Information Uncertainty Influences Learning Strategy from Sequentially Delayed Rewards

Sean R. Maulhardt[1], Alec Solway, and Caroline J. Charpentier[2]

Department of Psychology, University of Maryland

Author Note

We have no known conflict of interests to disclose.

Correspondence: [1]smaulhar@umd.edu, [2]ccharpen@umd.edu, Department of Psychology, University of Maryland, College Park, MD 20742.

Abstract

When receiving a reward after a sequence of multiple events, how do we determine which event caused the reward? This problem, known as temporal credit assignment, can be difficult for human solutions given a complex and uncertain environment. It's not clear whether people adjust their strategies to tackle this problem based on the uncertainty of the environment. To address this, we adapted a reward learning task that creates a temporal credit problem combining sequentially delayed rewards, intervening events, and varying uncertainty via the amount of information presented during feedback. Using computational modeling, two learning strategies were developed: eligibility trace whereby previously selected actions are updated as a function of the temporal sequence - and tabular update - whereby only systematically-related past actions (rather than unrelated intervening events) are updated. We hypothesized that reduced uncertainty would correlate with increased use of the tabular strategy, considering the model's capacity to incorporate additional feedback information. Our results supported this hypothesis. Both models effectively learned the task, and choices made by participants (N=142) were best explained by a hybrid model that combined both strategies. However, the tabular model outperformed under conditions of low uncertainty, as evidenced by more accurate predictions of participants' behavior and an increased tabular weight parameter. These findings provide new insights into the mechanisms implemented by humans to solve temporal credit assignment and how they adapt their strategy to the uncertainty and observability of the environment.

Keywords: credit assignment, reinforcement learning, delayed reward, temporal discounting, information uncertainty

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 abstraction of an environment, such as different ingredients for a dinner leading to a tasty meal. 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 mimics real-world decision-making processes in which reward uncertainty stresses a temporal CA 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). Additionally, RL offers a

normative framework for addressing the intricacies of gaining reward in sequential environments, presenting an avenue for comparing human behavior to RL agents that overcome long-time horizon contingencies (Daw et al., 2011; Gershman et al., 2014; Moran et al., 2019; Walsh & Anderson, 2011), and for developing AI agents that mimic human solutions in temporal CA (Nguyen et al., 2023). To handle changes in uncertainty, agents must adjust their manner of assigning credit through the implementation of different internal systems or modulating the components of such systems.

Uncertainty in CA manifests in two primary forms (Agogino & Tumer, 2004; Minsky, 1961; Sutton, 1984). First, there is temporal CA, which deals with the sequential challenge of identifying causal relationships between a series of actions and their outcomes. This contrasts with and leads into structural CA, where the focus is on the parallel challenge of discerning the contributions of various competing components within an internal system. Research to date has mostly focused on structural CA, showing how learning the contingencies or the structure in a transition matrix can then lead to properly assigning delayed feedback to its source (Gläscher et al., 2010; Lehmann et al., 2019; Moran et al., 2019; Walsh & Anderson, 2011). Furthermore, many recent RL experiments have uncovered that individuals learn by combining a strategy of naively adhering to the temporal sequence of rewards known as model-free (MF), and one of adopting an approach that considers the probabilistic transitions along the sequence to reward known as model-based (MB) (Daw et al., 2011; Gershman et al., 2014; Walsh & Anderson, 2011). We refrain from utilizing these terms, as our task does not allow clean distinctions (Daw et al., 2011; Moran et al., 2019). However, several aspects of the MF and MB distinction remain relevant when dealing with temporal CA through delayed rewards and intervening random decisions that might disrupt the systematic relationship (Kearns & Singh, 2000). A MB strategy can be more taxing as an additional layer of complexity is needed to weigh the correct states in a

temporal sequence, whereas a MF strategy might be advantageous when the underlying state space is not immediately observable. Thus, various problems of temporal CA can impact the identifiability of the structural solution and pose a question of the variability in human-implemented strategies.

Solutions that consider delayed feedback often entail embedding a temporal difference (TD) algorithm with an exponentially decaying eligibility trace but can either make use of an explicit transition matrix or not (Daw et al., 2011; Nguyen et al., 2023; Sutton et al., 1999; Walsh & Anderson, 2011). If the temporal sequence of events contains intervening events which disrupt the pairs temporal continuity, then an agent may erroneously credit feedback if they naively follow the sequence of transitions. The unpredictable nature of intervening events and their outcomes can lead to confusion when attempting to both retain and inhibit updates for the stimuli along similar abstractions. Such as in the N-back task, the agent must retain intervening events and prevent inference from past trials when presented with test trials (Kane et al., 2007). Routinely, independent tasks that measure the preference for delayed reward and working-memory demands of delayed events have shown how excessive cognitive demand can increase preference for immediacy of reward (Aranovich et al., 2016; Szuhany et al., 2018). However, a unified task that gives the participants the agency to use additional information to reduce information uncertainty could provide further insights into the tradeoff between value and time in a realistic decision environment (Solway et al., 2017).

Tanaka et al. (2009) introduced a task that creates a temporal CA problem with 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, which embed both systematically and randomly related signals. Over time, randomly-related signals become more infrequent to a systematic signal and result in eventually identifying the true relationship.

In the current study, we aimed 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. We introduced an experimental manipulation on the degree of information uncertainty through two forms of reward feedback: conjoint, like in the study described above, and disjoint, where immediate and delayed outcomes were presented separately. To account for separated knowledge, we developed a tabulation method that is designed to partition the task based on the systematic relationship of the time-horizon. This tabulation model only credits systematically-related timings, uses a value function that augments time as a dimension, and utilizes this new value function to generate separate prediction errors for immediate and delayed rewards. Environmental obstacles, such as through intervening events, can have direct costs on temporal contiguity and lead to inefficiencies of the proposed computational solution (Collins & Frank, 2012; Kearns & Singh, 2000; Nguyen et al., 2023). 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, complemented with one that uses a directed and systematic solution (Gläscher et al., 2010); in our case, an eligibility trace and our proposed tabulation mechanism, respectively. Here, we implemented each strategy (eligibility trace and tabular) as its own computational model, reflecting CA under different degrees of efficiency, and tested a hybrid computational model that combines both strategies.

Predictions

Due to the dominance of the eligibility trace in past research, we seek to understand if additional information that reduces uncertainty will change the agent's CA strategy to a solution that can partition the environment around a task-relevant time horizon. Previously, the task's structure may have prompted the usage of an eligibility trace due to the uncertainty associated with conjoint feedback (Tanaka et al., 2009). Here, we hypothesize that the uncertainty reduction associated with additional information will promote the use of the tabular strategy over eligibility trace, and that we will observe this effect in two ways. First, the tabular model will explain participants' choices better in the disjoint condition, where additional information is provided, while the eligibility model will explain participants' choice better in the conjoint condition. Second, the tabular weight will be higher in the disjoint than conjoint condition, while the eligibility weight will be higher in the conjoint than disjoint condition.

Methods

Participants

163 participants were recruited from Prolific Academic (https://prolific.com) 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 built with PsychoPy3 (Pierce et al., 2022). At the end of all materials, an exit survey was administered with questions assessing if they used external aids, the difficulty of the task on a scale of 0 (easy) to 100 (hard), $M_{difficulty}$ = 60.56, $SD_{difficulty}$ = 25.64, and two open-questions on noticing anything particular or the way they had learned the values.

CA Task

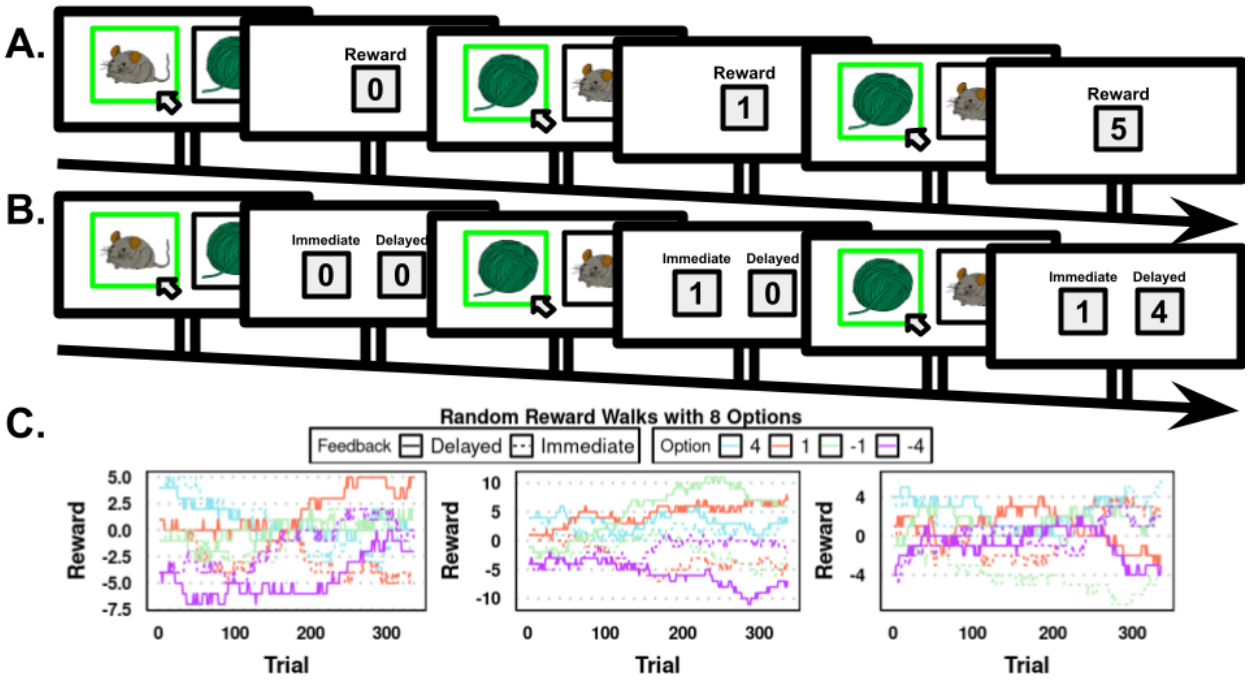
Participants were given explicit instructions, leading questions, and diagrams of the task structure. The leading questions were aimed at helping participants understand and therefore were intentionally thought-provoking. In sum, participants answered four instruction questions to help solidify their understanding ($M_{correct}$ = 56% $SD_{correct}$ = 19%). Upon answering one of the four instruction questions incorrectly, further instruction was provided. No participants 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 participants, 87% reported feeling comfortable and 13% feeling slight to moderate confusion. The specific object reward and whether the object had a delayed reward 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 (Figure 1A-B), 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 without replacement 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 (Figure 1C). The random walk conditions were chosen due to the limitations of the online software and were randomized for each participant (28.17% of participants had the same random walk for both conditions). However, the sequence of decision pairs was totally random. Depending on choice from previous trials, participants could receive none, immediate, delayed, or a conjoint immediate and delayed reward.

Figure 1

Experimental design and random walks



Note. A. Example trial sequence of the conjoint condition. Participants choose between two objects, then receive feedback. Here, a ball of yarn offers an immediate reward of '1', while a mouse provides a delayed reward of '4' after two trials. B. The disjoint condition presented on the same sequence of events but with dissociable feedback. C. This illustrates the three fixed random walk reward value patterns for all eight stimuli across trials. Each stimulus is linked to either an immediate (solid lines) or a delayed (dashed lines) reward. The starting reward values are 4, 1, -1, or -4, and they gradually drift throughout the task. Participants encounter a random sequence of unique pairs (28 in total) within one of the random walk conditions.

Conditions

Participants started in either the disjoint (n = 74) or conjoint (n = 70) condition. The conjoint condition gave participants both rewards conjoint together, such as receiving a delayed '4' reward from two trials back and an immediate '1' reward from the current trial, which would then be displayed as 5 (Figure 1A). 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 (Figure 1B).

Procedure

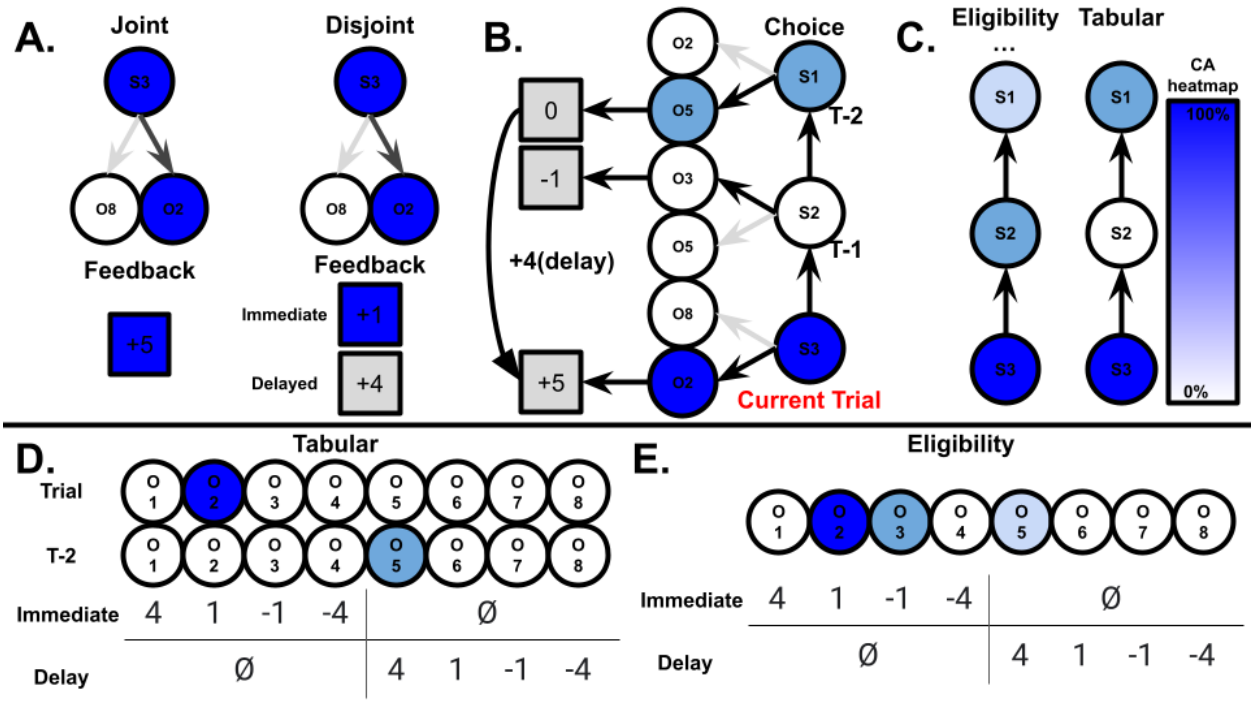
Each participant filled out a consent form that described the type of task they would receive compensation for. After each participant completed the CA task, they took an exit survey. Time taken, in minutes, for the CA task (Disjoint: $M_{end} = 25.27$, $SD_{end} = 8.77$, Conjoint: $M_{end} = 25.32$, $SD_{end} = 10.07$) 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.

Reinforcement Learning Models

To better illustrate the two learning models and their associated value functions, a graphical representation is provided in Figure 2.

Figure 2

Graphical representation of task conditions, learning models, and value functions



Note. A. Differences in feedback presentation based on the participant's condition and the outcome used to generate the prediction error for immediate feedback. B. Example sequence of two-alternative forced choice trials and the participant's selection (darker arrow) in the current trial (red), the previous trial (T-1), and trial-minus-two (T-2). The colors correspond to the tabular model, which updates the immediate choice (+1) and trial-minus-two choice (+4), each generating a prediction error to update the value function. C. Temporal sequence of assigning credit (shown as a blue heatmap). In this model, tabular skips assigning credit to the previous state (S2). The triple period signifies that credit assignment can extend beyond the three depicted states. Note that the extent to which past states are assigned credits in each model depends on the free parameter lambda: for eligibility, higher lambda values mean that credit extend further back in time (less decay), while for tabular, higher lambda values mean less discounting of the trial-minus-two state specifically. D-E. Value functions for the tabular model (D), which involves separate, independent, updates for the immediate and delayed chosen options, and for the eligibility trace (E), which utilizes a single prediction error for updates.

Eligibility Trace

The eligibility trace model (we use the abbreviated version 'elg' in different graphs) uses the Rescorla-Wagner learning rule (δ) to calculate the difference between reward and expected value (Rescorla & Wagner, 1972).

$$\delta_t = r_t(a) - v_t(a) \tag{1}$$

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.

$$et_t(a_i) = \begin{cases} 1 & if\, a_i = a_t \\ \lambda_{elg} et_{t-1}(a_i) & if\, a_i \neq a_t \end{cases} \tag{2}$$

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

$$v_{t+1}(a_i) \leftarrow v_t(a_i) + \alpha_{elg} \delta_t et_t(a_i) \tag{3}$$

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 (abbreviated 'tab') has an explicit representation of the temporal sequence (Sutton & Barto, 2018).

$$\delta_t(d) = r_t(a) - Q_t(d,a) \tag{4}$$

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

$$Q_{t+1}(d,a) \leftarrow \begin{cases} Q_t(d,a) + \alpha_{tab}\lambda_{tab}\delta_t(d) & if d = 2 \\ Q_t(d,a) + \alpha_{tab}\delta_t(d) & if d = 0 \end{cases} \tag{5}$$

Both RPEs are used to update the immediate choice and the choice from two trials ago. Both models have their own parameters used in the two separate models. Note that λ_{tab} is only 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. These decay rate parameters are not symmetrical between models, as the eligibility strategy decays the trace of the temporal sequence and the tabular strategy only downweighs the update for the two-trial back choice.

Decision Rule and Hybrid Model

Both models, eligibility (elg) and tabular (tab), are placed into a SoftMax function. For the independent models, there is a single strategy weight; whereas the hybrid model infuses these two strategies at decision time through two free parameters which reflect a weighing the strategy

contribution at decision time, referred to as strategy weight (β_{elg} and β_{tab}). These betas are used to weigh the value function calculated for each single-strategy model.

$$\pi_t(a) = \frac{exp\left[z(\beta_{elg} v_t(a) + \beta_{tab} \sum Q_t(d,a))\right]}{\sum_{i=0}^{n} exp\left[z(\beta_{elg} v_t(a_i) + \beta_{tab} \sum Q_t(d,a_i))\right]} \quad (6)$$

When considered alone, the hybrid model can be reduced to either the eligibility or tabular through a strategy weight of zero assigned to the other strategy. Finally, we incorporate for each model's value function a z-scored locally weighted smoothing function within 5-trials behind of current trial before SoftMax decision choice. By normalizing the data before feeding into the SoftMax, the hybrid model can account for different scales of the random walks and reduce the noise of sub-models operating on different scaled rewards.

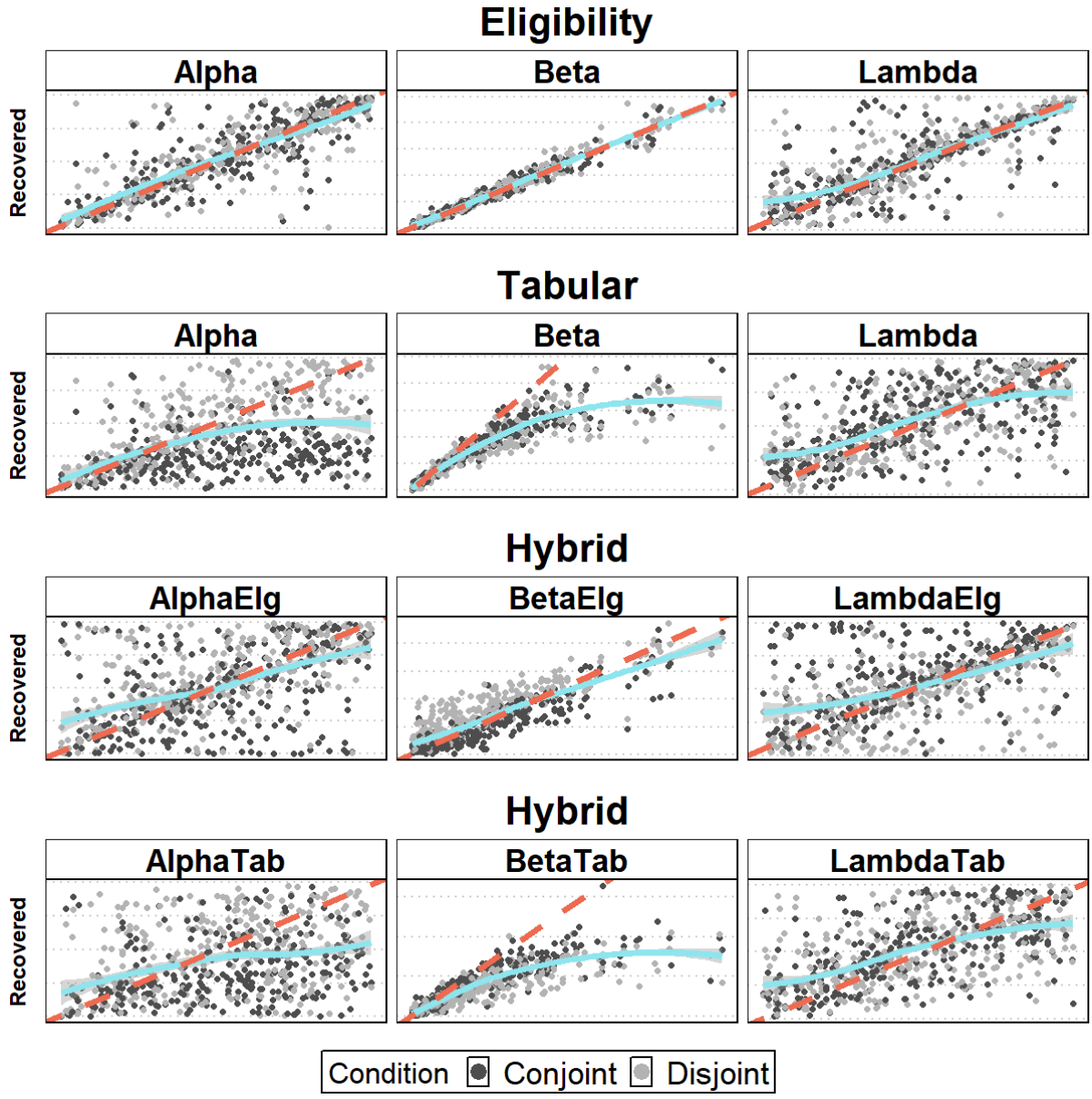
Parameter Recovery

To achieve proficient parameter recovery, we simulated 300 datasets using modest uncertainty to generate each of the 300 parameter sets. For each of model's learning rate and decay rate, a beta distribution was utilized ($\alpha = 1.25$, $\beta = 1.25$), while the decision weights used a gamma distribution ($\alpha = 1.25$, $\beta = 1$). The three generated parameters were held constant between the two independent models and feedback conditions. Additionally, for the hybrid model, these six generated parameters were held constant across both feedback conditions. Each of the generative models returned the simulated agent's choice, value functions, and probabilities. Afterwards, the model was optimized using an evolutionary strategy (R package 'DEoptim') on the simulated agent's choices to identify the correlation between generative and recovered parameters (Mullen et al., 2011). To avoid local minimums, we used a large smoothing value was set for the population size (750 for each model and 1500 for hybrid), a modest number of iterations (60 for each model and 120 for hybrid), and strategy #3 which provided a jitter to the inherited parameters from previous fit iteration. Additionally, priors were

used corresponding to the generated parameters distributions with modest uncertainty and the z-scored locally weighted smoothing function was used. Pearson's correlations between simulated and recovered parameters were then computed and transposed onto a scatterplot with a fitted line (Figure 3). For eligibility-conjoint, alpha (learning rate), $r(299) = .86$, beta (decision weight), $r(299) = .99$, and lambda (decay rate), $r(299) = .81$. For eligibility-disjoint, alpha, $r(299) = .84$, beta, $r(299) = .98$, and lambda, $r(299) = .83$. For tabular-conjoint, alpha, $r(299) = .33$, beta, $r(299) = .92$, and lambda, $r(299) = .61$. For tabular-disjoint, alpha, $r(299) = .64$, beta, $r(299) = .85$, and lambda, $r(299) = .64$. For hybrid-conjoint, alpha-eligibility, $r(299) = .48$, beta-eligibility, $r(299) = .9$, and lambda-eligibility, $r(299) = .5$, alpha-tabular, $r(299) = .33$, beta-tabular, $r(299) = .84$, and lambda-tabular, $r(299) = .63$. For hybrid-disjoint, alpha-eligibility, $r(299) = .59$, beta-eligibility, $r(299) = .89$, and lambda-eligibility, $r(299) = .6$, alpha-tabular, $r(299) = .32$, beta-tabular, $r(299) = .83$, and lambda-tabular, $r(299) = .54$. Noteworthy, beta-tabular appears to be underfitting the trend, along with some under-weighing and over-weighing based on the reward condition, such as in beta-eligibility for the hybrid model (see Figure 3).

Figure 3

Parameter recovery (N = 300) for all models, parameters, and conditions



Note. Parameter recovery was performed by simulating 300 choice datasets from random sets of parameter values, fitting those simulated datasets with the three models (Eligibility, top; Tabular, middle; Hybrid, bottom), and plotting the recovered parameter values as a function of the 'real' parameter values used to generate the data for disjoint (dark gray) and conjoint (light gray) conditions. The closer the loess line (blue-solid) is to the reference line (red-dotted) the better the fit of the generated parameters.

Analyses

To test our hypothesis, we ran the following analyses to understand how our conditions (conjoint vs. disjoint) and stage group order (disjoint-to-conjoint vs. conjoint-to-disjoint) influenced participants' choice and learning behavior. Initially, we performed basic descriptive statistics (Table 1) and regressions on experimental and constructed variables. Next, we aimed to validate the predictions of our models through two multilevel logistic regressions: one focusing on condition and delayed learning (Figure 4), and the other on how our RL models track participant choice behavior across combined conditions (Figure 5). Third, we compared our three models to determine which best fit the data for each participant and reward condition (Table 2). Lastly, we examined our independent and hybrid model parameters via correlations (Table 3) and t-tests (Table 4) and compared them in multilevel models looking at the interaction of reward condition and stage group order (Figure 6).

Posterior Predictive Checks: Model Validation Against Behavioral Data

To validate our eligibility and tabular models and understand their relation to behavioral data and the behavioral pattern they each capture, we defined behavioral measures that show whether participants were tracking delayed rewards correctly. Initially, we expected certain sequence patterns to reveal delayed reward learning through participants' stay or switch behavior, akin to the key behavioral signature of the two-step task (Daw et al., 2011). We selected sequences where participants chose a delayed option, and it reappeared as a potential choice three trials later. If participants are learning the contingencies properly, they should use feedback information two trials into the future to either stay or switch on the following trial, depending on the reward's valence, rather than relying on immediate or one-trial-forward feedback information (Figure 4A). We quantified this signature using a multilevel logistic regression from the 'lme4' package in R:

$$Stay \sim Condition * Time * Reward + (1 + Condition * Time * Reward \mid Subject) \quad (7)$$

We fit three logistic regressions from equation 9: one on the participants' data, and two on the data generated from eligibility and tabular models. Both eligibility and tabular used the participants' best fitting parameters to generate model-independent choice. The significance value of the interaction provides evidence claims on deviation from chance (.5) rather than comparisons between the three models (Figure 4C); nevertheless, the interaction plot with confidence intervals (Figure 4B) informs us of differences between models, as well as similarities and differences between each model predictions and participants' data.

Next, we examined whether participants were tracking the random-walk of rewards and how the models corresponded with our conditions. Similar to Tanaka et al. (2009), we initially focused on illustrating how participants were capable of tracking rewards in specific example pairs (Figure 5A). Following this, we demonstrated our hybrid model's capability to replicate these choice trajectories (Figure 5B). This analysis provides a foundational understanding of individual decision-making processes and the model's performance in isolated instances. However, for a comprehensive understanding of overall trial-by-trial accuracy and decision-making patterns in our two models (tabular and eligibility) under both conditions (disjoint and conjoint), we expanded our analysis. This broader approach involved aggregating data across all random walks and pairs, aiming to identify overarching trends and interactions not apparent in individual examples. We quantified participants' overall propensity of choice by simulating parameters for our two models, using a beta distribution ($\alpha = 1, \beta = 2$) for our learning rate, a gamma distribution ($\alpha = 3, \beta = 1$) for the inverse temperature parameter, and a beta distribution ($\alpha = 2, \beta = 3$) for the decay parameter. This maintained consistent parameters across the two models and kept a high inverse temperature parameter to reduce random noise. The two models then made selections on the same sequences seen by participants during the task. We fit a multilevel logistic regression (equation 10) predicting participants' choice on each trial from the

choice probability predicted by each model separately, both interacting with condition (conjoint and disjoint):

$$Choice \sim Condition * (Elg + Tab) + (1 + Condition * (Elg + Tab) \mid Subject) \quad (8)$$

This regression model allowed us to estimate and compare whether eligibility or tabular predictions were closer to participants' choices and whether that effect varied with the condition (Figure 5C).

Each multilevel model used the 'mixed' function from the 'afex' package, to calculate p-values using the Satterthwaite approximation for logistic regressions and Kenward-Roger approximation otherwise. These methods keep type I error rate from being inflated (Luke, 2017). However, despite the superior performance, the first behavioral check (equation 9) was too computationally taxing to implement this method.

Model Fitting Metrics

After quantifying behavioral signatures from participants' and model data, we quantified our model fits in each of our RL models. We fit each of the three models (eligibility, tabular, and hybrid) to participants' choice using the optimizer from R package 'DEoptim', separately for each condition. We provided four descriptive statistics of model fit per participant and condition (Table 2): the percentage of best-fitting model, the mean and standard deviation of negative loglikelihood, and pseudo-R2. Pseudo-R^2 was calculated by subtracting one from the ratio of the fitted model's negative log-likelihood to that of the null (or random) model for each participant's condition. Note, the current log-likelihoods are penalized for extreme values using priors, for each learning rate and decay parameters a beta distribution was used ($\alpha = 1.25$, $\beta = 1.25$) and for the inverse temperature parameter, a gamma distribution ($\alpha = 1.25$, $\beta = 1$).

Computational Parameters Across Conditions

We first looked at the bivariate correlations between the parameters partitioned on

condition (Table 3). Although these parameters are dependent on the model they are yoked to and violate independency assumptions of traditional statistics, we reasoned that comparing parameters across condition and stage group order (order of conditions) would still provide additional information to corroborate our hypothesis that tabular and eligibility weights should vary with information uncertainty and assess whether this effect was impacted by condition order. Additionally, exploring potential changes in decay and learning rate parameters may provide some insights into the cognitive mechanisms that drive the change in strategy – for example, increased or decrease reliance on a strategy could be accompanied by a modulation in either decay or learning rates. Checking for condition-driven changes in parameter values can help us dissociate between these mechanisms. Thus, we decided to rely on the parameters from our independent models since those were better recovered (Figure 3), although the differences in parameters from the hybrid model were similar. Transitioning from more biased statistics, such as those from a simplified model to a less biased model, like those partitioning variance, provides an ease of interpretation when considering various experimental conditions. T-tests give quick information about the initial decoupling of effects but suffer from inherent simplifying assumptions (Table 4). Multilevel models were used to handle the within-subject variation between the two conditions and provide an estimate for each model parameter across condition (conjoint, disjoint) and stage group order (disjoint-to-conjoint, conjoint-to-disjoint), utilizing a participant-level random intercept.

$$Parameter \sim Condition * Stage + (1 \mid Subject) \quad (9)$$

If one of the parameters, specifically learning rate for tabular, needed a more complex variance and correlation structures for the random effects, such as compound symmetry, the ‘nlme’ package was used instead of ‘lme4’. This can often happen when there is a negative relationship in the parameter between the two conditions that the participant is measured in.

For each of the multilevel models, the first measure of interest was the intraclass correlation (ICC). Typically, the ICC is calculated from an unconditioned model which includes only a fixed intercept and random participant intercepts. This model helps to estimate the variance attributable to differences between participants, thereby assessing the reliability of measurements within the same experimental condition across different subjects. In terms of this experiment, this coefficient would represent the correlation or consistency between the subject's conditions. Next, we compared models with and without an interaction between condition and stage order to test whether adding an interaction improved model fit. To do so, a likelihood ratio test using the 'anova' function in R was used, which follows a chi-square distribution. Last, the final statistic test used the interaction in the 'mixed' function, which should caution the interpretation of main effects.

Transparency and Openness

This study was approved by the University of Maryland College Park IRB (ID: 1155349-32). All code and cleaned data can be accessed to reproduce all statistics, tables, and figures at https://osf.io/rp56b/. This study was not preregistered.

Results

Decisions Favor Optimality in The First Disjoint Stage

Initial behavioral analysis sought to uncover participants' choice behavior from decision metrics, specifically optimal choice, delayed optimal choice, and average outcome received per trial (see Table 1 for means and standard deviations). Optimal choice is defined as the selection of the higher reward object. Delayed optimal choice subsets trials where a participant selected a delay, and the selection was the higher valued object. Average outcome is the selected reward of the choice regardless of delayed contingency. Noteworthy, a participant choosing the optimal option more often than chance (50%) reflects learning performance about the delayed

contingencies and random walk of reward. Optimal choice and delayed optimal choice were placed into a one-sample t-test for each of the four combinations of reward condition and stage (see all means and standard deviations in Table 1), resulting in performance for all conditions being significantly above chance, $\mu = .5$, $p < .001$. To quantify how performance varied across conditions, each variable was then placed into a linear regression with an interaction term between reward condition (disjoint, conjoint) and stage (first, second). Residual and QQ plots were all reasonable and did not merit the use of nonparametric tests. For optimal choice, there was a significant main effect of reward condition, $b = .1$, SE = .018, $t(280) = 5.32$, $p < .001$, $\omega_p^2 = .09$, a significant effect of stage, $b = .04$, SE = .018, $t(280) = 5.18$, $p = .04$, $\omega_p^2 = .01$, and a significant interaction, $b = -.1$, SE = .025, $t(280) = -4.05$, $p < .001$, $\omega_p^2 = .05$. For delayed optimal choice, there was a significant main effect of reward condition, $b = .09$, SE = .018, $t(280) = 4.93$, $p < .001$, $\omega_p^2 = .08$, a significant main effect of stage, $b = .04$, SE = .018, $t(280) = 2.31$, $p = .02$, $\omega_p^2 = .01$, and a significant interaction, $b = -.11$, SE = .026, $t(280) = -4.08$, $p < .001$, $\omega_p^2 = .05$. For average outcome, there was a significant main effect of reward condition, $b = .36$, SE = .15, $t(280) = 2.47$, $p = .01$, $\omega_p^2 = .02$, no effect of stage, $p = .23$, and a significant interaction, $b = -.42$, SE = .21, $t(280) = -2.00$, $p < .05$, $\omega_p^2 = .01$. Consequently, the observed order effect of stage and reward condition appears to have predominantly influenced these behavioral markers, as starting in the disjoint condition gave the highest numbers for all variables (Table 1). Participants performed best starting in the disjoint condition, but they also performed in the conjoint comparably to those who ended in the disjoint condition. Thus, the group that started in disjoint and ended in conjoint performed overall better than the group that started in conjoint and ended in disjoint.

Table 1

Descriptive statistics for behavioral performance

Disjoint-1	Mean	SD	**Disjoint-2**	Mean	SD
Outcome	0.85	3.18	Outcome	0.61	3.25
Optimal	0.7	0.46	Optimal	0.65	0.48
OptiDelay	0.82	0.39	OptiDelay	0.78	0.41
Conjoint-1	Mean	SD	**Conjoint -2**	Mean	SD
Outcome	0.48	3.17	Outcome	0.66	3.12
Optimal	0.61	0.49	Optimal	0.65	0.48
OptiDelay	0.77	0.42	OptiDelay	0.8	0.4

Note. Means and standard deviations (SD) are shown for outcome received (Outcome), proportion of choosing the optimal option (Optimal), proportion of choosing optimally when selecting between an immediate and delayed option (OptiDelay), across the two reward conditions (Disjoint and Conjoint) and current stage (1 and 2). Note that disjoint-1 and conjoint-2 represent the same group of participants (N=74) while conjoint-1 and disjoint-2 represent another group (N=70).

Tabular and Eligibility Best Map onto Disjoint and Conjoint Conditions, Respectively

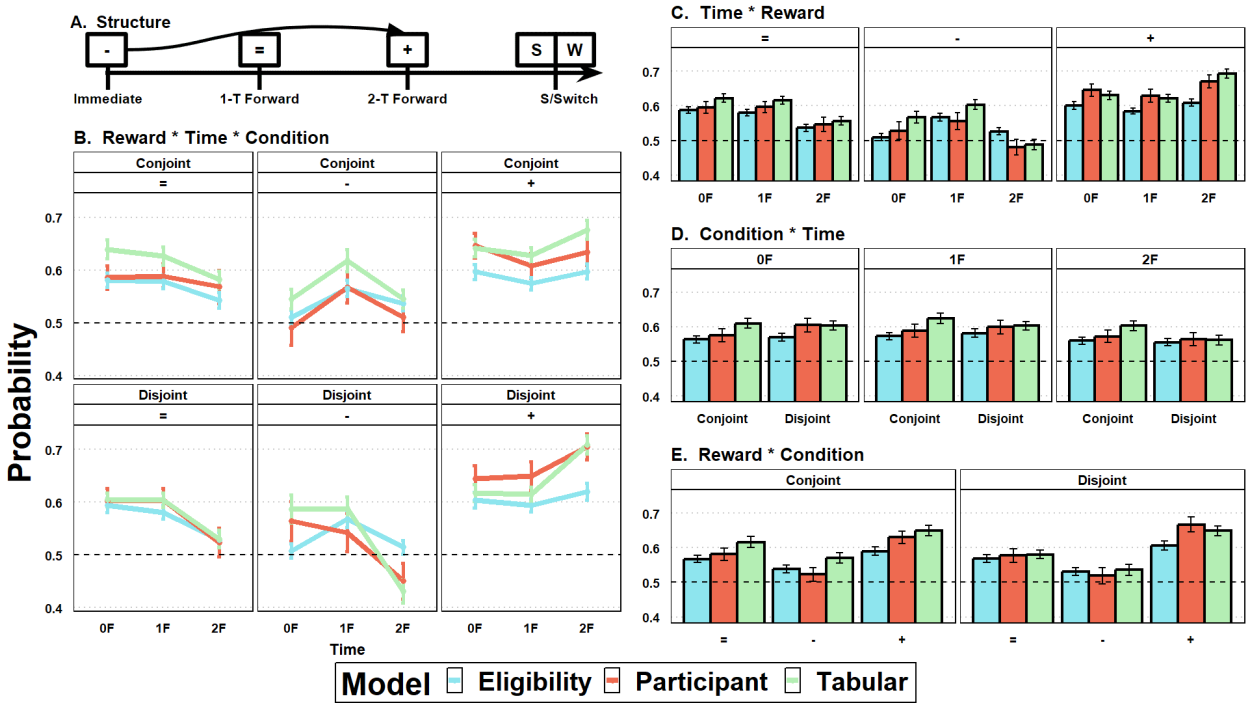
The logistic regression (equation 9) examines whether the reward feedback conditions (*Condition: conjoint, disjoint*) influenced participants' decision to stay on their delay choice three trials in the future (*Stay: 1 stay, 0 switch)*, while accounting for the reward valence (*Reward: + positive feedback, - negative feedback, = null feedback*) received from current, one, and two-trials forward (*Time: 0F, 1F, 2F, respectively*). For example, a participant might erroneously stay on their choice three trials in the future due to a positive immediate reward but should have correctly switched due to a negative reward two-trials in the future (Figure 4A). We fit three logistic regressions from equation 7: one on the participants' data, and two on the data generated from eligibility and tabular models (Figure 4B).

The three-way interaction (disjoint*+*2F) was significant for all three logistic regression models (Figure 4C), odds-ratios for participant, $b = 1.79$, SE =1.14, $Z = 4.55$, $p < .001$, 95% CI

[1.39, 2.3]; eligibility, $b = 1.19$, SE =1.08, $Z = 2.33$, $p = .02$, 95% CI [1.03, 1.37]; tabular, $b = 1.38$, SE =1.08, $Z = 4.34$, $p < .001$, 95% CI [1.19, 1.59]. For participants' data (Figure 4B – red line), the interaction was such that in the conjoint condition, the reward valence on choice (probability of stay difference between positive versus negative rewards) was strongest for immediate (0F). However, in the disjoint condition, the effect of reward valence on choice was strongest for two-trials forward (2F). Although both models were able to capture this effect, within the disjoint condition for the signal state (2F), tabular overlapped within participants' 95% CI for both negative (going below chance) and positive valence. Furthermore, this overlap suggests that the tabular model is better at predicting participants' tendency to repeat rewarded delayed choices in the disjoint rather than conjoint condition.

Figure 4

Behavioral signatures of learning and model predictions



Note. A. This diagram demonstrates the logical structure of a delayed choice that reappears after three trials, categorized as either 'stay' (S) or 'switch' (W). It includes feedback valence for immediate, one-trial, and two-trials forward choices. Correctly utilizing two-trials forward

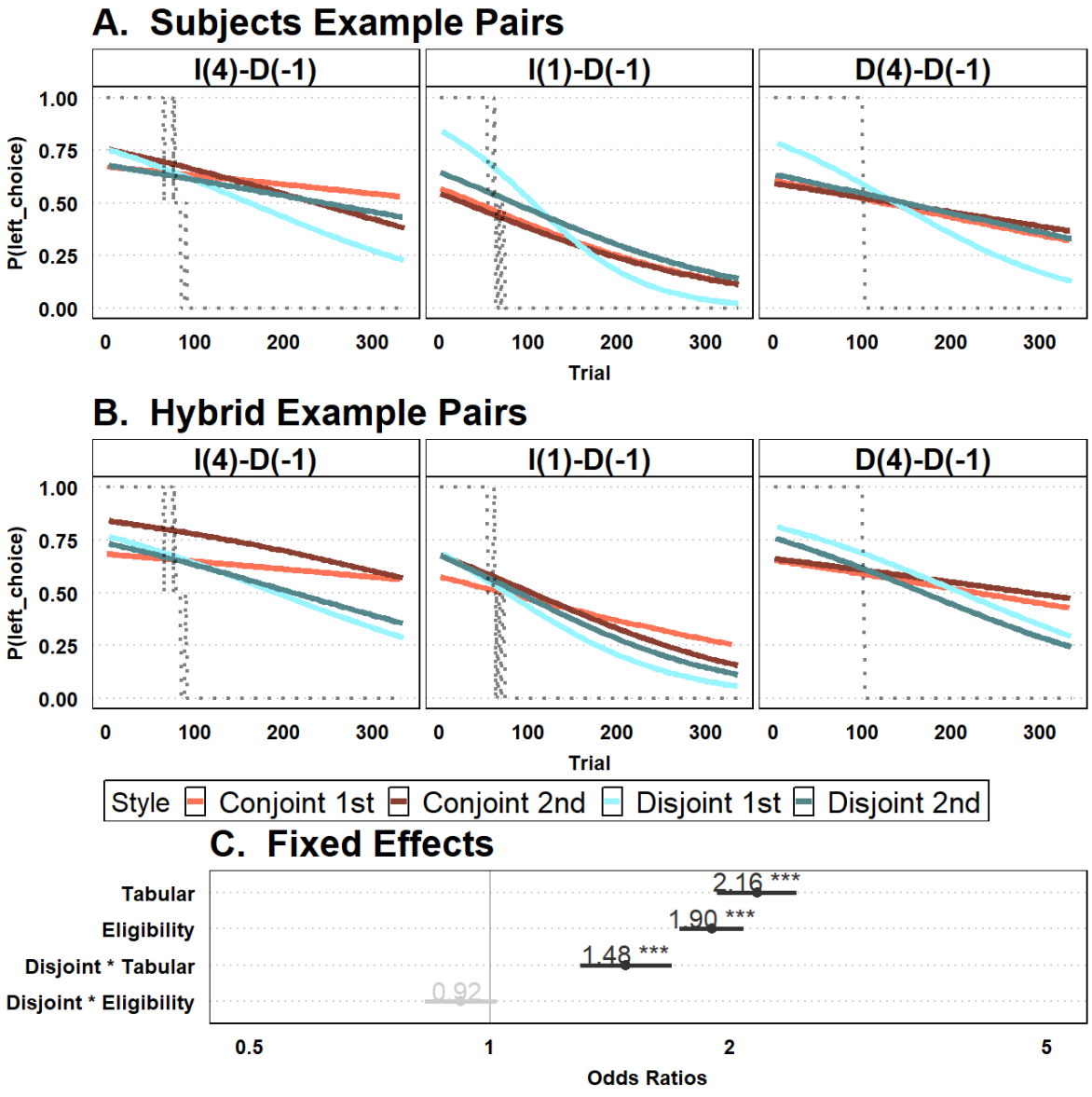
information (illustrated with a curved arrow) suggests staying with the initial choice. B. The probability of maintaining (i.e. staying on) a delayed choice three trials later was modeled using multilevel logistic regression, as a function of condition (Conjoint vs Disjoint), time (Immediate, 0F; one-trial forward, 1F; or two-trial forward, 2F), and reward (positive, +; negative, -; or null, =), as well as their interactions (equation 7). This regression was run on data from participants (red), as well as data generated by the tabular model (green) and data generated by the eligibility model (blue) for comparison. The logistic regression's estimated marginal means are displayed, accompanied by 95% confidence intervals for each condition. C-E. Collapsed two-way interactions for each of the combinations between the three variables, specifically time*reward collapsed across condition (C), condition*time collapsed across reward (D), and reward*condition collapsed across time (E). Bars represent 95% confidence intervals.

The second posterior predictive check analysis was set to reproduce the entire choice trajectories from the models. Specifically, we first showed participants' ability to track reward in some example pairs (Figure 5A), as well as how our hybrid model mimics this choice trajectory (Figure 5B). Next, to quantify the models' (eligibility and tabular) trial-by-trial predictive accuracy and compare it between conditions, we ran another multilevel logistic regression (equation 10), collapsing across all random walks and pairs. We were particularly interested in understanding how the predictions made by the tabular and eligibility models corresponded to disjoint and conjoint conditions, respectively. Specifically, we anticipated that the tabular model would predict participants' choices more accurately in the disjoint condition and the eligibility model would predict participants' choices more accurately in the conjoint condition. To test this, we employed the same task sequences observed by participants and utilized identically randomly sampled parameters to generate choice probabilities for both models (refer to Methods for further details), which were included as predictors of participants choice in the regression together with condition (equation 10, Figure 5C). Using the 'mixed' function in 'afex', we found a positive interaction between tabular predictions and condition, odds-ratio, $b = 1.48$, $F(1, 140.89) = 34.57$, $p < .001$, such that the predictions of the tabular model explained participants' choices better in the disjoint compared to the conjoint condition, as hypothesized. For the interaction between eligibility predictions and condition, we found a numeric reduction in the odds ratio, $b = .92$,

$F(1, 138.96) = 4.41$, $p = .037$, also confirming our hypothesis that eligibility predicts participants' choices more accurately in conjoint than the disjoint condition.

Figure 5

Model Prediction of trial-by-trial choice



Note. A. Average participants' propensity to choose left on three example pairs from one of the reward random walks, selected to show a clean reversal moving from left-to-right choice. The title identifies the reward contingency (I = Immediate, D = Delayed) with the initial value enclosed in paratheses (4 = initial start 4, -1 = initial start -1). The y-axis is the probability of selecting the left choice from the title, such as I(4)-D(1) illustrating the left choice to be the immediate with starting value 4. The grey dotted line tracks when the optimality of the reward

shifts from left (1.00) to equivalent (0.50) and then to right (0.00). Thus, participants in the different conditions (disjoint, conjoint) and different stages (1, 2) should follow the optimality of reward with participant data on top. Further, lighter colors belong to the same group as darker colors. B. The same analysis was performed on choice data generated by the hybrid model, using the best-fitting participant parameters. C. Fixed effects from a logistic regression predicting choice on each trial from tabular predictions, eligibility predictions, and their interaction with condition (equation 10), aligning with our hypothesis that tabular predicts behavior better in disjoint than joint condition (significant disjoint * tabular interaction), while eligibility does not (absence of disjoint * eligibility interaction). Dots and associated numbers represent the odds ratio for each effect; horizontal error bars represent 95% confidence intervals. *** $p < .001$.

Model Fits

In both conditions, our hybrid model showed better fits to the participants' data compared to either single strategy model (Table 2), suggesting that participants' behavior on this task is a mixture of eligibility and tabular learning overall. While some participants were better fitted with an independent-model eligibility strategy, an independent-model tabular strategy was the least likely to be implemented across all conditions. Participants who started in the disjoint condition had the best fit of the tabular model (in terms of likelihood and pseudo-R^2), which aligns with our previous statistics (Table 1) and hypotheses. Additionally, those who ended in the disjoint condition had similar fits to those in the conjoint condition; however, we did not perform any inferential tests to substantiate that finding.

Table 2

Model-fitting metrics

Disjoint-1	Eligibility	Tabular	Hybrid
Percent	18.92%	2.7%	78.38%
Mean	177	175.76	165.26
SD	29.19	37.19	38.33
Pseudo-R^2	.24	.245	.29

Disjoint-2	Eligibility	Tabular	Hybrid
Percent	32.35%	4.41%	63.24%
Mean	192.51	197.12	185.83
SD	27.77	29.61	32.91
Pseudo-R^2	.17	.15	.2
Conjoint-1	Eligibility	Tabular	Hybrid
Percent	39.71%	8.82%	51.47%
Mean	192.75	202.43	189.63
SD	29.02	27.44	31.67
Pseudo-R^2	.17	.13	.19
Conjoint-2	Eligibility	Tabular	Hybrid
Percent	31.08%	2.7%	66.22%
Mean	177.45	188.6	173.23
SD	39.14	37.03	40.93
Pseudo-R^2	.24	.19	.26

Note. Model-fitting metrics for each model across reward condition and stage. Percentages represent the proportion of individuals best fit by each model based on negative loglikelihoods. Mean and SD represent the mean standard deviations of negative loglikelihoods across participants, respectively. Pseudo-R^2 represents the proportion of variance accounted in participants' choices relative to a random model with higher values representing better fit.

Reinforcement Learning Model Parameters Across Conditions

Examining the correlation between parameters, we found that some of the individual model parameters were correlated with one another (Table 3). Noteworthy, correlations amongst the parameters in different conditions (between conjoint and disjoint) were surprisingly low but some were significant. The strongest correlations appeared in a relationship in the strategy

weights between eligibility and tabular; however, this did not hold across conditions. The strategy weight also represents the degree of decision stochasticity or reward sensitivity when selecting between two choices and thus would correlate between models. Other parameters seemed to tradeoff dependent on the relationship between learning rate, strategy weight, and decay weight.

Table 3

Correlation between independent model parameters

Parameter	BetaTab	BetaElg	LambdaTab	LambdaElg	AlphaTab	AlphaElg
BetaTab	**.45****	**.82****	**.21****	**.27****	.04	**-.24****
BetaElg	**.81****	**.37****	.12	.13	-.1	**-.4****
LambdaTab	.11	.03	.06	.09	-.06	.02
LambdaElg	**.26****	.07	.06	**.18****	.06	.03
AlphaTab	**-.33***	**-.4****	.02	-.01*	-.04	**.25****
AlphaElg	**-.43****	**-.44****	-.08	.02	**.62****	**.32****

Note. Diagonals are correlations between disjoint-conjoint, below are conjoint-conjoint correlations and above are disjoint-disjoint correlations. * $p < .05$, ** $p < .01$

Next, each of the RL hybrid model parameters were used in a combination of various statistical tests to assess whether each parameter differs between strategies (eligibility vs. tabular) and between conditions (conjoint vs. disjoint) (Table 4), as well as whether these effects of condition interacted with phase order (i.e. which condition was completed first) (Figure 6). First, we found that the tabular beta parameter, but not the eligibility beta parameter, varied with condition. Specifically, the weight of tabular was larger in the disjoint relative to conjoint condition, consistent with our hypothesis of increased reliance on tabular in the disjoint condition. Interestingly, the eligibility decay rate also varied with condition with more decay of

the trace in the conjoint condition. Finally, the learning rate in the tabular condition appeared to have larger updates of the prediction error in disjoint, but one should be mindful in interpreting this effect given the recovery rate of this parameter. All t-tests are shown in Table 4.

Table 4

Differences in model parameters by condition, for each strategy

Variable	df	t	p	95% CI	d
AlphaElg	141	1.26	.211	[-.02, .08]	.11
AlphaTab	141	-4.32	**<.001****	[-.21, -.08]	-.36
BetaElg	141	.28	.78	[-.08, .1]	.02
BetaTab	141	-4.04	**<.001*****	[-.18, -.06]	-.34
LambdaElg	141	-3.63	**<.001****	[-.18, -.05]	-.3
LambdaTab	141	.14	.89	[-.06, .07]	.01
Optimal	141	-4.53	**<.001*****	[-.06, -.03]	-.38
OptiDelay	141	-1.77	.08	[-.04, .00]	-.15

Note. Two-tailed paired t-tests were computed for individual parameters extracted from each independent model (Elg: Eligibility; or Tab: Tabular): learning rate (Alpha), decision weight (Beta) and decay rate (Lambda), comparing conjoint versus disjoint condition. The differences in proportion of optimal choices overall (Optimal) and optimal choices of delayed option in mixed choice types (OptiDelay) are also shown. df, degrees of freedom associated with each t-test; t, t-statistic, with negative numbers representing a larger mean in disjoint than conjoint condition; p, p-value, * $p < .05$ ** $p < .01$ *** $p < .001$, 95% CI, 95% confidence interval; d, effect size calculated as Cohen's d.

Model Parameter Mixed Effects Regression

For each parameter, we ran a mixed-effects linear regression predicting the parameter value from condition, stage group order, and their interaction. For beta-tabular, an unconditional model allowed an estimate of the ICC, accounting for 39.26% of the total variation between participants. The best fitting model held without an interaction term, showing a marginal

significant change in deviance compared to the model without an interaction, $\chi^2(1) = 3.74$, $p = .053$. The final noninteraction model showed a significant main effect of reward condition, $b = .11$, SE = .03, $F(1, 141) = 16.31$, $p < .001$, a significant main effect of stage group order, $b = .16$, SE = .04, $F(1, 140) = 12.95$, $p < .001$. The effect was such that beta-tabular was higher in the disjoint than joint condition, as predicted, but an additional effect of stage order as starting in disjoint led to greater effects than starting in conjoint (Figure 6A). However, slope differences between stage order groups had only marginal effects. For beta-eligibility, an unconditional model allowed an estimate of the ICC, accounting for 36.38% of the total variation between participants. The best fitting model held without an interaction term, showing a non-significant change in deviance when comparing models with and without the interaction term, $\chi^2(1) = .01$, $p = .92$. The final model showed no effect of reward condition, $p = .78$, but a significant effect of stage order, odds-ratio: $b = .21$, SE = .07, $F(1, 140) = 10.86$, $p < .001$. The main effect of stage order showed overall higher beta-eligibility in participants who started in the disjoint condition compared to those who started in the conjoint condition (Figure 6B).

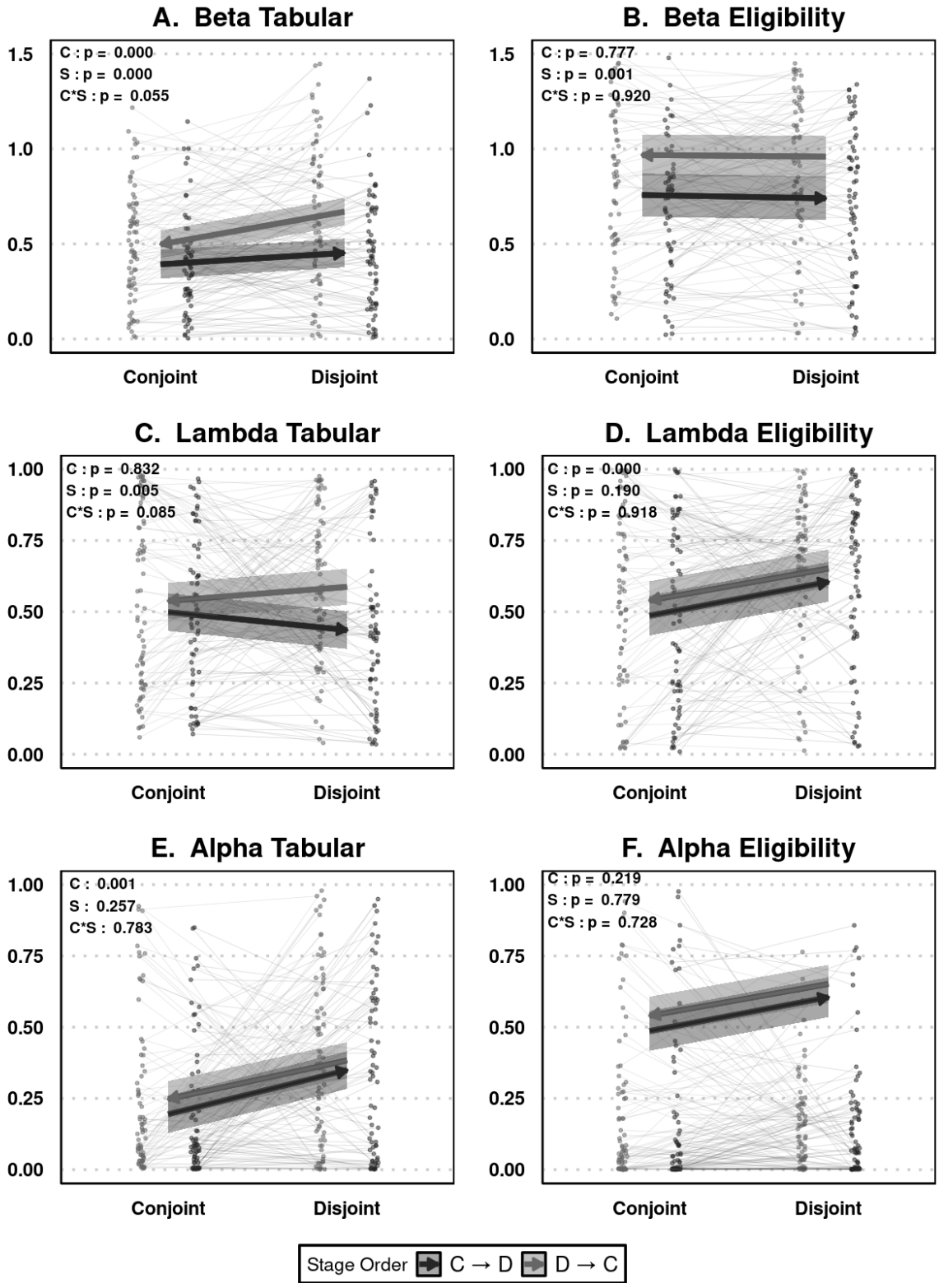
For lambda-tabular, an unconditional model allowed an estimate of the ICC, accounting for 6.15% of the total variation between participants. The best fitting model held without an interaction term, showing a marginal significant change in deviance compared to the model without an interaction, $\chi^2(1) = 3.02$, $p = .08$. There was no significant effect for reward condition, $p = .89$, but there was an effect of stage order, odds-ratio: $b = .1$, SE = .03, $F(1, 140) = 8.07$, $p = .005$ (Figure 6C). In effect, there was an increase in weighing the two-trial update when moving to the second stage, regardless of the condition. For lambda-eligibility, an unconditional model allowed an estimate of the ICC, accounting for 13.68% of the total variation between participants. The best fitting model consisted of a non-interaction model as the change in model deviance was not significant, $\chi^2(1) = .01$, $p = .92$. The non-interaction model showed a

significant effect of reward condition, odds-ratio: $b = .11$, SE = .03, $F(1, 141) = 13.15$, $p < .001$, and a non-significant effect of stage order, $p = .19$ (Figure 6D). Participants in the disjoint condition appeared to have lower rates of discounting delayed updates as compared to conjoint, indicating the eligibility trace decayed at a slower rate.

For alpha-tabular, the fit was based on a generalized least squares model with compound symmetry, as the within-subject correlation was negative, $\rho = -.05$. The negative correlation indicates that the within-subject effect appears to be inversely correlated, such that when a person has a higher learning rate, they are more likely to have a lower learning rate in the next stage or vice-versa. Maximum likelihood was used to compare models, which did not support adding the interaction term, likelihood ratio test: $\chi^2(1) = .08$, $p = .78$. The non-interaction model showed a significant effect of reward condition, $b = .15$, SE = .03, $\chi^2(1) = 18.65$, $p < .001$, and a non-significant effect of stage order, $p = .17$ (Figure 6E). Generally, the learning rate was higher for participants in the disjoint condition as compared to the conjoint condition. For alpha-eligibility, an unconditional model allowed an estimate of the ICC, accounting for 29.93% of the total variation between participants. The best fitting model consisted of a non-interaction model as the change in model deviance was not significant, $\chi^2(1) = .12$, $p = .73$. The non-interaction model showed no significant effects of reward condition, $p = .21$, or stage order, $p = .78$ (Figure 6F).

Figure 6

Effect of condition and stage on hybrid model parameters



Note. A mixed-effect linear regression model was applied to each RL parameter from their yoked individual model, predicting the parameter value from condition, stage order, and their interaction (equation 9). Specifically, this regression was performed on decision weight - Beta - for both tabular (A) and eligibility (B), decay rate - Lambda - for tabular (C), eligibility (D), as well as the learning rate - Alpha - across tabular (E) and eligibility (F). The results display estimated marginal means, marked with arrows pointing towards the participants' final condition (Conjoint and Disjoint) and the order of stages, namely from conjoint stage 1 to disjoint stage 2 (C → D, black) and from disjoint stage 1 to conjoint stage 2 (D → C, grey). Shaded areas represent 95% confidence intervals. Additionally, individual lines for each participant are shown. The top left corner of the results highlights the p-value associated with the main effect of the reward condition (C), the main effect of the stage (S), or their interaction (C * S).

Discussion

When presented with sequentially delayed rewards, the problem of credit assignment (CA) requires a person to engage in an intuitive or strategic solution (Minsky, 1961). This question of individual solutions was posed under a novel reward learning task under the guise of two differing structural strategies of temporal CA (Tanaka et al., 2009; Walsh & Anderson, 2014). To achieve this objective, we manipulated the degree of information in feedback presentation which modulated uncertainty through a partially and fully observable reward function. The eligibility trace, a viable and versatile solution for unobservable environments, was contrasted to our tabular model which differentiated credit assignment on the dimension of time (Tanaka et al., 2009; Walsh & Anderson, 2011). The eligibility trace updates with a single prediction error that decays signal and error credit towards the sequence of past actions. Contrasted to that, the tabular model employs two distinct prediction errors along the immediate and two-trial back timing but collapses across the time dimension rather than appropriately differentiating actions. Predictively, the learning efficiency differ between the two strategies but either strategy can fully observe the reward function over time; and consequently, we implemented random reward walks to prevent full learning (Tanaka et al., 2009). We predicted that in the disjoint condition, which made immediate versus delayed reward information

available, the tabular strategy would have greater utilization rates as to make use of that information, while in the conjoint condition, where no such detailed information was provided, eligibility trace may be defaulted to.

Our findings were consistent with this prediction across multiple analyses. First, the tabular model was found to capture clear patterns of behavior, specifically the tendency to repeat the choice of a delayed reward option in the disjoint versus joint condition, which the eligibility model did not capture. Second, predictions of the tabular model explained participants' choices better than eligibility model predictions in the disjoint condition, while the opposite was found in the conjoint condition. Third, parameters from the tabular and eligibility models provided more insights into the specific mechanisms deployed by participants to adapt to the change in uncertainty about delayed rewards. Regarding these mechanisms, the strategy weight of tabular was overall higher during the disjoint condition as compared to conjoint with an additional increase when starting in the disjoint condition. The decay rate only held differences between the stage order for each group, such that starting in the disjoint condition led to larger updates of the two-trial back option overall compared to starting in the conjoint condition. For eligibility, the effect of condition was only found in the decay rate, such that rates of CA were longer in the disjoint condition; whereas strategy weight remained constant across conditions was higher in the group of participants who started in the disjoint condition.

These effects are not without certain caveats for the parameter mechanisms. Despite the effect of condition on the tabular learning rate, we choose not to interpret this effect due to poor recovery of this parameter. We decided to interpret our inverse temperature parameters as strategy weights, but we note that they may also be confounded with an index of choice stochasticity. For reinforcement learning, these temperature parameters can further viewed under different paradigms, such as explore-exploit, stochastic-deterministic choice, or sensitivity to

value differences (Eckstein et al., 2022; Luce, 2005; Sutton & Barto, 2018). In our data, it is possible that the generally higher beta parameters in participants who started in the disjoint condition (compared to those who started in the conjoint condition), as shown by a main effect of stage order group on both beta tabular and beta eligibility, might reflect higher choice consistency (less stochasticity) in this group of participants. Generally, the order effects we observed for some of the parameters could reflect boredom with the task. That is, when starting in the conjoint condition, people were primed with more uncertainty that led to less effortful answers in the second stage. Comparatively, starting in the disjoint condition could result in better understanding of the task structure as to observe the distinction between immediate and delayed rewards.

The eligibility trace strategy weight did not substantially change between conditions, which may highlight its suitability across varying uncertainty. The eligibility trace mechanism 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, over time agents would fully observe the reward function rather than rely on cognitively dissociating the feedback. 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).

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; Nguyen et al., 2023). However, as the complexity of these environments increases, our understanding of effective strategies to navigate uncertainty remains limited (Gershman & Daw, 2017). 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, consider a multistep environment with fixed transitions and where the uncertainty is represented in a partially observable reward function. 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. The individual dilemma of determining the correct weight mixture of tracing and explicit transitions 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; Watkins, 1989). 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; 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 when appropriating the correct amount of complexity to an environmental problem.

Another popular tradeoff has been termed between MF and MB learners. However, the MF and MB distinction is carefully tied to a probabilistic stage transitions that can tease apart the influence of these computational systems (Daw et al., 2011) and their variation with uncertainty in the environment (Lee et al., 2014). Recent research has made significant progress in

distinguishing between MF and MB approaches to CA, revealing a dynamic interplay between retrospective and prospective processes (Deserno et al., 2021; Gershman et al., 2014; Moran et al., 2019, 2021; Shahar et al., 2019, 2021). This emerging body of work highlights how information that updates the current model can lead to retrospectively reassigning credit based on revealed structural information. However, when the state-transitions are randomly related to one another like in this task or in real world events that are independent from one another, then the structural solution is not entirely obvious. Regarding this, Tanaka et al.'s (2009) 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. How participants are solving this task remains a conundrum, where our two-condition system - one with disjoint reward and one conjoint - helped demonstrate an increased propensity to engage a prospective system given disjoint feedback.

The current strategies implemented in our eligibility and tabular models are restricted to mechanisms involved in RL. 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). Future studies might look into modifying these models for preservation, working memory, or learning rate asymmetry to potentially capture more of the decision variance (Collins & Frank, 2012; Niv et al., 2005; Sugawara & Katahira, 2021). 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 on the variety of individual strategies in overcoming CA in sequentially delayed rewards.

In summary, CA is a nontrivial problem and the mechanisms implemented in human solutions remain elusive. Despite the complex and demanding nature of the task, participants were able to overcome shifting rewards in a delayed repeat decision task with intervening events. As such, we manipulated the manner of feedback to incentivize participants to increase prospective processes in hope of characterizing differential strategy use related to information uncertainty. To this, we found evidence of increased weight of a decision strategy that was more efficient when additional information was provided and dependent on the order participants experienced the task. Thus, participants appear to choose to utilize this information or forgo it dependent on the observability of uncertain states. We hope that further investigations will examine new avenues in this experimental design and computational modeling decomposition of individual credit assignment strategies, as well as individual differences in the implementations of these strategies.

Acknowledgements

This work was supported by the University of Maryland start-up funds to AS and CJC. CJC is also supported by a National Institute of Mental Health R00 award (R00MH123669). A special thanks to Catherine Hartley and Gail Rosenbaum for their invaluable support and guidance during the initial development of this topic and methods.

References

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

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

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

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

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

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *eLife*, *11*,

e75474. https://doi.org/10.7554/eLife.75474

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.

925 Horan, W. P., Johnson, M. W., & Green, M. F. (2017). Altered experiential, but not hypothetical,
926 delay discounting in schizophrenia. *Journal of Abnormal Psychology*, *126*(3), 301–311.
927 https://doi.org/10.1037/abn0000249
928 Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory,
929 attention control, and the n-back task: A question of construct validity. *Journal of*
930 *Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622.
931 https://doi.org/10.1037/0278-7393.33.3.615
932 Kearns, M. J., & Singh, S. (2000). *Bias-Variance Error Bounds for Temporal Difference*
933 *Updates.* 142–147.
934 Kim, D., Jeong, J., & Lee, S. W. (2021). Prefrontal solution to the bias-variance tradeoff during
935 reinforcement learning. *Cell Reports*, *37*(13), 110185.
936 https://doi.org/10.1016/j.celrep.2021.110185
937 Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural Computations Underlying Arbitration
938 between Model-Based and Model-free Learning. *Neuron*, *81*(3), 687–699.
939 https://doi.org/10.1016/j.neuron.2013.11.028
940 Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019).
941 One-shot learning and behavioral eligibility traces in sequential decision making. *eLife*, *8*,
942 e47463. https://doi.org/10.7554/eLife.47463
943 Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Courier Corporation.
944 Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior*
945 *Research Methods*, *49*, 1494–1502.
946 Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, *49*(1), 8–30.
947 https://doi.org/10.1109/JRPROC.1961.287775
948 Moran, R., Dayan, P., & Dolan, R. J. (2021). Human subjects exploit a cognitive map for credit

949 assignment. *Proceedings of the National Academy of Sciences*, *118*(4), e2016884118.
950 https://doi.org/10.1073/pnas.2016884118
951 Moran, R., Keramati, M., Dayan, P., & Dolan, R. J. (2019). Retrospective model-based inference
952 guides model-free credit assignment. *Nature Communications*, *10*(1), Article 1.
953 https://doi.org/10.1038/s41467-019-08662-8
954 Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R Package
955 for Global Optimization by Differential Evolution. *Journal of Statistical Software*, *40*(6),
956 1–26. https://doi.org/10.18637/jss.v040.i06
957 Nguyen, T. N., McDonald, C., & Gonzalez, C. (2023). *Credit Assignment: Challenges and*
958 *Opportunities in Developing Human-like AI Agents* (arXiv:2307.08171). arXiv.
959 http://arxiv.org/abs/2307.08171
960 Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C.
961 (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention
962 Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.
963 https://doi.org/10.1523/JNEUROSCI.2978-14.2015
964 Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral*
965 *and Brain Functions*, *1*(1), 6. https://doi.org/10.1186/1744-9081-1-6
966 Pierce, J., Hirst, R., & MacAskill, Michael. (2022). *Building Experiments in PsychoPy* (Version
967 3) [Computer software]. https://uk.sagepub.com/en-gb/eur/building-experiments-in-
968 psychopy/book273700
969 Rescorla, R. A., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the
970 effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II* (pp.
971 64–99). Appleton-Century-Crofts, New York.
972 Shahar, N., Hauser, T. U., Moran, R., Moutoussis, M., NSPN consortium, Principal

973 investigators, Bullmore, E., Dolan, R. J., Goodyer, I., Fonagy, P., Jones, P., NSPN
974 (funded) staff, Moutoussis, M., Hauser, T., Neufeld, S., Romero-Garcia, R., Clair, M. S.,
975 Vértes, P., Whitaker, K., … Dolan, R. J. (2021). Assigning the right credit to the wrong
976 action: Compulsivity in the general population is associated with augmented outcome-
977 irrelevant value-based learning. *Translational Psychiatry*, *11*(1), 564.
978 https://doi.org/10.1038/s41398-021-01642-x
979 Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., NSPN
980 Consortium, & Dolan, R. J. (2019). Credit assignment to state-independent task
981 representations and its relationship with model-based decision making. *Proceedings of*
982 *the National Academy of Sciences*, *116*(32), 15871–15876.
983 https://doi.org/10.1073/pnas.1821647116
984 Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement
985 of causality by human subjects. *The Quarterly Journal of Experimental Psychology*,
986 *41*(2), 139–159.
987 Singh, S. P., & Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces.
988 *Machine Learning*, *22*(1), 123–158.
989 Solway, A., Lohrenz, T., & Montague, P. R. (2017). Simulating future value in intertemporal
990 choice. *Scientific Reports*, *7*(1), 43119. https://doi.org/10.1038/srep43119
991 Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and
992 perseverance in human reinforcement learning. *Scientific Reports*, *11*(1), 3574.
993 https://doi.org/10.1038/s41598-020-80593-7
994 Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. University of
995 Massachusetts Amherst.
996 Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition).

The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1–2), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

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

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

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