution to the bias-variance tradeoff during reinforcement learning. *Cell Reports*, *37*(13), 110185. https://doi.org/10.1016/j.celrep.2021.110185

Kirby, K. N. (1997). Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General*, *126*(1), 54.

Kobayashi, S., & Schultz, W. (2008). Influence of Reward Delays on Responses of Dopamine Neurons. *Journal of Neuroscience*, *28*(31), 7837–7846.

https://doi.org/10.1523/JNEUROSCI.1600-08.2008

Kvam, P. D., Romeu, R. J., Turner, B. M., Vassileva, J., & Busemeyer, J. R. (2021). Testing the factor structure underlying behavior using joint cognitive models: Impulsivity in delay discounting and Cambridge gambling tasks. *Psychological Methods*, *26*(1), 18–37. https://doi.org/10.1037/met0000264

Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. *eLife*, *8*, e47463. https://doi.org/10.7554/eLife.47463

MacKillop, J., Amlung, M. T., Few, L. R., Ray, L. A., Sweet, L. H., & Munafò, M. R. (2011). Delayed reward discounting and addictive behavior: A meta-analysis. *Psychopharmacology*, *216*(3), 305–321. https://doi.org/10.1007/s00213-011-2229-0

Mazur, J. E. (2013). An adjusting procedure for studying delayed reinforcement. In *The effect of delay and of intervening events on reinforcement value* (pp. 55–73). Psychology Press.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time Discounting for Primary Rewards. *Journal of Neuroscience*, *27*(21), 5796–5804. https://doi.org/10.1523/JNEUROSCI.4246-06.2007

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science*, *306*(5695), 503–507. https://doi.org/10.1126/science.1100907

McKerchar, T. L., Green, L., Myerson, J., Pickford, T. S., Hill, J. C., & Stout, S. C. (2009). A comparison of four models of delay discounting in humans. *Behavioural Processes*, *81*(2), 256–259. https://doi.org/10.1016/j.beproc.2008.12.017

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30.

https://doi.org/10.1109/JRPROC.1961.287775

Moran, R., Dayan, P., & Dolan, R. J. (2021). Human subjects exploit a cognitive map for credit assignment. *Proceedings of the National Academy of Sciences*, *118*(4), e2016884118. https://doi.org/10.1073/pnas.2016884118

Moran, R., Keramati, M., Dayan, P., & Dolan, R. J. (2019). Retrospective model-based inference guides model-free credit assignment. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-08662-8

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R Package for Global Optimization by Differential Evolution. *Journal of Statistical Software*, *40*(6), 1–26. https://doi.org/10.18637/jss.v040.i06

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Paglieri, F. (2013). The costs of delay: Waiting versus postponing in intertemporal choice. *Journal of the Experimental Analysis of Behavior*, *99*(3), 362–377.

Pierce, J., Hirst, R., & MacAskill, Michael. (2022). *Building Experiments in PsychoPy* (Version 3). https://uk.sagepub.com/en-gb/eur/building-experiments-in-psychopy/book273700

Rescorla, R. A., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II* (pp. 64–99). Appleton-Century-Crofts, New York.

Reynolds, B., Ortengren, A., Richards, J. B., & de Wit, H. (2006). Dimensions of impulsive behavior: Personality and behavioral measures. *Personality and Individual Differences*, *40*(2), 305–315. https://doi.org/10.1016/j.paid.2005.03.024

Schultz, W., Dayan, P., & Montague, P. R. (1997). *A Neural Substrate of Prediction and Reward*.

Seinstra, M. S., Sellitto, M., & Kalenscher, T. (2018). Rate maximization and hyperbolic discounting in human experiential intertemporal decision making. *Behavioral Ecology*, *29*(1), 193–203. https://doi.org/10.1093/beheco/arx145

Shahar, N., Hauser, T. U., Moran, R., Moutoussis, M., NSPN consortium, Principal investigators, Bullmore, E., Dolan, R. J., Goodyer, I., Fonagy, P., Jones, P., NSPN (funded) staff, Moutoussis, M., Hauser, T., Neufeld, S., Romero-Garcia, R., Clair, M. S., Vértes, P., Whitaker, K., … Dolan, R. J. (2021). Assigning the right credit to the wrong action: Compulsivity in the general population is associated with augmented outcome-irrelevant value-based learning. *Translational Psychiatry*, *11*(1), 564. https://doi.org/10.1038/s41398-021-01642-x

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., NSPN Consortium, & Dolan, R. J. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, *116*(32), 15871–15876. https://doi.org/10.1073/pnas.1821647116

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139–159.

Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces.

Machine Learning, *22*(1), 123–158.

Solway, A., Lohrenz, T., & Montague, P. R. (2017). Simulating future value in intertemporal choice. *Scientific Reports*, *7*(1), 43119. https://doi.org/10.1038/srep43119

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Szuhany, K. L., MacKenzie Jr, D., & Otto, M. W. (2018). The impact of depressed mood, working memory capacity, and priming on delay discounting. *Journal of Behavior Therapy and Experimental Psychiatry*, *60*, 37–41.

Tanaka, S. C., Shishida, K., Schweighofer, N., Okamoto, Y., Yamawaki, S., & Doya, K. (2009). Serotonin Affects Association of Aversive Outcomes to Past Actions. *Journal of Neuroscience*, *29*(50), 15669–15674. https://doi.org/10.1523/JNEUROSCI.2799-09.2009

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189.

Walsh, M. M., & Anderson, J. R. (2011). Learning from delayed feedback: Neural responses in temporal credit assignment. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(2), 131–143. https://doi.org/10.3758/s13415-011-0027-0

Walsh, M. M., & Anderson, J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological Bulletin*, *140*(2), 466–486. https://doi.org/10.1037/a0033455

Wason, P. C., & Evans, J. St. B. T. (1974). Dual processes in reasoning? *Cognition*, *3*(2), 141–154. https://doi.org/10.1016/0010-0277(74)90017-1

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation*, *14*(4), 406–432.

https://doi.org/10.1016/0023-9690(83)90025-5

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, *33*(5), 869–880. https://doi.org/10.1016/j.cct.2012.05.004