A new model of decision processing in instrumental learning tasks

Steven Miletić[a]*, Russell J. Boag[a], Anne C. Trutti[a,b], Birte U. Forstmann[a], Andrew Heathcote[a,c]

[a] University of Amsterdam, Department of Psychology, Nieuwe Achtergracht 129B, Amsterdam, The Netherlands
[b] Leiden University, Department of Psychology, Wassenaarseweg 52, Leiden, The Netherlands
[c] University of Newcastle, School of Psychology, Newcastle, Australia

*Correspondence concerning this article should be addressed to Steven Miletić, Nieuwe Achtergracht 129B, 1001NK Amsterdam, The Netherlands. E: s.miletic@uva.nl

Abstract

Learning and decision making are interactive processes, yet cognitive modelling of error-driven learning and decision making have largely evolved separately. Recently, evidence accumulation models (EAMs) of decision making and reinforcement learning (RL) models of error-driven learning have been combined into joint RL-EAMs that can in principle address these interactions. However, we show that the most commonly used combination, based on the diffusion decision model (DDM) for binary choice, consistently fails to capture crucial aspects of response times observed during reinforcement learning. We propose a new RL-EAM based on an advantage racing diffusion (ARD) framework for choices among two or more options that not only addresses this problem but captures stimulus difficulty, speed-accuracy trade-off, and stimulus-response-mapping reversal effects. The RL-ARD avoids fundamental limitations imposed by the DDM on addressing effects of absolute values of choices, as well as extensions beyond binary choice, and provides a computationally tractable basis for wider applications.

Keywords: Decision making, reinforcement learning, evidence-accumulation models, speed-accuracy trade-off, reversal learning.

Learning and decision-making are mutually influential cognitive processes. Learning processes refine the internal preferences and representations that inform decisions, and the outcomes of decisions underpin feedback-driven learning (Bogacz and Larsen, 2011). Although this relation between learning and decision-making has been acknowledged (Bogacz and Larsen, 2011; Dayan and Daw, 2008), the study of cognitive processes underlying feedback-driven learning on the one hand, and of perceptual and value-based decision-making on the other, have progressed as largely separate scientific fields. In the study of error-driven learning (O'Doherty et al., 2017; Sutton and Barto, 2018), the decision process is typically simplified to soft-max, a descriptive model that offers no process-level understanding of how decisions arise from representations, and ignores choice response times (RTs). In the study of decision-making using evidence-accumulation models (EAMs; Donkin and Brown, 2018; Forstmann et al., 2016; Ratcliff et al., 2016), tasks are typically designed to minimize the influence of learning, and residual variability caused by learning is treated as noise.

Recent advances (Fontanesi et al., 2019a, 2019b; Luzardo et al., 2017; McDougle and Collins, 2020; Miletić et al., 2020; Millner et al., 2018; Pedersen et al., 2017; Pedersen and Frank, 2020; Sewell et al., 2019; Sewell and Stallman, 2020; Shahar et al., 2019; Turner, 2019) have emphasized how both modelling traditions can be combined in joint models of reinforcement learning (RL) and evidence-accumulation decision-making processes, providing mutual benefits for both fields. Combined models generally propose that value-based decision-making and learning interact as follows: For each decision a subject gradually accumulates evidence for each choice option by sampling from a distribution of memory representations of the subjective value (or *expected reward*) associated with each choice option (known as *Q-values*). Once a threshold level of evidence is reached, they commit to the decision and initiate a corresponding motor process. The response triggers feedback, which is used to update the internal representation of subjective values. The next time the subject encounters the same choice options, this updated internal representation changes evidence accumulation.

The RL-EAM framework has many benefits (Miletić et al., 2020). It allows for studying a rich set of behavioral data simultaneously, including entire RT distributions and trial-by-trial dependencies in choices and RTs. It posits a theory of evidence accumulation that assumes a memory representation of rewards is the source of evidence, and it formalizes how these memory representations change due to learning. It complements earlier work connecting theories of reinforcement learning and decision-making (Bogacz and Larsen, 2011; Dayan and Daw, 2008) and their potential neural implementation in basal ganglia circuits (Bogacz and Larsen, 2011), by presenting a measurement model that can be fit to, and makes predictions about, behavioral data. Adding to benefits in terms of theory building, the RL-EAM framework also has potential to improve parameter recovery properties compared to standard RL models (Shahar et al., 2019), and allows for the estimation of single-trial parameters of the decision model, which can be crucial in the analysis of neuroimaging data.

An important challenge of this framework is the number of modeling options in both the fields of reinforcement learning and decision-making. Even considering only model-free (as opposed to model-based (Daw and Dayan, 2014)) reinforcement learning, there exists a variety of learning rules (e.g., Palminteri et al., 2015; Rescorla and Wagner, 1972; Rummery and Niranjan, 1994; Sutton, Richard, 1988), as well as the possibility of multiple learning rates for positive and negative prediction errors (Christakou et al., 2013; Daw et al., 2002; Frank et al., 2009; Gershman, 2015; Haughey et al., 2007; Niv et al., 2012), and many additional concepts, such as eligibility traces to allow for updating of previously visited states (Barto et al., 1981; Bogacz et al., 2007). Similarly, in the decision-making literature, there exists a wide range of evidence-accumulation models, including most prominently the diffusion decision model (DDM; Ratcliff, 1978; Ratcliff et al., 2016) and race models such as the linear ballistic accumulator model (LBA; Brown and Heathcote, 2008) and racing diffusion (RD) models (Boucher et al., 2007; Hawkins and Heathcote, 2020; Leite and Ratcliff, 2010; Logan et al., 2014; Purcell et al., 2010; Ratcliff et al., 2011; Tillman et al., 2020).

The existence of this wide variety of modelling options is a double-edged sword. On the one hand, it highlights the success of the general principles underlying both modelling traditions (i.e., learning from prediction errors and accumulate-to-threshold decisions) in explaining behavior, and it allows for studying specific learning/decision-making phenomena. On the other hand, it constitutes a bewildering combinatorial explosion of potential RL-EAMs; here we provide empirical grounds to navigate this problem with respect to EAMs.

The DDM is the dominant EAM as currently used in reinforcement learning(Fontanesi et al., 2019a, 2019b; Millner et al., 2018; Pedersen et al., 2017; Pedersen and Frank, 2020; Sewell et al., 2019; Sewell and Stallman, 2020; Shahar et al., 2019), but this choice is without experimental justification. Furthermore, the DDM has several theoretical drawbacks, such as its inability to explain multi-alternative decision-making and its strong commitment to the accumulation of the evidence *difference*, which leads to difficulties in explaining behavioral effects of absolute stimulus and reward magnitudes without additional mechanisms (Fontanesi et al., 2019a; Ratcliff et al., 2018; Teodorescu et al., 2016). Here, we compare the performance of different decision-making models in explaining choice behavior in a variety of instrumental learning tasks. Models that fail to capture crucial aspects of performance run the risk of producing misleading psychological inferences. For EAMs, the full RT distribution (i.e., its level of variability and skew) have proven to be crucial. Hence, it is important to assess which RL-EAMs are able to capture not only learning-related changes in choice probabilities and mean RT, but also the general shape of the entire RT distribution and how it changes with learning. Further, in order to be held forth as a general modeling framework, it is important to capture how all of these measures interact with key phenomena in the decision-making and learning literature.

We compare the RL-DDM with two RL-EAMs based on a racing accumulator architecture (Figure 1). All RL-EAMs assume evidence accumulation is driven by Q-values, which change based on error-driven learning as governed by the classical State-Action-Reward-State-Action (SARSA; Rummery and Niranjan, 1994) update rule. Rather than a two-sided DDM process (Figure 1A), the alternative models adopt a neurally plausible RD architecture (Ratcliff et al., 2007), which conceptualize decision making as a statistically independent race between single-sided diffusive accumulators, each collecting evidence for a different choice option. The first accumulator to reach its threshold triggers motor processes that execute the corresponding decision. The alternative models differ in how the mean values of evidence are constituted. The first model, the RL-RD (Figure 1B), postulates accumulators are driven by the expected reward for their choice, plus a stimulus-independent baseline (c.f. an *urgency* signal; Miletić and Van Maanen, 2019). The second model, the RL-ARD (advantage racing diffusion), uses the recently proposed *advantage* framework (Van Ravenzwaaij et al., 2020), assuming that each accumulator is driven by weighted combination of three terms: the *difference* ("advantage") in mean reward expectancy of one choice option over the other, the *sum* of the mean reward expectancies, and the urgency signal. In perceptual choice the advantage term consistently dominates the sum term by an order of magnitude (Van Ravenzwaaij et al., 2020), but the sum term is necessary to explain the effects of absolute stimulus magnitude. We also fit a limited version of this model, RL-lARD, with the weight of the sum term set to zero to test whether accounting for the influence of the sum is necessary even when reward magnitude is not manipulated, as was the case in our experiments. The importance of sum and advantage terms is also quantified by their weights as estimated in full RL-ARD model fits.

For all models, we first test how well they account for RT distributions (central tendency, variability, and skewness of RTs), accuracies, and learning-related changes in RT distributions and accuracies in a typical instrumental learning task (Frank, 2004). In this experiment we also manipulated difficulty, that is, the magnitude of the difference in average reward between pairs of options. In two further experiments we test the ability of the RL-EAMs to capture key behavioral phenomena in the decision-making and reinforcement-learning literatures, respectively, speed-accuracy trade-off (SAT), and reversals in reward contingencies. Again, these tests required a comprehensive account of not only choice probabilities but also the full distribution of RT, and learning-related changes thereof.

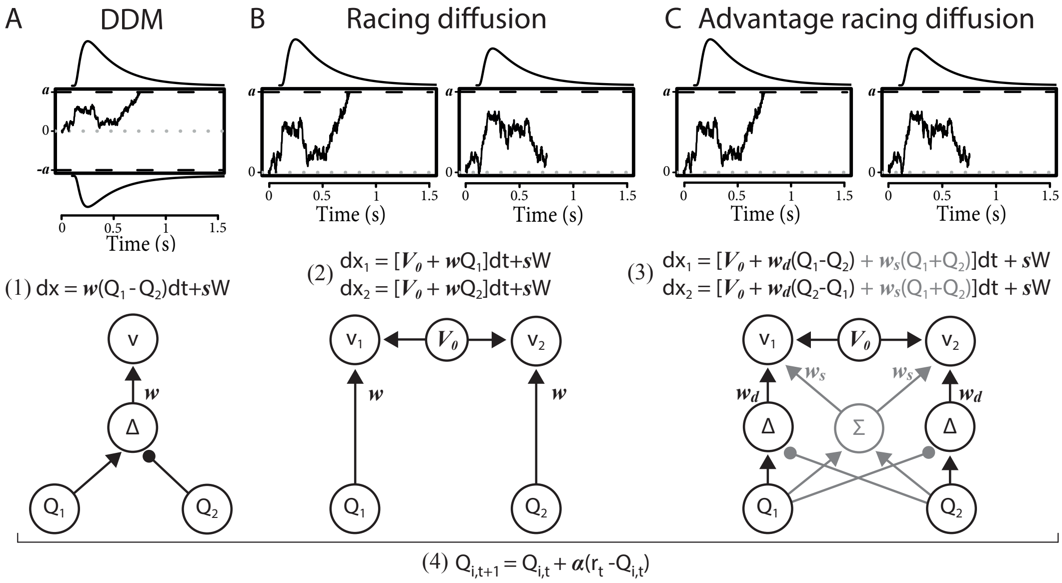
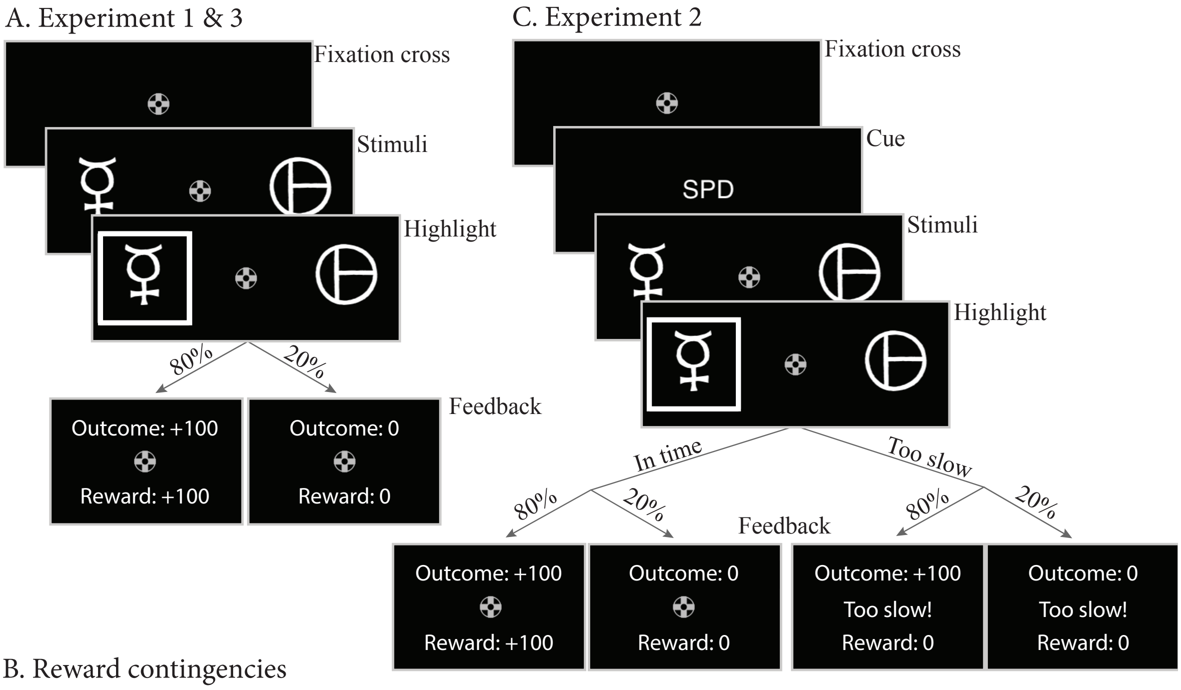

A
DDM
B
Racing diffusion
C Advantage racing diffusion
a
0
-a
a
0
0
0.5
1
1.5
Time (s)
(1) $dx = \boldsymbol{w}(Q_1 - Q_2)dt + \boldsymbol{s}W$
(2) $dx_1 = [\boldsymbol{V_0} + \boldsymbol{w}Q_1]dt + \boldsymbol{s}W$
$dx_2 = [\boldsymbol{V_0} + \boldsymbol{w}Q_2]dt + \boldsymbol{s}W$
(3) $dx_1 = [\boldsymbol{V_0} + \boldsymbol{w_d}(Q_1 - Q_2) + \boldsymbol{w_s}(Q_1 + Q_2)]dt + \boldsymbol{s}W$
$dx_2 = [\boldsymbol{V_0} + \boldsymbol{w_d}(Q_2 - Q_1) + \boldsymbol{w_s}(Q_1 + Q_2)]dt + \boldsymbol{s}W$
v
v_1
v_2
V_0
w
w_d
w_s
Δ
Σ
Q_1
Q_2
(4) $Q_{i,t+1} = Q_{i,t} + \boldsymbol{\alpha}(r_t - Q_{i,t})$

Figure 1. Comparison of the decision-making models. Bottom graphs visualize how Q-values are linked to accumulation rates. Top panel illustrates the evidence-accumulation process of the DDM (panel A) and racing diffusion (RD) models (panels B and C). Note that in the race models there is no lower bound. Equations 1-3 formally link Q-values to evidence-accumulation rates. In the RL-DDM, the difference (Δ) in Q-values is accumulated, weighted by free parameter w, plus additive within-trial white noise W with standard deviation s. In the RL-RD, the (weighted) Q-values for both choice options are independently accumulated. An evidence-independent baseline urgency term, V_0 (equal for all accumulators), further drives evidence accumulation. In the RL-ARD models, the advantages (Δ) in Q-values are accumulated as well, plus the evidence-independent baseline term V_0. The grey icons indicate the influence of the Q-value *sum* (Σ) on evidence accumulation, which is not included in the limited variant of the RL-ARD. In all panels, bold-italic faced characters indicate parameters. Q_1 and Q_2 are Q-values for both choice options, which are updated according to a SARSA learning rule (equation (4) at the bottom of the graph), with learning rate α.

Results

In the first experiment, participants made decisions between four sets of two abstract choice stimuli, each associated with a fixed reward probability (Figure 2A). On each trial, one choice option always had a higher expected reward than the other; we refer to this choice as the 'correct' choice. After each choice, participants received feedback in the form of points. Reward probabilities, and therefore choice difficulty, differed between the four sets (Figure 2B). In total, data from 55 subjects were included in the analysis, each performing 208 trials (see methods).

Throughout, we summarize RT distributions by calculating the 10th, 50th (median) and 90th percentiles separately for correct and error responses. The median summarizes central tendency, the difference between 10th and 90th percentiles summarizes variability and the larger difference between the 90th and 50th percentiles than between the 50th and 10th percentiles summarizes the positive skew that is always observed in RT distributions. To visualize the effect of learning, we divided all trials in 10 bins (approximately 20 trials each), and calculated accuracy and the RT percentiles per bin. Note that model fitting was not based on these data summaries. Instead, we used hierarchical Bayesian methods to fit models to the data from every trial and participant simultaneously. We compared model fits informally using posterior predictive distributions—calculating the same summary statistics on data generated from the fitted model as we did for the empirical data—and formally using the Bayesian Predictive Information Criterion (BPIC; Ando, 2007). The former method allows us to assess the absolute quality of fit (Palminteri et al., 2017) and detect misfits; the latter provides a model-selection criterion that trades off quality of fit with model complexity (lower BPICs are preferred), ensuring that a better fit is not only due to greater model flexibility.



B. Reward contingencies

Experiment 1					Experiment 2					Experiment 3								
											Acquisition				Reversal			
Set 1	80%	☿	⊖	20%	Set 1	80%	☿	⊖	20%	Set 1	80%	☿	⊖	20%	20%	☿	⊖	80%
Set 2	70%	☨	♆	30%	Set 2	70%	☨	♆	30%	Set 2	70%	☨	♆	30%	30%	☨	♆	70%
Set 3	65%	⚳	⚚	35%	Set 3	65%	⚳	⚚	35%									
Set 4	60%	⊛	⌖	40%														

Figure 2. Paradigms for all experiments. A: Example trial for experiment 1 and 3. Each trial starts with a fixation cross, followed by the presentation of the stimulus (until choice is made or 2.5 s elapses), a brief highlight of the chosen option, and probabilistic feedback. Reward probabilities are summarized in B. Percentages indicate the probabilities of receiving +100 points for a choice (with 0 otherwise). The actual symbols used differed between experiments and participants. In experiment 3, the acquisition phase lasted 61-68 trials (uniformly sampled each block), after which the reward contingencies for each stimulus set reversed. C: Example trial for experiment 2, which added a cue prior to each trial (‘SPD’ or ‘ACC’), and had feedback contingent on both the choice and choice timing. In the SPD condition, RTs under 700 ms were considered in time, and too slow otherwise. In the ACC condition, choices were in time as long as they were made in the stimulus window of 1.5 s. Positive feedback “Outcome: +100” and “Reward: +100” were shown in green letters, negative feedback (“Outcome: 0”, “Reward: 0”, and “Too slow!”) were shown in red letters.

We first examine results aggregated over difficulty conditions. The posterior predictives of all four RL-EAMs are shown in Figure 3, with the top row showing accuracies, and the middle and bottom rows correct and error RT distributions (parameter estimates for all models can be found in the supplementary materials). The RL-DDM generally explains the learning-related increase in accuracy well, and if only the central tendency were relevant it might be considered to provide an adequate account of RT, although correct median RT is systematically under-estimated. However, RT variability and skew are severely over-estimated. The RL-RD largely overcomes the RT distribution misfit, but it overestimates RTs in the first trial bins, and while capturing an increase in accuracy over trials, it is systematically underestimated. The RL-ARD models provide the best explanation of all key aspects of the data: except for a slight underestimation of accuracy in early trial bins (largely shared with the RL-DDM), they capture accuracy well, and like the RL-RD, they capture the RT distributions well, but without overpredicting the RTs in the early trials. The two RL-ARD models do not differ greatly in fit, except that the limited version slightly underestimates the decrease in RT with learning.

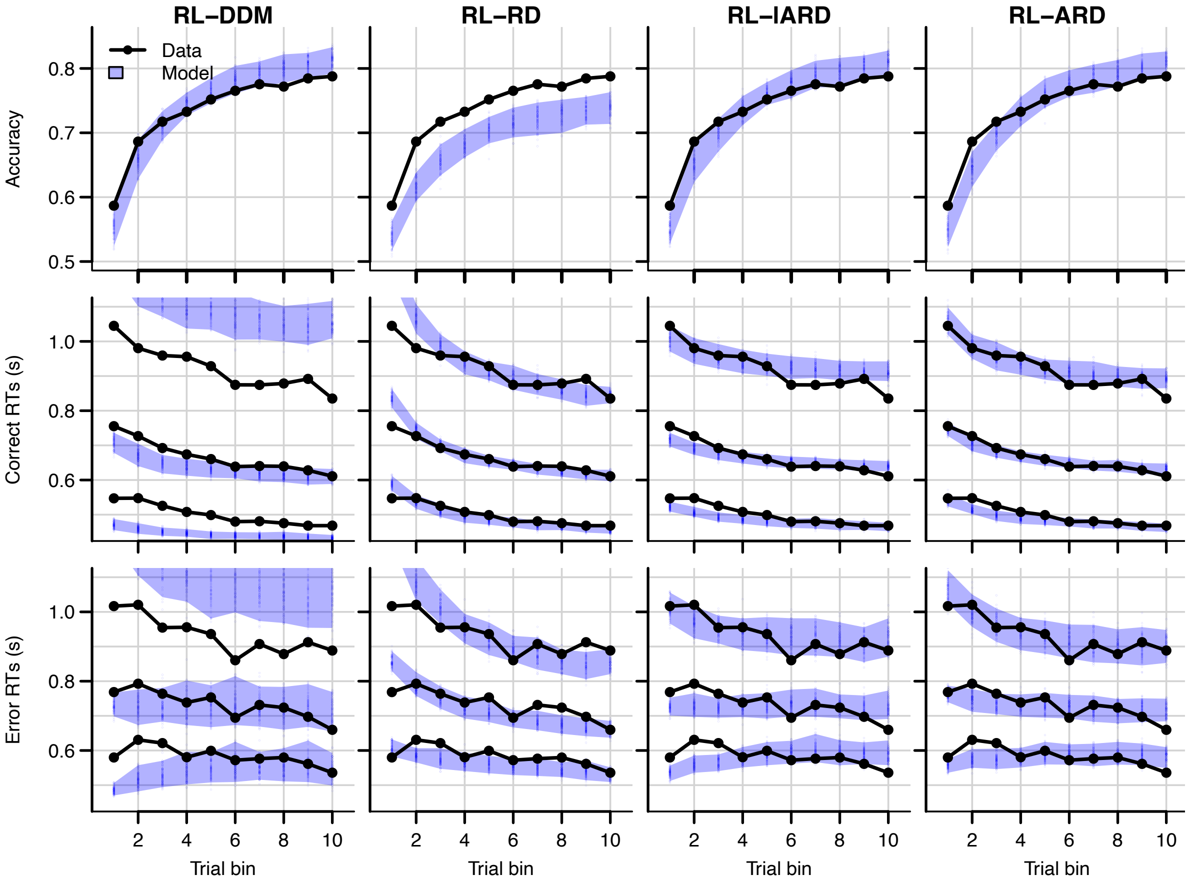


Figure 3. Comparison of posterior predictive distributions of the four RL-EAMs. Data (black) and posterior predictive distribution (blue) of the RL-DDM (left column), RL-RD, RL-lARD, and RL-ARD (right column). Top row depicts accuracy over trial bins. Middle and bottom row show 10th, 50th, and 90th RT percentiles for the correct (middle row) and error (bottom row) response over trial bins. Shaded areas correspond to the 95% credible interval of the posterior predictive distributions. All data are collapsed across participants and difficulty conditions.

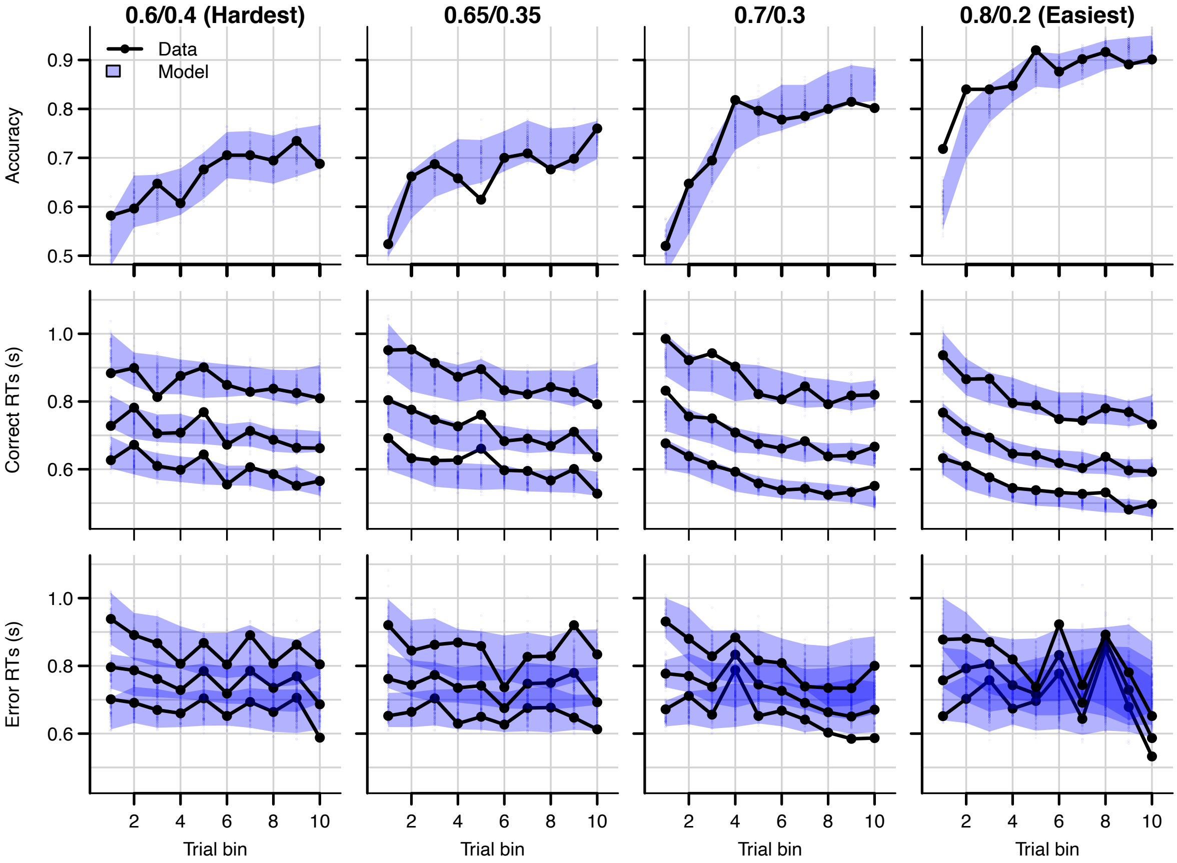


Figure 4. Data (black) and posterior predictive distribution of the RL-ARD (blue), separately for each difficulty condition. Column titles indicate the reward probabilities, with 0.6/0.4 being the most difficult, and 0.8/0.2 the easiest condition. Top row depicts accuracy over trial bins. Middle and bottom rows show 10th, 50th, and 90th RT percentiles for the correct (middle row) and error (bottom row) response over trial bins. Shaded areas correspond to the 95% credible interval of the posterior predictive distributions. All data and fits are collapsed across participants.

Figure 4 shows the data and RL-ARD model fit separated by difficulty (see supplementary materials for equivalent RL-DDM fits, which again fail to capture RT distributions). The RL-ARD model displays the same excellent fit as to data aggregated over difficulty, except that it underestimates accuracy in early trials in the easiest condition (Figure 4, bottom right panel). Further inspections of the data revealed that 17 participants (31%) reached perfect accuracy in the first bin in this condition. Likely, they guessed correctly on the first occurrence of the easiest choice pair, repeated their choice, and received too little negative feedback in the next repetitions to change their choice strategy. Supplementary materials show that, with these 17 participants removed, the overestimation is largely mitigated. SARSA assumes learning from feedback, and so cannot explain such high early accuracies. Working memory processes could have aided performance in the easiest condition, since the total number of stimuli pairs was limited and feedback was quite reliable, making it relatively easy to remember correct-choice options (Collins and Frank, 2012a, 2018; McDougle and Collins, 2020).

Reward magnitude and Q-value evolution

Q-values represent the participants' internal beliefs about how rewarding each choice option is. The RL-lARD and RL-DDM assume drift rates are driven only by the difference in Q-values (Figure 5), and both underestimate the learning-related decrease in RTs. Similar RL-DDM underestimation has been detected before (Pedersen et al., 2017), with the proposed remedy being a decrease in the decision bound with time (but with no account of RT distributions). The RL-ARD explains the additional speed-up through the increasing *sum* of Q-values over trials (Figure 5C), which in turn increases drift rates (Figure 5D). In line with observations in perceptual decision-making (Van Ravenzwaaij et al., 2020), the effect of the expected reward magnitude on drift rate is smaller (on average, $w_S = 0.36$) than that of the Q-value difference ($w_D = 2.25$) and the urgency signal ($V_0 = 2.45$). Earlier work using an RL-DDM (Fontanesi et al., 2019a) showed that higher reward magnitudes decrease RTs in reinforcement learning paradigms. There, the reward magnitude effect on RT was accounted for by allowing the threshold to change as a function of magnitude. However, this requires participants to rapidly adjust their threshold based on the identity of the stimuli, something that is usually not considered possible in EAMs (Donkin et al., 2011; Ratcliff, 1978). The RL-ARD avoids this problem, with magnitude effects entirely mediated by drift rates, and our result show expected reward magnitudes influence RTs due to learning even in the absence of a reward magnitude manipulation. Because the sum affects each accumulator equally, it changes RT with little effect on accuracy.

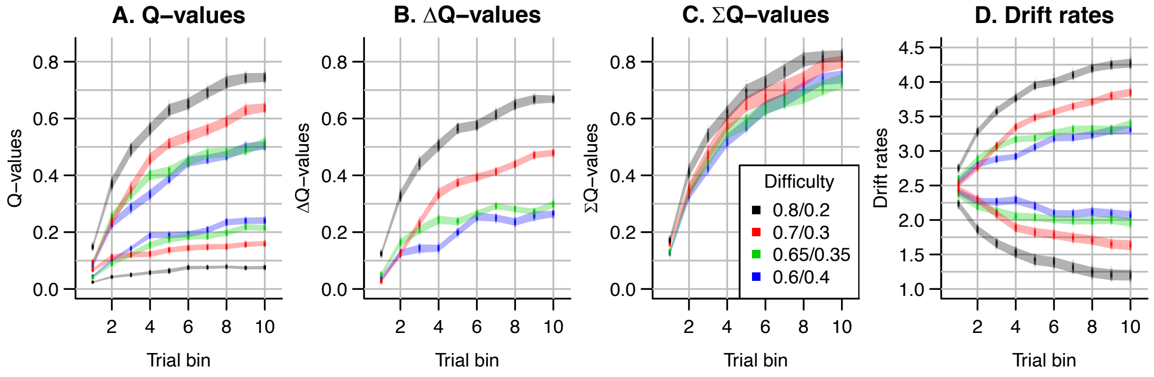


Figure 5. The evolution of Q-values and their effect on drift rates in the RL-ARD. A depicts raw Q-values, separate for each difficulty condition (colors). B and C depict the Q-value differences and the Q-value sums over time. The drift rates (D) are a weighted sum of the Q-value differences and Q-value sums, plus an intercept.

Speed-accuracy trade-off

Speed-accuracy trade-off (SAT) refers to the ability to strategically trade-off decision speed for decision accuracy (Bogacz et al., 2010; Pachella and Pew, 1968; Ratcliff and Rouder, 1998). As participants can voluntarily trade speed for accuracy, RT and accuracy are not independent variables, so analysis methods considering only one of these variables while ignoring the other (e.g., soft-max, which only focuses on choice accuracy) can be misleading. EAMs simultaneously consider RTs and accuracy and allow for estimation of SAT settings. The classical explanation in the DDM framework (Ratcliff and Rouder, 1998) holds that participants adjust their SAT by changing the decision threshold: increasing thresholds require a participant to accumulate more evidence, leading to slower but more accurate responses.

Empirical work draws a more complex picture. Several papers suggest that in addition to thresholds, drift rates (Arnold et al., 2015; Heathcote and Love, 2012; Ho et al., 2012; Rae et al., 2014; Sewell and Stallman, 2020) and sometimes even non-decision times (Arnold et al., 2015; Voss et al., 2004) can be affected. Increases in drift rates in a race model could indicate an urgency signal, implemented by drift gain modulation, with qualitatively similar effects to collapsing thresholds over the course of a decision (Cisek et al., 2009; Hawkins et al., 2015; Miletić, 2016; Miletić and Van Maanen, 2019; Murphy et al., 2016; Thura and Cisek, 2016; Trueblood et al., 2020; van Maanen et al., 2019). In cognitively demanding tasks, it has been shown that two distinct components of evidence accumulation (quality and quantity of evidence) are affected by SAT manipulations, with quantity of evidence being analogous to an urgency signal (Boag et al., 2019b, 2019a). Recent evidence suggests that different SAT manipulations can affect different psychological processes: cue-based manipulations that instruct participants to be fast or accurate, lead to overall threshold adjustments, whereas deadline-based manipulations lead to a collapse of thresholds (Katsimpokis et al., 2020).

Here, we apply an SAT manipulation in an instrumental learning task (Figure 2C). This paradigm differed from experiment 1 by the inclusion of a cue-based instruction to either stress response *speed* ('SPD') or response *accuracy* ('ACC') prior to each choice (randomly interleaved). Furthermore, on speed trials, participants had to respond within 0.7 s to receive a reward. Feedback was determined based on both the choice's probabilistic outcome ('+100' or '+0') and the RT: On trials where participants responded too late, they were additionally informed of the reward associated with their choice, had they been in time, so that they always received the feedback required to learn from their choices. After exclusions (see methods), data from 19 participants (324 trials each) were included in the analyses.

To illustrate the importance of simultaneously analyzing RTs and choice behavior, we first test whether a soft-max model (which ignores RTs) is able to capture the behavioral changes in choice probability due to the manipulation. We fit two soft-max models to the data: One with a single inverse temperature parameter, and one with an inverse temperature parameter per SAT condition. The soft-max model with separate parameters per condition was outperformed by a model with a single parameter ($\Delta BPIC = 11$), indicating that a researcher using soft-max would have concluded that there was no difference in choice behavior between conditions. Clearly, the difference in accuracy (and RTs) did indicate there were differences in behavior (see supplementary materials for formal statistical tests), showing that soft-max fails to capture a strong and well-known phenomenon of decision-making.

Next, we compared the RL-DDM and RL-ARD, and in light of the multiple psychological mechanisms potentially affected by the SAT manipulation, we allowed different combinations of threshold, drift rate, and for the RL-ARD urgency, to vary with the SAT manipulation. We fit three RL-DDM models, varying either threshold, the Q-value weighting on the drift rates parameter (Sewell and Stallman, 2020), or both. For the RL-ARD, we fit all seven possible models with different combinations of the threshold, urgency, and drift rate parameters free to vary between SAT conditions.

Formal model comparison (see Supplementary Table S1 for all BPIC values) indicated that the RL-ARD model combining response caution and urgency effects provides the best explanation of the data, in line with earlier research in non-learning contexts (Katsimpokis et al., 2020; Miletić and Van Maanen, 2019; Rae et al., 2014; Thura and Cisek, 2016). The advantage for the RL-ARD was substantial; the best RL-DDM (with only a threshold effect) performed worse than the worst RL-ARD model. The data and posterior predictive distributions of the best RL-DDM model and the winning RL-ARD model are shown in Figure 6. As in experiment 1, the RL-DDM failed to capture the shape of RT distributions, although it fit the SAT effect on accuracy and median RTs. The RL-ARD model provides a much better account of the RT distributions, including the differences between SAT conditions. In supplementary materials we show that adding non-decision time variability to the RL-DDM mitigates some of the misfit of the RT distributions, although it still consistently under-predicted the 10th percentile in the accuracy condition. Further, this model was still substantially outperformed by the RL-ARD in formal model selection (ΔBPIC = 209), and non-decision time variability was estimated as much greater than what is found in non-learning context, raising the question of its psychological plausibility.

Both RL-DDM and RL-ARD models tended to underestimate RTs and choice accuracy in the early trial bins in the accuracy emphasis condition. As in experiment 1, working memory may have contributed to the accurate but slow responses in the first trial bin for the accuracy condition (Collins and Frank, 2018, 2012b; McDougle and Collins, 2020).

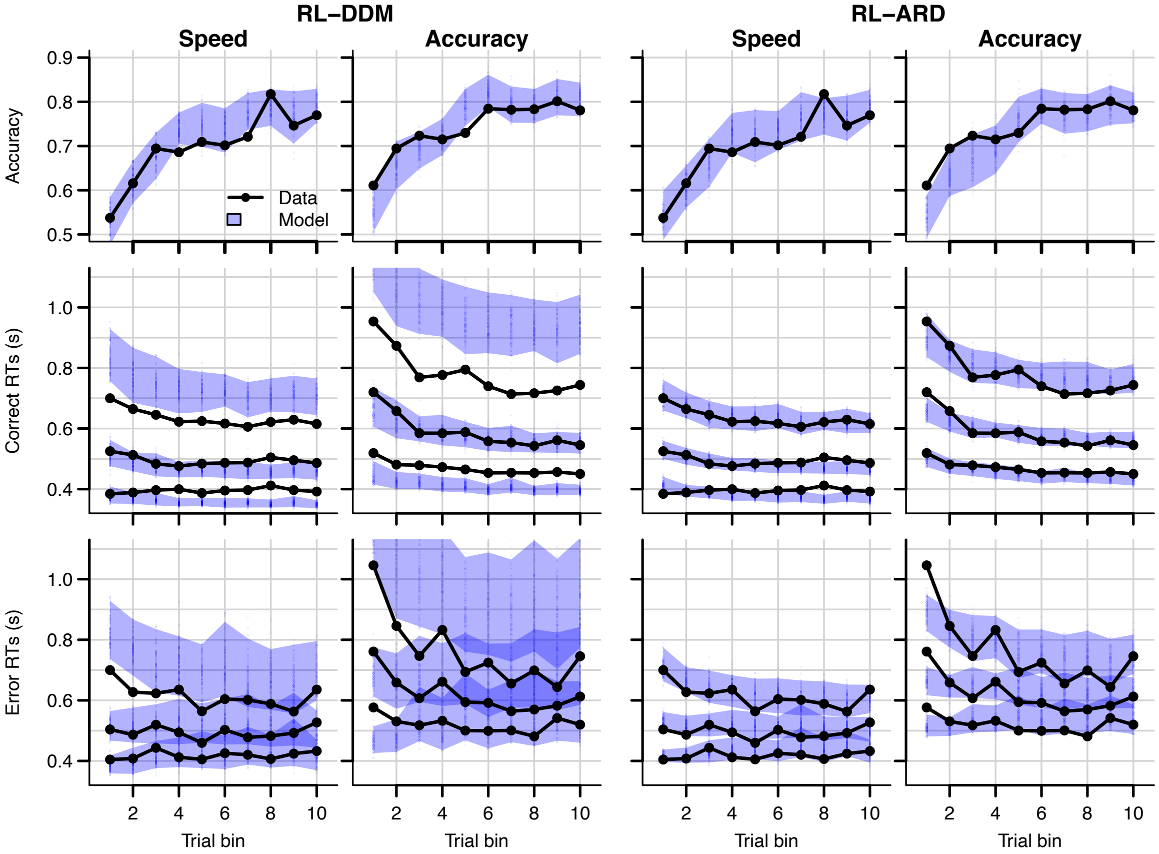


Figure 6. Data (black) and posterior predictive distributions (blue) of the best-fitting RL-DDM (left columns) and the winning RL-ARD model (right columns), separate for the speed and accuracy emphasis conditions. Top row depicts accuracy over trial bins. Middle and bottom row show 10th, 50th, and 90th RT percentiles for the correct (middle row) and error (bottom row) response over trial bins. Shaded areas in the middle and right column correspond to the 95% credible interval of the posterior predictive distribution.

Reversal learning

Finally, we tested whether the RL-ARD can capture changes in accuracy and RTs caused by a perturbation in the learning process due to reversals in reward contingencies. In the reversal learning paradigm (Behrens et al., 2007; Costa et al., 2015; Izquierdo et al., 2017) participants first learn a contingency between choice options and probabilistic rewards (the acquisition phase) that is then suddenly reversed without any warning (the reversal phase). If the link between Q-values and decision mechanisms as proposed by the RL-ARD underlies decisions, the model should be able to account for the behavioral consequences (RT distributions and decisions) of Q-value changes induced by the reversal.

Our reversal learning task had the same general structure as experiment 1 (Figure 1), except for the presence of reversals. 47 participants completed four blocks of 128 trials each. Within each block, two pairs of stimuli were randomly interleaved. Between trials 61 and 68 (uniformly sampled) in each block, the reward probability switched between stimuli, such that stimuli that were correct during acquisition were incorrect after reversal (and vice versa). Participants were not informed of the reversals prior to the experiment, but many reported noticing them.

Data and the posterior predictive distributions of the RL-DDM and the RL-ARD models are shown in Figure 7. Both models captured the change in choice proportions after the reversal reasonably well, although they underestimate the speed of change. In supplementary materials we show that the same is true for a standard soft-max model, suggesting that the learning rule is the cause of this problem. Recent evidence indicates that, instead of only estimating expected values of both choice options by error-driven learning, participants may additionally learn the task structure, estimate the probability of a reversal occurring and adjust choice behavior accordingly. Such a model-based learning strategy could increase the speed with which choice behavior changes after a reversal (Costa et al., 2015; Izquierdo et al., 2017; Jang et al., 2015), but as yet a learning rule that implements this strategy has not been developed.

The change in RT around the reversal was less marked than the change in choice probability. Once again, the RL-DDM overestimates variability and skew. Both models fit the effects of learning and reversal similarly, but the fastest responses for the RL-DDM decrease much too quickly during initial learning and the reduction in speed for the slowest responses due to the reversal is strongly overestimated. The RL-ARD provides a much better account of the shape of the RT distributions, and furthermore captures the increase in entire RT *distributions* (instead of only the median) after the reversal point. Formal model comparison also very strongly favors the RL-ARD over the RL-DDM ($\Delta BPIC = 4051$). Supplementary materials provide model comparisons to RL-DDMs with between-trial variability parameters, which lead to the same conclusion.

A notable aspect of the data is that choice behavior stabilizes approximately 20 trials after the reversal, whereas RTs remain high compared to just prior to the reversal point for up to ~40 trials. The RL-ARD explains this behavior through relatively high Q-values for the choice option that was correct during the acquisition (but not reversal) phase (i.e., choice A). Figure 8 depicts the evolution of Q-values, Q-value differences and sums, and drift rates in the RL-ARD model. The Q-values for both choice options increase until the reversal (Figure 8A), with a much faster increase for Q_A. At the reversal Q_A decreases and Q_B increases, but as Q_A

decreases faster than Q_B increases there is a temporary decrease in Q-value sums (Figure 8C). After approximately 10 trials post-reversal, Q_B is higher than for Q_A, which flips the sign of the Q-value differences (Figure 8B). However, Q_A *after* the reversal remains higher than the Q_B *before* the reversal, which causes the (absolute) Q-value differences to be lower after the reversal than before. As a consequence, the drift rates for B after the reversal remain lower than the drift rates for A before the reversal, which increases RT. Clearly, it is important to take account of the sum of inputs to accumulators as well as the difference between them in order to provide an accurate account of the effects of learning.

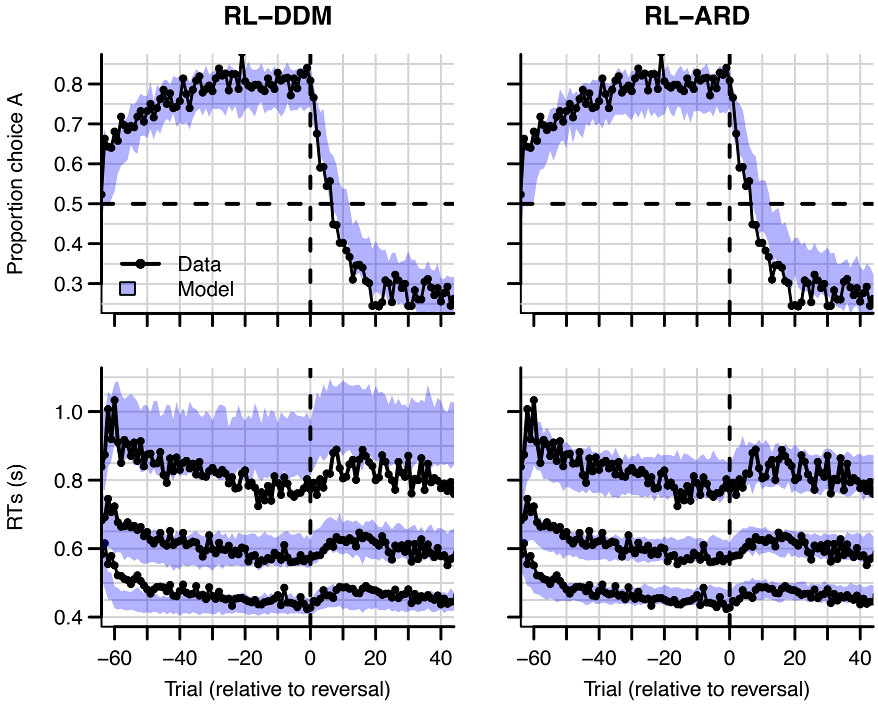


Figure 7. Experiment 3 data (black) and posterior predictive distributions (blue) for the RL-DDM (left) and RL-ARD (right). Top row: choice proportions over trials, with choice option A defined as the high-probability choice before the reversal in reward contingencies. Bottom row: 10th, 50th, and 90th RT percentiles. The data are ordered relative to the trial at which the reversal first occurred (trial 0, with negative trial numbers indicated trials prior to the reversal). Shaded areas correspond to the 95% credible interval of the posterior predictive distributions.

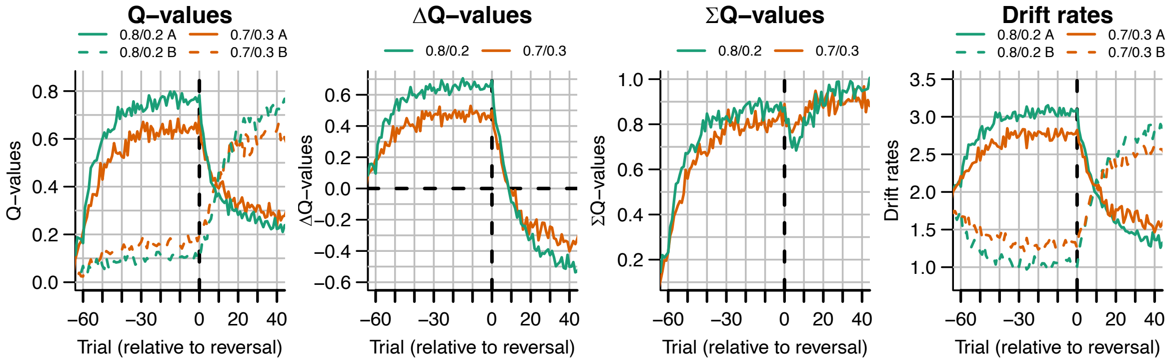


Figure 8. The evolution of Q-values and their effect on drift rates in the RL-ARD in experiment 3, aggregated across participants. Left panel depicts raw Q-values, separate for each difficulty condition (colors). The second and third panel depict the Q-value differences and the Q-value sums over time. The drift rates (right panel) are a weighted sum of the Q-value differences and Q-value sums, plus an intercept. Choice A (solid lines) refers to the option that had the high probability of reward during the acquisition phase, and choice B (dashed lines) to the option that had the high probability of reward after the reversal.

Discussion

We compared combinations of different evidence-accumulations models with a simple SARSA (Rummery and Niranjan, 1994) reinforcement learning rule (RL-EAMs). The comparison tested the ability of the RL-EAMs to provide a comprehensive account of behavior in learning contexts, not only in terms of the choices made but also the full distribution of the times to make them (RT). We examined a standard instrumental learning paradigm (Frank, 2004) that manipulated the difference in rewards between binary options (i.e., decision difficulty). We also examined two elaborations of that paradigm testing key phenomena from the decision-making and learning literatures, speed-accuracy trade-offs (SAT), and reward reversals, respectively. Our benchmark was the dual threshold Diffusion Decision Model (Ratcliff, 1978) (DDM), which has been used in almost all previous RL-EAM research, but has not been compared to other RL-EAMs, and has not been thoroughly evaluated on its ability to account for RT distributions in learning tasks. Our comparison used several different racing diffusion (RD) models, where decisions depend on the winner of a race between single barrier diffusion processes.

The RL-DDM provided a markedly inferior account to the other models, consistently overestimating RT variability and skew. As these aspects of behavior are considered critical in evaluating models in decision-making literature (Forstmann et al., 2016; Ratcliff and McKoon, 2008; Voss et al., 2013), our results question whether the RL-DDM provides an adequate model of instrumental learning. Furthermore, the DDM carries with it two important theoretical limitations. First, it can only address binary choice. This is unfortunate given that perhaps the most widely used clinical application of reinforcement learning, the Iowa gambling task (Bechara et al., 1994), requires choices among four options. Second, the input to the DDM combines the evidence for each choice (i.e., "Q" values determined by the learning rule) into a single difference, and so requires extra mechanisms to account for known effects of overall reward magnitude (Fontanesi et al., 2019a). Although there are potential ways that the RL-DDM might be modified to account for magnitude effects, such as increasing between-trial drift rate variability in proportion to the mean rate (Ratcliff et al., 2018), its inability to extend beyond binary choice remains an enduring impediment.

The best alternative model that we tested, the RL-ARD (advantage racing diffusion), which is based on the recently proposed advantage accumulation framework (Van Ravenzwaaij et al., 2020), remedied all of these problems. The input to each accumulator is the weighted sum of three components: stimulus independent "urgency", the difference between evidence for the choice corresponding to the accumulator and the alternative (the advantage), and the sum of the evidence over accumulators. The urgency component had large effect in all fits and played a key role in explaining the effect of speed-accuracy trade-offs. The advantage component, which is similar to the input to the DDM, was strongly supported over a model in which each accumulator only receives evidence favoring its own choice. The sum component provided a simple and theoretically transparent way to deal with reward magnitude effects in instrumental learning. Despite having the weakest effect among the three components, the sum was clearly necessary to provide an accurate fit to our data, even though we did not manipulate reward magnitude. It also played an important role in explaining the effect of reward reversals.

It is perhaps surprising that the RL-DDM consistently overestimated RT variability and skewness given that the DDM typically provides much better fits to data from perceptual decision-making tasks without learning. The inclusion of between-trial variability in non-decision times partially mitigated the misfit but required an implausibly high non-decision time variability, and model comparisons still favored the RL-ARD. Previous work on the RL-DDM did not investigate this issue. In many RL-DDM papers, RT distributions are either not visualized at all, or are plotted using (defective) probability density functions on top of a histogram of RT data, making it hard to detect misfit, particularly with respect skew due to the slow tail of the distribution. One exception is Pedersen et al. (2020), whose quantile-based plots show the same pattern that we found here of over-estimated variability and skewness for more difficult choice conditions, despite including between-trial variability in non-decision times. In a non-learning context, it has been shown that the DDM overestimates skewness in high-risk preferential choice data (Dutilh and Rieskamp, 2016). Together these results suggest that decision processes in value-based decision in general, and instrumental learning tasks in particular, may be fundamentally different from a two-sided diffusion process, and instead better captured by a race model such as the RL-ARD.

In the current work, we chose to use racing diffusion processes over the more often used LBA models for reasons of parsimony: error-driven learning introduces between-trial variability in accumulation rates, which are explicitly modelled in the RL-EAM framework. As the LBA includes between-trial variability in drift rates as a free parameter, multiple parameters can account for the same variance. Nonetheless, exploratory fits (see the supplementary materials) confirmed our expectation that an RL-ALBA (Advantage ALBA) model fit the data of experiment 1 well, although formal model comparisons preferred the RL-ARD. Future work might consider completely replacing one or more sources of between trial variability in the LBA with structured fluctuations due to learning and adaption mechanisms.

The parametrization of the ARD model used in the current paper followed van Ravenzwaaij et al.'s (2020) proposed ALBA model. This parametrization interprets the influence on drift rates in terms of advantages and magnitudes. However, as both the weights on Q-value differences and sums (w_D and w_S) are freely estimated parameters, the equations that define the drift rates can be rearranged as follows:

$$\begin{aligned} dx_1 &= [V_0 + w_e Q_1 - w_i Q_2]dt + sW \\ dx_2 &= [V_0 + w_e Q_2 - w_i Q_1]dt + sW \end{aligned} \tag{5}$$

Where w_e equals the sum of w_D and w_S in the parametrization of Equation (4), and w_i equals the difference between w_D and w_S. This re-parametrization shows that each drift rate is determined by an excitatory influence w_e of the Q-value associated with the accumulator, and an inhibitory influence w_i of the Q-value associated with the other accumulator. Turner (2019) proposed that inhibition plays an important role in learning tasks. Although the locus of inhibition is different in the two models, there are clear parallels that bear further investigation.

A limitation of the current work is that we collapsed across blocks in analyzing the data of experiments 2 and 3. However, in more detailed explorations (see supplementary materials for details) there were indications of second-order changes across blocks. In experiment 2, participants were faster in the first trial bin of the second and third block compared to the first block, suggesting additional practice or adaptation effects at the beginning of the experiment. In experiment 3, participants slowed down, and learned the reversal faster, after the first block. This suggests they learned about the presence of reversals in the first block and applied a different strategy in the later blocks. Although it is known that participants increase their learning rates in volatile environments (Behrens et al., 2007), this by itself does not explain a decrease in response speed. Potentially, if participants understood the task structure after the first block, model-based strategies, such as estimating the probability of a reversal having occurred, also slowed down responses.

Although the account of data provided by the RL-ARD model was generally quite accurate, some elements of misfit suggest the need for further model development. RT and accuracy were underestimated in the initial trials of the easiest condition in experiment 1, in the accuracy emphasis condition in experiment 2, and prior to reversals in experiment 3. Furthermore, the RL-ARD model underestimated the speed with which choice probability changed after reversal of stimulus-response mappings. These misfits point to a limited ability to capture the learning-related changes in behavior. This is to some degree unsurprising, since we used a very simple model of error-driven learning. Future work might explore more sophisticated mechanisms, such as multiple learning rates (Daw et al., 2002; Fontanesi et al., 2019a; Gershman, 2015; Pedersen et al., 2017) or different learning rules (Fontanesi et al., 2019b, 2019a). Furthermore, there is clearly a role for working memory in some reinforcement learning tasks (Collins and Frank, 2018, 2012b), likely explaining the accurate but slow responses we observed in the early trial bins for easy conditions.

In summary, we believe that the ARD decision mechanism provides a firm basis for further explorations of the mutual benefits that arise from the combination of reinforcement learning and evidence-accumulation models, providing constraint that is based on a more comprehensive account of data than has been possible in the past. As it stands, the RL-ARD's parameter recovery properties are good even with relatively low trial numbers, making it a suitable measurement model for simultaneously studying learning and decision-making processes, and inter-individual differences therein. Further, the advantage framework extends to multiple choice while maintaining analytical tractability and addressing key empirical phenomena in that domain, such as Hick's Law and response-competition effects (Van Ravenzwaaij et al., 2020), enabling future applications to clinical settings, such as in the Iowa gambling task (Bechara et al., 1994).

Methods

Experiment 1

Participants

61 participants (mean age 21y [SD 2.33], 47 women, 56 right handed) were recruited from the subject pool of the department of Psychology, University of Amsterdam, and participated for course credits. All participants had normal or corrected-to-normal vision and gave written informed consent prior to the experiment onset. The study was approved by the local ethics committee.

2.1.2 Task

The task was an instrumental probabilistic learning task (Frank, 2004). On each trial, the subject was presented with two abstract symbols (a 'stimulus pair') representing two choice options (see Figure 2A for an example trial). Each choice option had a fixed probability of being rewarded with points when chosen, with one choice option always having a higher probability of being rewarded than the other. The task is to discover, by trial and error, which choice options are most likely to lead to rewards, and thereby to collect as many points as possible.

After a short practice block to get familiar with the task, participants completed one block of 208 trials. Four different pairs of abstract symbols were included, each presented 52 times. Stimulus pairs differed in their associated reward probabilities: 0.8/0.2, 0.7/0.3, 0.65/0.35, and 0.6/0.4. The size of the reward, if obtained, was always the same: '+100' (or '+0' otherwise). Reward probabilities were chosen such that they differed only in the between-choice difference in reward probability, leading to varying choice difficulties while keeping the mean reward magnitude fixed.

Participants were instructed to earn as many points as possible, and to always respond before the deadline of 2 seconds. Feedback consisted of two parts: an 'outcome' and a 'reward'. The outcome corresponded to the probabilistic outcome of the choice, whereas the reward corresponded to the actual number of earned points. When participants responded before the deadline, the reward was equal to the outcome. If they were too late, the outcome was shown to allow participants to learn from their choice, but the reward they received was set to 0 to encourage responding in time. Participants received a bonus depending on the number of points earned (maximum +0.5 course credits, mean received +0.24). The task was coded in PsychoPy (Peirce et al., 2019). After this block, participants performed two more blocks of the same task with different manipulations, which are not of current interest.

Exclusion

Six participants were excluded from analysis: One reported, after the experiment, not to have understood the task, one reported a technical issue, and four did not reach an above-chance accuracy level as determined by a binomial test (accuracy cut-off 0.55, corresponding to $p < 0.05$). The final sample thus consisted of 55 subjects (14 men, mean age 21 years old [SD 2.39], 51 right-handed).

Cognitive modelling

The main analysis consists of fitting four RL-EAMs to the data and comparing the quality of the fits penalized by model complexity. We compared four different decision models: the DDM (Ratcliff, 1978), a racing diffusion (Boucher et al., 2007; Logan et al., 2014; Purcell et al., 2010; Turner, 2019) model, and two Advantage Racing Diffusion (ARD; Van Ravenzwaaij et al., 2020) models (see Figure 1 for an overview). Whereas the former is a two-sided diffusion process, the latter three models employ a race architecture.

For all models we used the State-Action-Reward-State-Action (SARSA; Rummery and Niranjan, 1994) update rule as a learning model:

$$Q_{i,t+1} = Q_{i,t} + \alpha(r_t - Q_{i,t}) \tag{4}$$

where $Q_{i,t}$ is the value representation of choice option i on trial t, α the learning rate, and r_t the reward on trial t. The difference between the actual reward and the value representation of the chosen stimulus, $r_t - Q_{i,t}$, is known as the reward prediction error. The learning rate controls the speed at which Q-values change in response to the reward prediction error, with larger learning rates leading to stronger fluctuations. In this model, only the Q-value of the chosen option is updated.

RL-EAM 1: RL-DDM

In the first RL-EAM, we use the DDM (Ratcliff, 1978) as a choice model (Figure 1, left column). The DDM assumes that evidence accumulation is governed by:

$$dx = vdt + sW$$

v is the mean speed of evidence accumulation (the *drift rate*), and s is the standard deviation of the within-trial accumulation white noise (W). The RL-DDM assumes that the drift rate depends linearly on the difference of value representations:

$$v_t = w(Q_{t,1} - Q_{t,2})$$

w is a weighting variable, and $Q_{t,1}$ and $Q_{t,2}$ are the Q-values for both choice options per trial, which change each trial according to Equation 4. Hence,

$$dx = w(Q_1 - Q_2)\,dt + sW \tag{1}$$

The starting point of evidence accumulation, z, lies between decision boundaries a and $-a$. Here, as in earlier RL-DDM work (Fontanesi et al., 2019a, 2019b; Pedersen et al., 2017), we assume an unbiased start of the decision process (i.e., z = 0). Evidence accumulation finishes when threshold a or $-a$ is reached, and the decision for the choice corresponding to Q_1 or Q_2, respectively, is made. The response time is the time required for the evidence-accumulation process to reach the bound, plus an intercept called the non-decision time ($t0$). The non-decision time is the sum of the time required for perceptual encoding and the time required for the execution of the motor response. Parameter s was fixed to 1 to satisfy scaling constraints (Donkin et al., 2009; van Maanen and Miletić, 2020). In total, this specification of the RL-DDM has 4 free parameters ($\alpha, w, a, t0$). In the supplementary materials, we fit an RL-DDM specification with between-trial variabilities in start point, drift rate, and non-decision time.

RL-EAM 2: RL-RD

The RL-RD (Figure 1, middle panel) assumes that two evidence accumulators independently accrue evidence for one choice option each, both racing towards a common threshold a (assuming no response bias). The first accumulator to hit the bound wins, and the corresponding decision is made. For each choice option i, the dynamics of accumulation are governed by:

$$dx_i = [V_0 + wQ_i]dt + sW \tag{2}$$

V_0 is a parameter specifying the drift rate in the absence of any evidence, w a weighting parameter, and s the standard deviation of within-trial noise. As such, the mean speed of accumulation (the drift rate v_i) is the sum of two independent factors: an evidence-independent baseline speed V_0, and an evidence-dependent weighted Q-value, wQ_i. Since V_0 is assumed to be identical across accumulators, and governs the speed of accumulation unrelated to the amount of evidence, we interpret this parameter as an additive urgency signal (Miletić and Van Maanen, 2019), with conceptually similar behavioral effects as collapsing bounds (Hawkins et al., 2015). Similar to the DDM, a non-decision time parameter accounts for the time for perceptual encoding and the motor response time. Parameter s was fixed to 1 to satisfy scaling constraints (Donkin et al., 2009; van Maanen and Miletić, 2020). In total, the RL-RD has 5 free parameters $(\alpha, w, a, v0, t0)$.

Each accumulator's first passage times are Wald (also known as inverted Gaussian) distributed (Anders et al., 2016). In an independent race model, each accumulator's first passage time distribution is normalized to the probability of the response with which it is associated (Brown and Heathcote, 2008; Turner, 2019).

RL-EAM 3 & 4: RL-ARD

Thirdly, we fit two racing diffusion models based on an advantage race architecture (Van Ravenzwaaij et al., 2020). An advantage race model using an LBA has been shown to provide a natural account for multi-alternative choice phenomena such as Hick's law, as well as stimulus magnitude effects in perceptual decision-making. Like in the RL-RD, accumulators race towards a common bound, but the speed of evidence accumulation v_i depends on multiple factors: first, as in the RL-RD, the evidence-independent speed of accumulation V_0; second, the *advantage* of the evidence for one choice option over the other (c.f. the DDM, where the difference between evidence for both choice options is accumulated); and third, the *sum* of the total available evidence. Combined, for two accumulators in the RL-EAM framework, this leads to:

$$\begin{aligned} dx_1 &= [V_0 + w_d(Q_1 - Q_2) + w_s(Q_1 + Q_2)]dt + sW \\ dx_2 &= [V_0 + w_d(Q_2 - Q_1) + w_s(Q_1 + Q_2)]dt + sW \end{aligned} \tag{3}$$

In the original work proposing the advantage accumulation framework (Van Ravenzwaaij et al., 2020), it was shown that the w_d parameter had a much stronger influence on evidence-accumulation rates than the w_s parameter. Therefore, we first fixed the w_s parameter to 0, to test whether the accumulation of *differences* is sufficient to capture all trends in the data. We term this model the RL-lARD (l = limited), which we compare to the RL-ARD in which we fit w_s as a free parameter.

As previously, parameter s was fixed to 1 to satisfy scaling constraints (Donkin et al., 2009; van Maanen and Miletić, 2020). The RL-ARD also has a threshold, non-decision time, and learning rate parameter, totaling five $(\alpha, w_d, a, V_0, t0)$ and 6 free parameters $(\alpha, w_d, w_s, a, V_0, t0)$ for the RL-lARD and RL-ARD, respectively.

Bayesian hierarchical parameter estimation, posterior predictive distributions, model comparisons

We estimated group-level and subject-level posterior distributions of each model's parameter using a combination of differential evolution (DE) and Markov-chain Monte Carlo sampling (MCMC) with Metropolis-Hastings (Ter Braak, 2006; Turner et al., 2013). Sampling settings were default as implemented in the Dynamic Models of Choice *R* software (Heathcote et al., 2019): The number of chains, D, was three times the number of free parameters. Cross-over probability was set to $2.38/\sqrt{D}$ at the subject-level and $U[0,1]$ at the group level. Migration probability was set to 0.05 during burn-in only. Convergence was assessed using visual inspection of the chain traces and Gelman-Rubin diagnostic (Brooks and Gelman, 1998; Gelman and Rubin, 1992) (individual and multivariate potential scale factors < 1.03 in all cases).

Hierarchical models were fit assuming independent normal population ("hyper") distributions for each parameter. For all models, we estimated the learning rate on a probit scale (mapping [0, 1] onto the real domain), with a normal prior $\alpha \sim \Phi(\mathcal{N}(-1.6, 5))$ (Spektor and Kellen, 2018). Prior distributions for all estimated hyper-mean decision-related parameters were vague. RL-EAMs, the threshold parameter $\mathrm{a} \sim \mathcal{N}(3,5)$ truncated at 0, and $\mathrm{t0} \sim \mathcal{N}(0.3, 0.5)$ truncated at 0.025 s and 1 s (all estimation was carried out on the seconds scale). For the RL-DDM, $\mathrm{w} \sim \mathcal{N}(2,5)$. For the RL-RD, $\mathrm{w} \sim \mathcal{N}(9,5)$, and for the RL-ARD models, $\mathrm{w}_D \sim \mathcal{N}(9,5)$ and $\mathrm{w}_S \sim \mathcal{N}(0,3)$. For the hyper-SD, a $\Gamma(1,1)$ distribution was used as prior. Plots of superimposed prior and posterior hyper-distributions confirmed that these prior setting were not influential.

In initial explorations, we also freely estimated the Q-values at trial 0. However, in the RL-EAMs, the posterior distributions for these Q-values consistently converged on 0, which was therefore subsequently used as a fixed value for all results reported here. For the the soft-max fits, they were set to 0.5 as often used in reinforcement learning models of two-choice tasks (Apps et al., 2015; Collins and Frank, 2018, 2012b; Fontanesi et al., 2019a; McDougle and Collins, 2020; Pedersen and Frank, 2020). Including the initial Q-values as a free parameter in the soft-max models of experiment 2 led to the same conclusions.

To visualize the quality of model fit, we took 100 random samples from the estimated parameter posteriors and simulated the experimental design with these parameters. For each behavioral measure (e.g., RT quantiles, accuracy), credible intervals were estimated by taking the range between the 2.5% and 97.5% quantiles of the averages over participants.

To quantitatively compare the fit of different models, penalized by their complexity, we used the Bayesian predictive information criterion(BPIC; Ando, 2007). The BPIC is an analogue of the Bayesian information criterion (BIC), but (unlike the BIC) suitable for models estimated using Bayesian methods. Compared to the deviance information criterion (DIC; Spiegelhalter et al., 2002), the BPIC penalizes model complexity more strongly to prevent over-fitting (c.f. AIC vs. BIC). Lower BPIC values indicate better trade-offs between fit quality and model complexity.

Experiment 2

Participants

23 participants (mean age 19 years old [SD 1.06 years], 7 men, 23 right-handed) were recruited from the subject pool of the Department of Psychology of the University of Amsterdam and participated for course credits. Participants did not participate in experiment 1 or 3. All participants had normal or corrected-to-normal vision and gave written informed consent prior to the experiment onset. The study was approved by the local ethics committee.

Task

Participants performed the same task as in experiment 1, with the addition of an SAT manipulation (Figure 2C). The SAT manipulation included both an instructional cue and a response deadline. Prior to each trial, a cue instructed participants to emphasize either decision speed (‘SPD’) or decision accuracy (‘ACC’) in the upcoming trial, and in speed trials, participants did not earn points if they were too late (> 700 ms). As in experiment 1, after each choice participants received feedback consisting of two components: an outcome and a reward. The outcome refers to the outcome of the probabilistic gamble, whereas the reward refers to the number of points participants actually received. If participants responded in time, the reward was equal to the outcome. In speed trials, participants did not earn points if they responded later than 700 ms after stimulus onset, even if the outcome was +100. On trials where participants responded too late, they were additionally informed of the reward that was associated with their choice, had they been in time. This way, even when participants are too late, they still receive the feedback that can be used to learn from their choices.

The deadline manipulation was added because we hypothesized that instructional cues alone would not be sufficient to persuade participants to change their behavior in the instrumental learning task, since that task specifically requires them to accumulate points. If the received number of points was independent of response times, the optimal strategy to collect most points would be to ignore the cue and focus on accuracy only.

Participants performed 324 trials divided over 3 blocks. Within each block, three pairs of stimuli were shown, with associated reward probabilities of 0.8/0.2, 0.7/0.3, and 0.6/0.4. Speed and accuracy trials were randomly interleaved. Figure 2C depicts the sequence of events in each trial. As this experiment also served as a pilot for an fMRI experiment, we added fixation crosses between each phase of the trial, with jittered durations. A pre-stimulus fixation cross lasted 0.5, 1, 1.5, or 2 s; fixation crosses between cue and stimulus, between stimulus and highlight, and between highlight and feedback lasted 0, 0.5, 1, or 1.5 s; and an inter-trial interval fixation cross lasted 0.5, 1, 1.5, 2, 2.5 seconds. Each trial took 7.5 seconds. The experiment took approximately 45 minutes.

Exclusion

Four participants did not reach above-chance performance as indicated by a binomial test (cut-off 0.55, $p < 0.05$), and were excluded from further analyses. The final sample thus consisted on 19 participants (mean age 19 years old [SD 1.16 years], 6 men, 19 right-handed). For one additional participant, a technical error occurred after the first block. This participant was included in the analyses, since the Bayesian estimation framework naturally down-weighs the influence of participants with fewer trials.

Cognitive modelling

First, we tested whether a standard soft-max model is able to capture the difference in choice behavior. Soft-max is given by:

$$P_{i,t} = \frac{\exp \beta Q_{i,t}}{\sum_j^J \exp \beta Q_{j,t}} \qquad (5)$$

where $P_{i,t}$ is the probability of choosing option i on trial t, J is the total number of choice options, and β is a free parameter often called the inverse temperature. The inverse temperature is often interpreted in terms of the exploration/exploitation trade-off (Daw et al., 2006), with higher values indicating more exploitation. In two-choice settings, Equation 5 can be re-written as:

$$P_{2,t} = \frac{1}{1 + \exp \beta (Q_{1,t} - Q_{2,t})} \qquad (6)$$

which highlights that the choice probability is driven by the *difference* in Q-values, weighted by the inverse temperature parameter. We hierarchically fit two soft-max models using the same parameter estimation methods as in experiment 1. One model assumed a single β parameter, the other model assumed a β parameter per SAT condition. Priors for the hypermean were set to $\beta \sim N(1{,}5)$ truncated at 0, and for the hyperSD $\Gamma(1{,}1)$.

Next, we fit three RL-DDMs and seven RL-ARDs. The three RL-DDM models varied either threshold, the Q-value weighting on the drift rates parameter (Sewell and Stallman, 2020), or both. The seven RL-ARD allowed all unique combinations of the threshold, urgency, and drift rate parameters free to vary between the speed and accuracy conditions.

For the accuracy condition, we used the same priors as in experiment 1. In the speed condition, the parameters that were free to vary were estimated as proportional differences from the accuracy conditions; specifically: $a_{spd} = \left(1 + m_{a,spd}\right) * a_{acc}$, $V_{0_{spd}} = \left(1 + m_{V_0,spd}\right) * V_{0_{acc}}$, and $v_{i_{spd}} = \left(1 + m_{v,spd}\right) * v_{i_{acc}}$. The prior used was $\mathcal{N}(0, 5)$ for the hypermean and $\Gamma(1, 1)$ for the hyperSD of all parameters m, truncated at -1.

Experiment 3

Participants

47 participants (mean age 21 years old [SD 2.81 years], 16 men, 40 right-handed) were recruited from the subject pool of the Department of Psychology of the University of Amsterdam and participated for course credits. Participants did not participate in experiment 1 or 2. All participants had normal or corrected-to-normal vision and gave written informed consent prior to the experiment onset. The study was approved by the local ethics committee.

Task

The reversal learning task had the same general task structure as experiment 1. Participants completed four blocks of 128 trials each, totaling 512 trials. Within each block, two pairs of stimuli were randomly interleaved, with associated reward probabilities of 0.8/0.2 and 0.7/0.3. Between trials 61 and 68 (uniformly sampled) of each block, the reward probability switched between stimuli, such that the stimulus with a pre-reversal reward probability of 0.8/0.7 had a post-reversal reward probability of 0.2/0.3 (and vice versa). Participants were not informed of the reversals prior to the experiment, but many reported noticing them.

In addition to the reversal learning task, the experimental session also contained a working memory task that is not of current interest. 30 participants performed the reversal learning task before the working memory task, and 17 participants afterwards. The entire experiment took approximately one hour.

Cognitive modelling

The RL-DDM and RL-ARD were fit to the data using the same methods as in experiment 1.

Data availability statement

All data are available on OSF (https://osf.io/ygrve/).

Code availability statement

All analysis code is available on OSF (https://osf.io/ygrve/).

Acknowledgements

We thank Barbara Mathiopoulou and Chris Riddell for their help collecting the data. This work was supported by an NWO-VICI grant (BUF), an ABC VIP grant and ARC DP150100272 and DP160101891 grants (AH).

References

Anders R, Alario F, Van Maanen L. 2016. The Shifted Wald Distribution for Response Time Data Analysis. *Psychol Methods* **21**:309–327.

Ando T. 2007. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* **94**:443–458. doi:10.1093/biomet/asm017

Apps MAJ, Lesage E, Ramnani N. 2015. Vicarious reinforcement learning signáis when instructing others. *J Neurosci* **35**:2904–2913. doi:10.1523/JNEUROSCI.3669-14.2015

Arnold NR, Bröder A, Bayen UJ. 2015. Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychol Res* **79**:882–898. doi:10.1007/s00426-014-0608-y

Barto AG, Sutton RS, Brouwer PS. 1981. Associative search network: A reinforcement learning associative memory. *Biol Cybern* **40**:201–211. doi:10.1007/BF00453370

Bechara a, Damasio a R, Damasio H, Anderson SW. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**:7–15. doi:10.1016/0010-0277(94)90018-3

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. 2007. Learning the value of information in an uncertain world. *Nat Neurosci* **10**:1214–1221. doi:10.1038/nn1954

Boag RJ, Strickland L, Heathcote A, Neal A, Loft S. 2019a. Cognitive Control and Capacity for Prospective Memory in Complex Dynamic Environments. *J Exp Psychol Gen* **148**:2181–2206. doi:10.1037/xge0000599

Boag RJ, Strickland L, Loft S, Heathcote A. 2019b. Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition* **191**:103974. doi:10.1016/j.cognition.2019.05.011

Bogacz R, Larsen T. 2011. Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural Comput* **23**:817–851. doi:10.1162/NECO_a_00103

Bogacz R, McClure SM, Li J, Cohen JD, Montague PR. 2007. Short-term memory traces for action bias in human reinforcement learning. *Brain Res* **1153**:111–121. doi:10.1016/j.brainres.2007.03.057

Bogacz R, Wagenmakers E-J, Forstmann BU, Nieuwenhuis S. 2010. The neural basis of the speed-accuracy tradeoff. *Trends Neurosci* **33**:10–6. doi:10.1016/j.tins.2009.09.002

Boucher L, Palmeri TJ, Logan GD, Schall JD. 2007. Inhibitory Control in Mind and Brain : An Interactive Race Model of Countermanding Saccades **114**:376–397. doi:10.1037/0033-295X.114.2.376

Brooks SP, Gelman A. 1998. General Methods for Monitoring Convergence of Iterative Simulations. *J Comput Graph Stat* **7**:434–455. doi:10.1080/10618600.1998.10474787

Brown SD, Heathcote A. 2008. The simplest complete model of choice response time: Linear ballistic accumulation. *Cogn Psychol* **57**:153–178. doi:10.1016/j.cogpsych.2007.12.002

Christakou A, Gershman SJ, Niv Y, Simmons A, Brammer M, Rubia K. 2013. Neural and Psychological Maturation of Decision-making in Adolescence and Young Adulthood. *J Cogn Neurosci* **25**:1807–1823. doi:10.1162/jocn_a_00447

Cisek P, Puskas GA, El-Murr S. 2009. Decisions in changing conditions: the urgency-gating model. *J Neurosci* **29**:11560–71. doi:10.1523/JNEUROSCI.1844-09.2009

Collins AGE, Frank MJ. 2018. Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc Natl Acad Sci* **115**:2502–2507. doi:10.1073/pnas.1720963115

Collins AGE, Frank MJ. 2012a. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur J Neurosci* **35**:1024–1035. doi:10.1111/j.1460-9568.2011.07980.x

Collins AGE, Frank MJ. 2012b. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur J Neurosci* **35**:1024–1035. doi:10.1111/j.1460-9568.2011.07980.x

Costa VD, Tran VL, Turchi J, Averbeck BB. 2015. Reversal learning and dopamine: A Bayesian perspective. *J Neurosci* **35**:2407–2416. doi:10.1523/JNEUROSCI.1989-14.2015

Daw ND, Dayan P. 2014. The algorithmic anatomy of model-based evaluation. *Philos Trans R Soc B Biol Sci* **369**. doi:10.1098/rstb.2013.0478

Daw ND, Kakade S, Dayan P. 2002. Opponent interactions between serotonin and dopamine. *Neural Networks* **15**:603–616. doi:10.1016/S0893-6080(02)00052-7

Daw ND, O’Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006. Cortical substrates for exploratory decisions in humans. *Nature* **441**:876–879. doi:10.1038/nature04766

Dayan P, Daw ND. 2008. Decision theory, reinforcement learning, and the brain. *Cogn Affect Behav Neurosci* **8**:429–453. doi:10.3758/CABN.8.4.429

Donkin C, Brown SD. 2018. Response Times and Decision-Making, Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience. doi:10.1002/9781119170174.epcn509

Donkin C, Brown SD, Heathcote A. 2011. Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *J Math Psychol* **55**:140–151. doi:10.1016/j.jmp.2010.10.001

Donkin C, Brown SD, Heathcote A. 2009. The overconstraint of response time models: Rethinking the scaling problem. *Psychon Bull Rev* **16**:1129–1135. doi:10.3758/PBR.16.6.1129

Dutilh G, Rieskamp J. 2016. Comparing perceptual and preferential decision making. *Psychon Bull Rev* **23**:723–737. doi:10.3758/s13423-015-0941-1

Fontanesi L, Gluth S, Spektor MS, Rieskamp J. 2019a. A reinforcement learning diffusion decision model for value-based decisions. *Psychon Bull Rev*. doi:10.3758/s13423-018-1554-2

Fontanesi L, Palminteri S, Lebreton M. 2019b. Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cogn Affect Behav Neurosci* **19**:490–502. doi:10.3758/s13415-019-00723-1

Forstmann BU, Ratcliff R, Wagenmakers E-J. 2016. Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annu Rev Psychol* **67**:641–666. doi:10.1146/annurev-psych-

122414-033645

Frank MJ. 2004. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science (80-)* **306**:1940–1943. doi:10.1126/science.1102941

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat Neurosci* **12**:1062–1068. doi:10.1038/nn.2342

Gelman A, Rubin DB. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Stat Sci* **7**:457–472. doi:10.1214/ss/1177011136

Gershman SJ. 2015. Do learning rates adapt to the distribution of rewards? *Psychon Bull Rev* **22**:1320–1327. doi:10.3758/s13423-014-0790-3

Haughey HM, Hutchison KE, Curran T, Frank MJ, Moustafa AA. 2007. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc Natl Acad Sci* **104**:16311–16316. doi:10.1073/pnas.0706111104

Hawkins GE, Forstmann BU, Wagenmakers E-J, Ratcliff R, Brown SD. 2015. Revisiting the Evidence for Collapsing Boundaries and Urgency Signals in Perceptual Decision-Making. *J Neurosci* **35**:2476–2484. doi:10.1523/JNEUROSCI.2410-14.2015

Hawkins GE, Heathcote A. 2020. Racing Against The Clock: Evidence-Based Vs. Time-Based Decisions. *Psychol Rev*.

Heathcote A, Lin YS, Reynolds A, Strickland L, Gretton M, Matzke D. 2019. Dynamic models of choice. *Behav Res Methods* **51**:961–985. doi:10.3758/s13428-018-1067-y

Heathcote A, Love J. 2012. Linear deterministic accumulator models of simple choice. *Front Psychol* **3**:1–19. doi:10.3389/fpsyg.2012.00292

Ho T, Brown SD, Van Maanen L, Forstmann BU, Wagenmakers E-J, Serences JT. 2012. The Optimality of Sensory Processing during the Speed-Accuracy Tradeoff. *J Neurosci* **32**:7992–8003. doi:10.1523/JNEUROSCI.0340-12.2012

Izquierdo A, Brigman JL, Radke AK, Rudebeck PH, Holmes A. 2017. The neural basis of reversal learning: An updated perspective. *Neuroscience* **345**:12–26. doi:10.1016/j.neuroscience.2016.03.021

Jang AI, Costa VD, Rudebeck PH, Chudasama Y, Murray EA, Averbeck BB. 2015. The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. *J Neurosci* **35**:11751–11760. doi:10.1523/JNEUROSCI.1594-15.2015

Katsimpokis D, Hawkins GE, Van Maanen L. 2020. Not all Speed-Accuracy Trade-Off Manipulations Have the Same Psychological Effect. *Comput Brain Behav*. doi:10.1007/s42113-020-00074-y

Leite FP, Ratcliff R. 2010. Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, Psychophys* **72**:246–273. doi:10.3758/APP.72.1.246

Logan GD, Van Zandt T, Verbruggen F, Wagenmakers EJ. 2014. On the ability to inhibit thought and action: General and special theories of an act of control. *Psychol Rev* **121**:66–95. doi:10.1037/a0035230

Luzardo A, Alonso E, Mondragón E. 2017. A Rescorla-Wagner drift-diffusion model of conditioning and timing. *PLOS Comput Biol* **13**:e1005796. doi:10.1371/journal.pcbi.1005796

McDougle SD, Collins AGE. 2020. Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychon Bull Rev*. doi:10.3758/s13423-020-01774-z

Miletić S. 2016. Neural Evidence for a Role of Urgency in the Speed-Accuracy Trade-off in Perceptual Decision-Making. *J Neurosci* **36**:5909–5910. doi:10.1523/JNEUROSCI.0894-16.2016

Miletić S, Boag RJ, Forstmann BU. 2020. Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia* **136**. doi:10.1016/j.neuropsychologia.2019.107261

Miletić S, Van Maanen L. 2019. Caution in decision-making under time pressure is mediated by timing ability. *Cogn Psychol* **110**:16–29. doi:10.1016/j.cogpsych.2019.01.002

Millner AJ, Gershman SJ, Nock MK, den Ouden HEM. 2018. Pavlovian Control of Escape and Avoidance. *J Cogn Neurosci* **30**:1379–1390. doi:10.1162/jocn_a_01224

Murphy PR, Boonstra E, Nieuwenhuis S. 2016. Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nat Commun* **7**:1–14. doi:10.1038/ncomms13526

Niv Y, Edlund JA, Dayan P, O'Doherty JP. 2012. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *J Neurosci* **32**:551–562. doi:10.1523/jneurosci.5498-10.2012

O'Doherty JP, Cockburn J, Pauli WM. 2017. Learning, Reward, and Decision Making. *Annu Rev Psychol* **68**:73–100. doi:10.1146/annurev-psych-010416-044216

Pachella RG, Pew RW. 1968. Speed-Accuracy Tradeoff in Reaction Time: Effect of Discrete Criterion Times. *J Exp Psychol* **76**:19–24. doi:10.1037/h0021275

Palminteri S, Khamassi M, Joffily M, Coricelli G. 2015. Contextual modulation of value signals in reward and punishment learning. *Nat Commun* **6**. doi:10.1038/ncomms9096

Palminteri S, Wyart V, Koechlin E. 2017. The Importance of Falsification in Computational Cognitive

Modeling. *Trends Cogn Sci* **21**:425–433. doi:10.1016/j.tics.2017.03.011

Pedersen ML, Frank MJ. 2020. Simultaneous Hierarchical Bayesian Parameter Estimation for Reinforcement Learning and Drift Diffusion Models: a Tutorial and Links to Neural Data. *Comput Brain Behav*. doi:10.1007/s42113-020-00084-w

Pedersen ML, Frank MJ, Biele G. 2017. The drift diffusion model as the choice rule in reinforcement learning. *Psychon Bull Rev* **24**:1234–1251. doi:10.3758/s13423-016-1199-y

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: Experiments in behavior made easy. *Behav Res Methods* **51**:195–203. doi:10.3758/s13428-018-01193-y

Purcell BA, Heitz RP, Cohen JY, Schall JD, Logan GD, Palmeri TJ. 2010. Neurally constrained modeling of perceptual decision making. *Psychol Rev* **117**:1113–1143. doi:10.1037/a0020311

Rae B, Heathcote A, Donkin C, Averell L, Brown SD. 2014. The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *J Exp Psychol Learn Mem Cogn* 1–39. doi:10.1037/a0036801

Ratcliff R. 1978. A theory of memory retrieval. *Psychol Rev* **85**:59–108.

Ratcliff R, Hasegawa YT, Hasegawa RP, Childers R, Smith PL, Segraves MA. 2011. Inhibition in superior colliculus neurons in a brightness discrimination task? *Neural Comput* **23**:1790–1820. doi:10.1162/NECO_a_00135

Ratcliff R, Hasegawa YT, Hasegawa RP, Smith PL, Segraves MA. 2007. Dual Diffusion Model for Single-Cell Recording Data From the Superior Colliculus in a Brightness-Discrimination Task. *J Neurophysiol* **97**:1756–1774. doi:10.1152/jn.00393.2006

Ratcliff R, McKoon G. 2008. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput* **20**:873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff R, Rouder JN. 1998. Modeling Response Times for Two-Choice Decisions. *Psychol Sci* **9**:347–356.

Ratcliff R, Smith PL, Brown SD, McKoon G. 2016. Diffusion Decision Model: Current Issues and History. *Trends Cogn Sci* **20**:260–281. doi:10.1016/j.tics.2016.01.007

Ratcliff R, Voskuilen C, Teodorescu A. 2018. Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cogn Psychol* **103**:1–22. doi:10.1016/j.cogpsych.2018.02.002

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class Cond II Curr Res Theory* **21**:64–99. doi:10.1101/gr.110528.110

Rummery GA, Niranjan M. 1994. On-Line Q-Learning Using Connectionist Systems.

Sewell DK, Jach HK, Boag RJ, Van Heer CA. 2019. Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychon Bull Rev*. doi:10.3758/s13423-019-01570-4

Sewell DK, Stallman A. 2020. Modeling the Effect of Speed Emphasis in Probabilistic Category Learning. *Comput Brain Behav* **3**:129–152. doi:10.1007/s42113-019-00067-6

Shahar N, Hauser TU, Moutoussis M, Moran R, Keramati M, Consortium NSPN, Dolan RJ. 2019. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Comput Biol* **15**:1–25. doi:10.1371/journal.pcbi.1006803

Spektor MS, Kellen D. 2018. The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychon Bull Rev* **25**:2047–2068. doi:10.3758/s13423-018-1446-5

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. 2002. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Statistical Methodol* **64**:583–639.

Sutton, Richard S. 1988. Learning to Predict by the Method of Temporal Differences. *Mach Learn* **3**:9–44. doi:10.1023/A:1018056104778

Sutton RS, Barto AG. 2018. Reinforcement Learning: An Introduction, 2nd ed, MIT Press. Cambridge, MA: MIT press.

Teodorescu AR, Moran R, Usher M. 2016. Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychon Bull Rev* **23**:22–38. doi:10.3758/s13423-015-0858-8

Ter Braak CJF. 2006. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Stat Comput* **16**:239–249. doi:10.1007/s11222-006-8769-1

Thura D, Cisek P. 2016. Modulation of Premotor and Primary Motor Cortical Activity during Volitional Adjustments of Speed-Accuracy Trade-Offs. *J Neurosci* **36**:938–956. doi:10.1523/JNEUROSCI.2230-15.2016

Tillman G, Van Zandt T, Logan GD. 2020. Sequential sampling models without random between-trial variability: the racing diffusion model of speeded decision making. *Psychon Bull Rev*. doi:10.3758/s13423-020-01719-6

Trueblood JS, Heathcote A, Evans NJ, Holmes WR. 2020. Urgency, Leakage, and the Relative Nature of Information Processing in Decision-making. *Psychol Rev* 706291. doi:10.1101/706291

Turner BM. 2019. Toward a Common Representational Framework for Adaptation. *Psychol Rev*. doi:10.1037/rev0000148

Turner BM, Sederberg PB, Brown SD, Steyvers M. 2013. A method for efficiently sampling from distributions with correlated dimensions. *Psychol Methods* **18**:368–384. doi:10.1037/a0032222

van Maanen L, Miletić S. 2020. The interpretation of behavior-model correlations in unidentified cognitive models. *Psychon Bull Rev*. doi:10.3758/s13423-020-01783-y

van Maanen L, van der Mijn R, van Beurden MHPH, Roijendijk LMM, Kingma BRM, Miletić S, van Rijn H. 2019. Core body temperature speeds up temporal processing and choice behavior under deadlines. *Sci Rep* **9**:10053. doi:10.1038/s41598-019-46073-3

Van Ravenzwaaij D, Brown SD, Marley AAJ, Heathcote A. 2020. Accumulating advantages: A new conceptualization of rapid multiple choice. *Psychol Rev* **127**:186–215. doi:10.1037/rev0000166

Voss A, Nagler M, Lerche V. 2013. Diffusion models in experimental psychology: A practical introduction. *Exp Psychol* **60**:385–402. doi:10.1027/1618-3169/a000218

Voss A, Rothermund K, Voss J. 2004. Interpreting the parameters of the diffusion model: An empirical validation. *Mem Cognit* **32**:1206–1220. doi:10.3758/BF03196893