Reinforcement learning under uncertainty: expected versus unexpected uncertainty and state versus reward uncertainty

Adnane Ez-zizi[a,*], Simon Farrell[b], David Leslie[c], Gaurav Malhotra[d], Casimir J.H. Ludwig[d]

[a] *School of Engineering, Arts, Science and Technology, University of Suffolk, Ipswich, UK*
[b] *School of Psychological Science, University of Western Australia, Perth, Australia*
[c] *Department of Mathematics and Statistics, Lancaster University, Lancaster, UK*
[d] *School of Psychological Science, University of Bristol, Bristol, UK*

Abstract

Two prominent types of uncertainty that have been studied extensively are expected and unexpected uncertainty. Studies suggest that humans are capable of learning from reward under both expected and unexpected uncertainty when the source of variability is the reward. How do people learn when the source of uncertainty is the environment's state and the rewards themselves are deterministic? How does their learning compare with the case of reward uncertainty? The present study addressed these questions using behavioural experimentation and computational modelling. Experiment 1 showed that human subjects were generally able to use reward feedback to successfully learn the task rules under state uncertainty, and were able to detect a non-signalled reversal of stimulus-response contingencies. Experiment 2, which combined all four types of uncertainties—expected versus unexpected uncertainty, and state versus reward uncertainty—highlighted key similarities and differences in learning between state and reward uncertainties. We found that subjects performed significantly better in the state uncertainty condition, primarily because they explored less and improved their state disambiguation. We also show that a simple reinforcement learning mechanism that ignores state uncertainty and updates the state-action value of only the identified state accounted for the behavioural data better than both a Bayesian reinforcement learning model that keeps track of belief states and a model that acts based on sampling from past experiences. Our findings suggest a common mechanism supports reward-based learning under state and reward uncertainty.

Keywords: expected and unexpected uncertainty, reinforcement learning, Bayesian reinforcement learning, sampling-based learning

1. Introduction

Humans are capable of dealing with different types of uncertainty when learning (Soltani & Izquierdo, 2019) and making decisions (Platt & Huettel, 2008). Two forms of uncertainty that have attracted interest in the psychology and neuroscience literature include reward uncertainty and state uncertainty. Reward uncertainty occurs when reward outcomes are generated stochastically from a probability distribution. State uncertainty (also often referred to as perceptual uncertainty; Bruckner et al., 2020) arises when the agent cannot tell for sure what the current state of the world is: its perceptual system might be noisy (e.g., noise in the brain or in sensors);

*Corresponding author
Email address: a.ez-zizi@uos.ac.uk (Adnane Ez-zizi)

the observations are ambiguous (e.g. reading road signs in challenging weather conditions) or the information about the state is incomplete (e.g. playing card games without knowing the cards dealt to the other opponents or their play strategies).

In addition, uncertainty can also be categorised into expected uncertainty and unexpected uncertainty. Uncertainty is considered expected when the rules governing the environment are stochastic but predictable, due, for example, to rewards being generated from a stable probability distribution (Bland & Schaefer, 2012). This definition covers both the case where the uncertainty arises from reward outcomes (we refer to this here as expected reward uncertainty; also sometimes called risk as, for example, in Bruckner et al., 2022) and the case where the uncertainty arises from the environment's states (called here expected state uncertainty; though in some previous work, expected uncertainty has been equated with stochasticity arising from reward outcomes as in Soltani & Izquierdo, 2019). Though definitions of unexpected uncertainty differ, this type of uncertainty refers to the case where there is a sudden and fundamental change in the environment such as when there is a non-signalled change in the rules of the task, which invalidates the previous acquired rules (Bland & Schaefer, 2012). Another well-studied uncertainty type that is closely related to unexpected uncertainty but fundamentally different is volatility. Volatility occurs when there are frequent non-signalled changes and the frequency of these changes varies across time. More precisely, it captures the variance in the frequency of fundamental changes (Bland & Schaefer, 2012; Piray & Daw, 2021).

Several authors have proposed psychological and neural accounts of reward-based learning under uncertainty when the source of uncertainty lies in the reward outcomes (i.e., when rewards are stochastic; for reviews, see Bland & Schaefer, 2012; Soltani & Izquierdo, 2019; Bruckner et al., 2022). For instance, Yu & Dayan (2005) developed an influential account of the computation of expected and unexpected uncertainty in reward outcomes, where the signalling of these forms of uncertainty is mediated by different neurotransmitter systems (acetylcholine and norepinephrine for expected and unexpected uncertainty, respectively). Other authors have assessed the extent to which people adjust their learning rate according to the volatility of the reward environment, and compared human behaviour to that of an (approximately) optimal Bayesian learner (Behrens et al., 2007; Nassar et al., 2010; Mathys et al., 2011; for a more recent theoretical work, see also Piray & Daw, 2020, 2021).

While such studies have provided important insights into the psychological and neural responses to *reward* uncertainty, little is known about how people learn from reward when the source of uncertainty is the environment's state and rewards themselves are deterministic (though see Bruckner et al., 2020, for a recent attempt to explore reward-based learning under state uncertainty). This situation arises when the agent cannot tell for sure what the current state of the world is, and presents the agent with a difficult challenge: should an unexpected reward (or lack of an expected reward) be used to update action values for what the agent believes to be the current environmental state, or does the unexpected reward outcome carry information about the accuracy of identification of the uncertain state itself? Unexpected uncertainty (e.g. in the form of an unexpected change in the state-action mappings) further adds a challenge: an unexpected reward outcome may be down to identifying the state incorrectly or to a change in the environment.

To illustrate, consider an animal that has to decide which food patch to forage in. The animal needs to quickly assess the nutritional value and predation risk of the patch based on uncertain sensory cues (whether visual, auditory or olfactory), even before feeding begins (Mella et al., 2018). For simplicity, say the animal has to learn to choose between two types of patches, A or B, each of which provides reward with a fixed probability (expected reward uncertainty). Assume that patch A is, on average, more rewarding (e.g. more nutritious or safer) for the animal than

patch B and that the animal can only detect the type of patch based on the sensory cues with a certain fixed probability (expected state uncertainty). Framing this as a reinforcement learning problem, the animal's task here would be to learn the value of each patch type (i.e. its long term expected reward, also known as state-value) so that it can maximise its reward from the food patch visits.

One main problem that the animal faces under (expected) state uncertainty conditions is: which state should an obtained reward be assigned to when updating the state-values? Should it be assigned to the patch type that was perceived by the animal as the true state, or to both patch types proportionally to the animal's beliefs about the true state? This problem is compounded under unexpected uncertainty. For instance, suppose there is an unexpected change in the reward distributions of the two food patches, so patch B becomes better than patch A (perhaps patches of type A suddenly start attracting more predators or have become less nutritious due to some special environmental conditions)? Now the lack of an expected reward may be down to either identifying the state incorrectly, or to a change in the environment so that the state-action mapping has changed. How can the animal learn the correct state-action mapping under expected state uncertainty and adapt to the unexpected uncertainty arising from a change in the state-action mapping? How can RL models account for learning under these forms of uncertainty?

Larsen et al. (2010) proposed two plausible models for how agents might learn from reward under state uncertainty. One model was a simple RL model that assumes that people commit to one state being the true state (e.g. animal identifying one patch type as the true state based on uncertain sensory cues), then update the state-action value of only that identified state. Larsen et al. (2010) also suggested a more complex Bayesian RL model where the learner is assumed to update the state-action value of each possible state according to the posterior probability of that state being the true state, given the observed reward. The main difference between these two models is that the Bayesian approach uses the current state-action values along with the observed reward to help inform the beliefs about the true state, whereas the simple model just uses the state information from the environment to make a best guess about the true state. Accordingly, under the simple model, only one state-action value will be updated, whereas the Bayesian model can use the observed reward to update information about all states.

Larsen and colleagues showed that the two models behave differently under expected and unexpected uncertainty. Under expected uncertainty, Larsen et al. proved that the Bayesian model converges to the correct state-action values, whereas the non-Bayesian model converged to incorrect values unless the expected rewards for the different state-action pairs were all equal. Under unexpected uncertainty—such as a sudden reversal of stimulus-response contingencies—the simple RL model adapted to the environmental change by updating its state-action value estimates to match the environmental change (even though these value estimates were not correct), while the Bayesian RL model did not (unless the state uncertainty was made small; see Larsen et al., 2010, for further details).

Both models are candidates for model-free reward-based learning under the different types of uncertainty we have reviewed above. Following on from Larsen et al.'s (2010) work and predictions, we had several aims in this study. First, we wanted to understand whether human participants use reward feedback to update both their state estimates and state-action mapping—as stipulated by the Bayesian RL model—or whether they update the state-action mapping for only the estimated state, as assumed in a simple RL model. Second, if participants behave in line with the Bayesian RL model, they are likely to fail to adapt to unexpected uncertainty (i.e. a change in the state-action mapping). Therefore, we wanted to see whether participants will successfully adapt to unexpected uncertainty. Third, we wanted to assess people's learning under state uncertainty to that under reward uncertainty, allowing a direct comparison of how people

deal with expected and unexpected uncertainty in states versus rewards.

A final aim was to compare the two RL models to another type of model that has emerged as a plausible mechanism for learning under uncertainty: models that act based on sampling from past experiences. Sampling models assume that the learner chooses actions that have produced the best payoff in small samples of past trials, usually taken from the recent past (e.g., Stewart et al., 2006; Chen et al., 2011). Recent research has suggested a superiority of such models over classic RL models in explaining participants' learning behaviour in tasks involving reward uncertainty (Hochman & Erev, 2013; Plonsky et al., 2015; Bornstein et al., 2017; Bornstein & Norman, 2017). For instance, Bornstein et al. (2017) developed a simple sampling model that assumes that the decision maker computes the values of the different choice options by using one outcome sampled from the past. This model fit participants' choices and neural signals better than a classic temporal difference-based RL model. They also showed that the sampling process could be biased by reminding participants of particular trial outcomes, which cannot be accounted for under the RL framework.

In considering the sampling models, we also entertained a variety of Bayesian (approximate-) normative models of learning under uncertainty. These include, for example, the hierarchical Bayesian model of Behrens et al. (2007), the approximate-normative Bayesian model of Payzan-LeNestour & Bossaerts (2011), the hierarchical Gaussian filter of Mathys et al. (2014) and the more recent Kalman-based models of Piray & Daw (2020, 2021). Although these models are relevant to the study of human learning under uncertainty, these have been built primarily to study learning under volatility. Given that our manipulation of unexpected uncertainty takes the form of a single change in the state-action mapping, we do not consider volatility in this paper.

We set up two experiments to investigate how people deal with expected and unexpected state and reward uncertainty. In the first experiment, we focus on state uncertainty, and present a psychophysical reward task in which the states are signalled imperfectly with noisy stimuli so that participants cannot identify the true state with certainty. The states perfectly determine the rewards delivered by different actions. This paradigm allows us to test whether participants can learn the correct rules under this state uncertainty condition. We also introduce a switch (i.e., reversal) of stimulus-response contingencies for some participants without telling them, to see if they can adapt to the unsignalled change (unexpected uncertainty) despite the presence of state uncertainty. In the second experiment, we tackle the question of whether learning under state uncertainty and reward uncertainty is supported by the same mechanism, by adding a condition in which the source of uncertainty is the reward signal instead of the state cues (i.e. the reward function is stochastic while states are deterministic).

We used these data to compare the three computational learning approaches—Bayesian RL, non-Bayesian RL and sampling models—in their account of the learning of different participants. Based on the theoretical predictions from Larsen et al. (2010) it might be expected that in the state uncertainty condition, the pattern of learning of the participants who adapt to the switch will be better described by the simple RL in comparison with the Bayesian model, while the learning behaviour of those who fail to adapt to the switch, or adapt only slowly, will be better explained by the Bayesian RL model. We expected the sampling model of Chen et al. (2011) to be able to handle a variety of adaptation patterns due to its greater flexibility (at the expense of increased computational complexity).

The contribution of this work is two-fold. First, our work offers behavioural and computational insights into how people learn under conditions of (expected) state uncertainty. Second, we compared how people learn under expected and unexpected uncertainty when the source of uncertainty lies in the reward versus state. That is, we assessed whether the mechanisms for

dealing with expected and unexpected reward uncertainty, generalise to situations in which the uncertainty arises from the state.

2. Experiment 1

In most previous demonstrations of reward-based learning under expected and unexpected uncertainty, the stimuli were perceptually unambiguous, while the reward was uncertain (being drawn from a probability distribution). In contrast, in Experiment 1, we paired a deterministic reward function with ambiguous stimuli to study the effect of a rule reversal on the ability to perform reward-based learning under state uncertainty.

2.1. Participants and design

Twenty-four participants were tested in Experiment 1. Participants were students from the University of Bristol, or other volunteers recruited using the University email announcement system. Twelve of the participants served as a control group for whom we did not switch the state-action mapping rules; for the remaining twelve participants we switched the mapping half way through the task. All participants had normal or corrected-to-normal vision. Participants were reimbursed a fixed amount of £4 for participation, and in addition earned a variable amount of up to £5, which depended on their performance on all trials (£0.025 per correct response). The study was approved by the Faculty of Science Human Research Ethics Committee at the University of Bristol.

2.2. Material and procedures

2.2.1. Experimental setup and stimuli

Participants were seated in front of a 21-inch Viewsonic G225fB monitor with a resolution of 1024×768 resolution and a refresh rate of 85 Hz. The experiment was programmed and run using Matlab with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). The viewing distance was set to ~ 57 cm.

In each trial, participants saw a patch filled with a variable proportion of white and black dots, in the middle of a gray background (see Figure 1). When the majority of the dots were white, the patch appeared lighter than the background ("light" state). On the other hand, when the majority of the dots were black, the patch was darker than the background ("dark" state). More precisely, the patch contained 400 by 400 white or black pixels ($15 \times 15°$. The luminances for the black, grey and white pixels were respectively equal to $0.1cd/m^2$, $42cd/m^2$ and $85cd/m^2$, so that the black and white pixels were approximately equidistant from the background luminance.

As the stimulus was noisy, the learner could only identify the correct state with a certain probability ρ called level of state uncertainty. In our experiment, we aimed to have ρ around 0.65. Due to individual variation in visual sensitivity, we conducted a preliminary measurement to calibrate the proportion of light and dark pixels in the stimuli per participant, so that performance was at the desired level of ρ.

2.2.2. Procedure

The procedure was the same for the control and test group, except that for the control group, we did not switch the state-action mapping rules at any point. The control condition was set up to examine whether people are able to perform reward-based learning under state uncertainty (similarly, results from the test group before the switch can be used to answer this question), whereas the test condition should inform us about people's ability to respond to an unsignalled reversal in state-action mapping rules in the presence of state uncertainty, and how this change

would affect their learning performance. Participants from both groups completed two phases: a calibration phase and a learning phase.

Calibration phase

This phase aimed to determine, for each participant, the proportion of white/black pixels that corresponds to the level of state uncertainty $\rho \approx 0.65$. That is, we aimed to identify the mixture of dark and light pixels that the participant could accurately identify with an accuracy of 65%). For that, we measured a full psychometric function—that is, the proportion of "light" responses as a function of the proportion of white dots—using the method of constant stimuli with 9 proportions of white dots that varied between 0.4 and 0.6 (0.4 corresponding to an easily identifiable dark state, and 0.6 corresponding to an easily identifiable light state) spaced at 0.02 intervals (Kingdom & Prins, 2009). There were 25 presentations at each of the nine contrast levels, for a total of 225 trials. Note that the positions of the white and black pixels varied randomly from trial to trial, hence two stimuli can have the same proportion of white and black dots, but might appear different.

In each trial (Figure 1), after being presented with the stimulus, participants used the mouse to click on one of two shapes (triangle or rectangle), depending on whether they believed that the state of the patch was light or dark. The (initial) mapping of shape to light/dark response was counterbalanced across participants, so that half of the participants were asked to click on the triangle if they thought the patch was light and the rectangle if the patch was dark; the other half had to use the reversed mapping. To minimise spatial response bias, on each trial the two response stimuli were positioned randomly on a virtual circle and were always on two opposing circle quarters, with the mouse cursor initially placed at the centre of the circle.

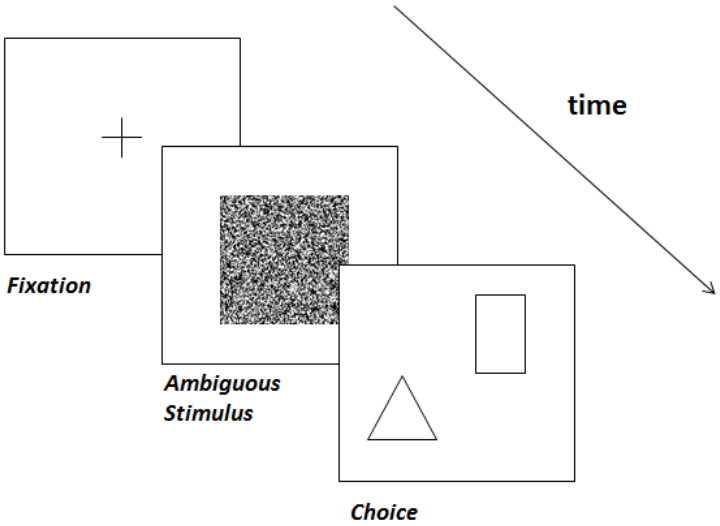

Figure 1: Sequence in one trial from the calibration phase. Firstly, the subject saw a fixation cross on the centre of the screen for 506 ms (43 frames). Then the stimulus was displayed for 247 ms (21 frames). After that, the subject had to choose between two shapes depending on the state of the patch. The next trial started following a blank screen that was displayed for 506 ms (43 frames). The centres of the response shapes—equilateral triangle of side length 175 px (a visual angle of 7°) or a rectangle of height 350 px (13°) and width 175 px (7°)— were positioned randomly on a virtual circle of radius 250 px (10°).

Once participants had completed all 225 trials, a cumulative Gaussian psychometric function was fit to the obtained frequencies of light responses for each participant using the Palamedes Toolbox in Matlab (Prins & Kingdom, 2018); an example is shown in Figure 2. Using each

participant's fitted function, proportions of white dots that corresponded to the level of state uncertainty $\rho = 0.65$ for both the light and dark state were computed (for the dark state, this was obtained from the figure by computing the proportion of white dots that corresponded to a frequency of light response equal to $1 - \rho$).

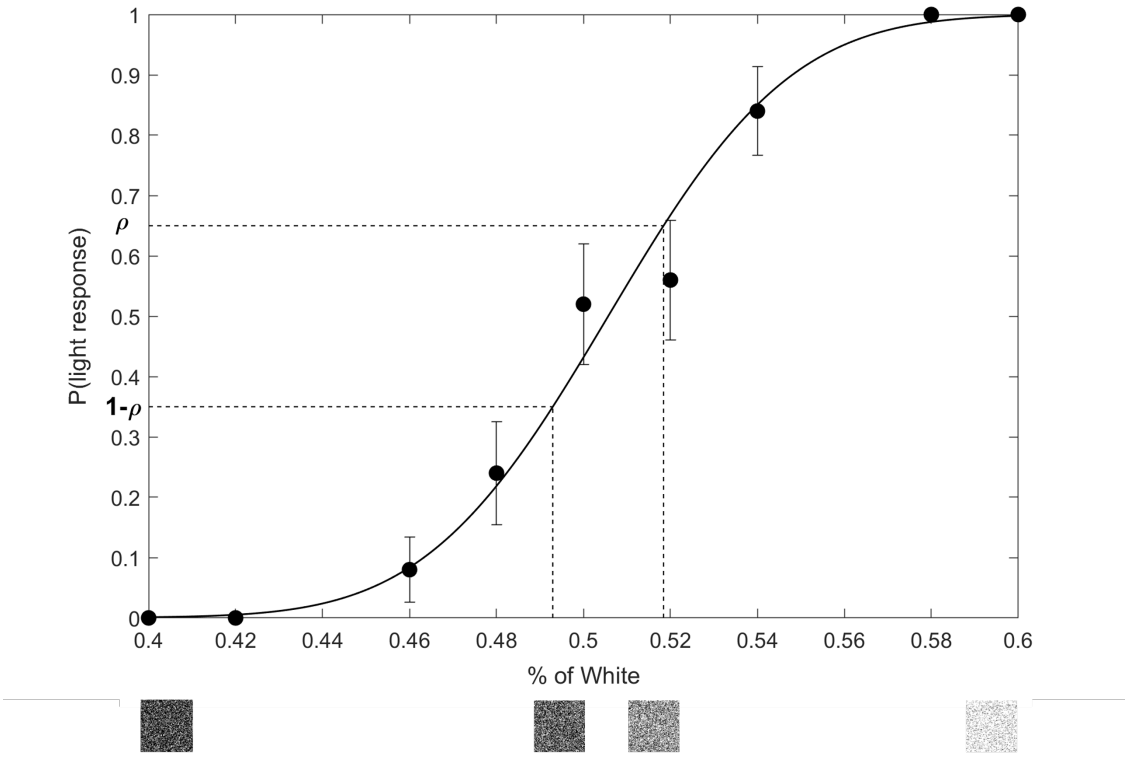

Figure 2: Example data and fitted psychometric function. Frequency of light responses as a function of the proportion of white dots for nine contrast levels. On the x-axis, as the percentage of white dots increases, the stimulus would appear lighter to participants. The error bars represent binomial standard error bars. The solid curve shows the fitted psychometric function. The dashed lines illustrate how the ρ-thresholds for both the light and dark states are computed using the the fitted function.

Learning phase

The learning phase was used to train participants on the state-value associations. A schematic trial sequence is shown in Figure 3. Presentation of the stimulus and choice options were as in the calibration phase. In the learning phase, however, participants were not told about the correct mapping between light or dark states and the two response options, so the mapping needed to be learned. In order to avoid any interference from the state-action mapping used in the calibration phase, the new response options were a circle and square instead of rectangle and triangle (the mapping of shape to correct response was counterbalanced across participants, as in the calibration phase). Each trial provided a reward opportunity: in trials where the subject chose the correct response, they received positive reward feedback (£0.025) in the form of a moneybag icon that appeared for 1.52 s (129 frames). In trials where the incorrect response was selected, no reward feedback was given and the next trial was presented immediately after the response and the inter-trial interval of 506 ms (43 frames). Subjects were assigned the task of maximising their rewards, and at the end of the experiment were paid the total accumulated

reward across all trials in addition to their reimbursement. The learning phase contained 200 trials.

Participants in the control condition were exposed to the same reward mapping throughout the task. In the test condition, the mapping rules were switched (without signalling this change to participants) at the 101st trial, so that if the correct (reward-generating) response for the light state (resp., dark state) before the switch was circle (resp., square), after the switch the correct response would become square (resp., circle). Participants were not given any instructions about whether or not a switch might occur prior to the start of the experiment.

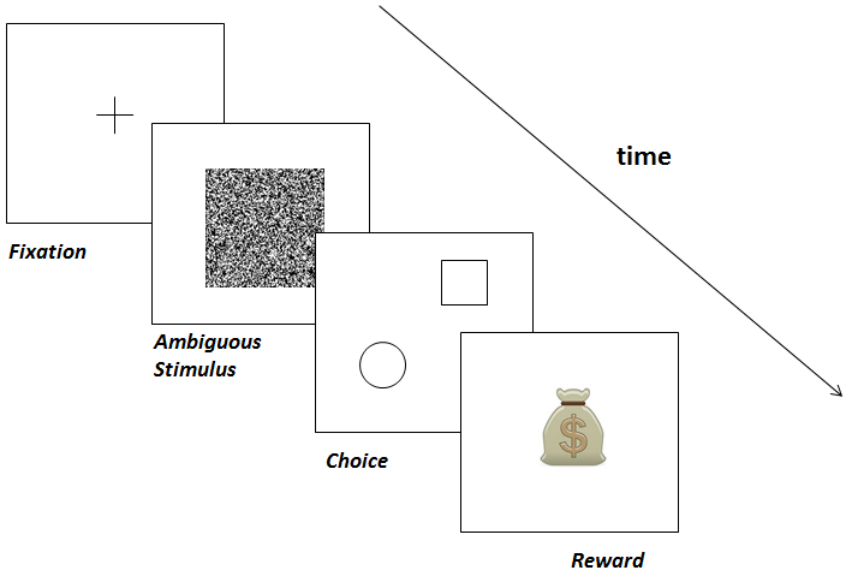

Figure 3: The sequence in one trial from the learning phase looked similar to that from the calibration phase. The main difference here is that participants received a positive reward feedback (winning £0.025) if they responded correctly. Another difference is that the response shapes were changed to a circle and a square instead of a triangle and rectangle. The circle's diameter and square's side length were both equal to 175px (i.e. within a visual angle of 7°).

2.3. Analysis

2.3.1. Learning curves

To visually evaluate the dynamics of participants' learning performance and their ability to adapt to the switch in the mapping rules, we calculated a learning curve for each participant by computing a moving average of the proportion of rewarded responses with a window size of 30 trials over the course of the learning phase (i.e., the averaging was performed for every 30 trials on the binary variable that encodes whether a response was rewarded or not). Since the learning phase consisted of 200 trials, there are 171 values of average values for each participant, corresponding to the 171 windows of length 30, and the sequence of these 171 values is henceforth called the participant's learning curve.

2.3.2. Statistical analyses

To evaluate participants' learning statistically and understand what factors influenced their learning performance, we used generalised additive mixed modelling.

Generalised additive models (GAMs; Hastie & Tibshirani, 1990; Wood, 2017) are the natural choice for modelling a response variable that varies in a complex non-linear fashion with time

as is the case for our data. These models can adapt to a much wider range of shapes than polynomial regression, with the latter being restrictive, especially with regards to the location of the inflexion points (for example, in cubic polynomial regression the location of the first inflexion point will often dictate the location of the other inflexion point). This is an important consideration to take into account for our data since for a participant who adapts to the switch, the fitted curve should have an inflection point around the 101st trial to account for the decrease in performance after the switch, while the location of the second inflection point could be at any of the next 100 trials depending on how fast the participant adapts to the switch. GAMs overcome this restriction by dividing the data range into smaller intervals then applying local polynomial regression separately in each of those intervals.

Concretely, the regression equation in a GAM can be formally expressed as:

$$g(E(Y)) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p) + \epsilon \tag{1}$$

where Y is the response variable, which follows a distribution from the exponential family (e.g., binomial distribution), $E()$ is the expected value operator, $X_1, ..., X_p$ are the explanatory variables, g is a link function (e.g. logistic), β_0 is the intercept, ϵ is a random error that is assumed to have a constant variance and a null mean. $f_1, ..., f_p$ are unknown smooth functions to be estimated from the data along with β_0 as would be the case in a standard generalised linear model (GLM). Each smooth function—also called a smoothing spline—is composed of a weighted sum of basis functions:

$$f(X) = \beta_1 b_1(X) + \cdots + \beta_q b_q(X) \tag{2}$$

where the basis functions are piecewise-defined polynomial functions. An example of smoothing bases are cubic regression splines which are formed by connecting polynomial segments of degree 3. The points where the segments connect are called knots, whose number determines the wiggliness of the fitted curve and can be set either manually or using cross validation with the objective of minimizing a penalized residual sum of squares.

Our analysis also required the inclusion of random effects since our data contained repeated measurements of the same participants at multiple trials. GAMs, similarly to GLMs, can be extended with random effects with even more flexibility with regards to how they can be captured. Our generalised additive mixed effects models (GAMMs) thus contained by-participant factor smooths for *trial order* (the non-linear counterpart to random slopes and random intercepts). We started with a saturated model that contained all fixed effects of interest, including their interactions, as well as the (random) by-participant factor smooths for trial order. From the saturated model, we sequentially removed fixed effects using a backward-selection approach based on the Akaike Information Criterion (AIC), that is, we removed variables that led to the largest decrease in AIC and we stopped when the AIC could no longer be improved.

The fixed structure of the statistical model incorporated the *experimental condition* factor (control versus test condition); and *state match* factor: whether or not the current (true) state matches the previous (true) state. We introduced *state match* as it generates a prediction that could help discern between our computational models. Under the simple RL framework (and similarly for the sampling model), updates occur at the level of state-action pairs (see Section 4.1.2 for more details), which should result in the dependence of reward acquisition on the nature of the previous state. This is because visiting a state and receiving reward feedback on that visit is likely to improve the learner's estimate of the action values associated with that state, which in turn, should promote the selection of the most rewarding action on the next visit of that state. In contrast, if the state is not visited, the learner's knowledge about how to deal with that state is not updated, and hence no improvement should be observed. Under the Bayesian RL framework,

state-action values of both states are updated (according to their posterior probability of being the true state; more details are provided in Section 4.1.1). Here, the state match effect should be attenuated since reward information—which can potentially improve learning—will flow to the next state irrespective of whether the next state matches the previous one. In other words, under the Bayesian RL model, the positive effect of re-encountering a state on learning performance is likely to be weaker than under the simple RL model because updates will take place for all states.

To account for the possible interaction between experimental condition, state match and trial order, we included the non-linear interaction by-(condition×state match) factor smooth for trial order [1]. We also added the factor *reward acquisition at t-1* (whether or not reward was received in the previous trial), primarily to remove autocorrelation in the residuals, but also to assess whether learning effectively took place (a positive effect would provide support for this). The GAMMs were run in R version 4.1.2 (R Core Team, 2021) using the *bam()* function from the mgcv package version 1.8-40 (Wood, 2017). Interaction graphics were generated using the itsadug package version 2.4.1 (van Rij et al., 2020).

2.4. Results

We first analysed the behavioural results to see (1) how well participants learned the task in the control condition (and test condition prior to the switch) given the expected state uncertainty, and (2) whether they managed to adapt to the unexpected uncertainty in the test condition despite the presence of state uncertainty. We then performed a more in-depth analysis using GAMMs to assess participants' learning statistically and understand the impact of the unexpected reversal and other factors on their learning performance.

2.4.1. Learning performance under expected uncertainty

Figure 4 shows that participants from both the control and test (before the switch) group performed, in general, at the level of the state uncertainty, and that they were able to learn the correct state-action mapping despite the presence of state uncertainty.

Learning was confirmed at the individual level when we compared the proportion of rewarded responses on the first block of 20 trials to the proportion of rewarded responses on the last block of 20 trials preceding the "switch trial" (i.e., the 101^{th} trial in both conditions). In fact, as can be seen from Figure 5 (left pane), before the "switch trial", 20 out of 24 participants from both groups showed improved performance by the end of the first 100 trials, as indicated by points above the equality line in the figure (note also that the performance of one participant did not increase in the last 20 trials because their reward acquisition rate was already above 75% in the first 20 trials). Also, 20 participants reached a reward acquisition rate above chance in the last block of 20 trials preceding the switch. This suggests that the majority of our participants not only managed to learn the task (without the switch), but were able to do so within 100 trials.

2.4.2. Learning performance under unexpected uncertainty

Participants were also able, in general, to adapt to the unsignalled change in state-action mapping as illustrated in Figure 4 (solid red learning curve). As expected, reward proportion dropped to the level $1 - \rho$ after the switch, and seems to have recovered slowly to arrive again at a level near of the level of state uncertainty $\rho = 0.65$. Figure 5 (right pane) indicates that

[1] Effectively, we coded the three-way non-linear interaction as a two-way non-linear interaction by combining condition and state match into a single variable that we denote as *condition & state match*, since three-way non-linear interactions are not directly supported in the software implementation of GAMMs that we used

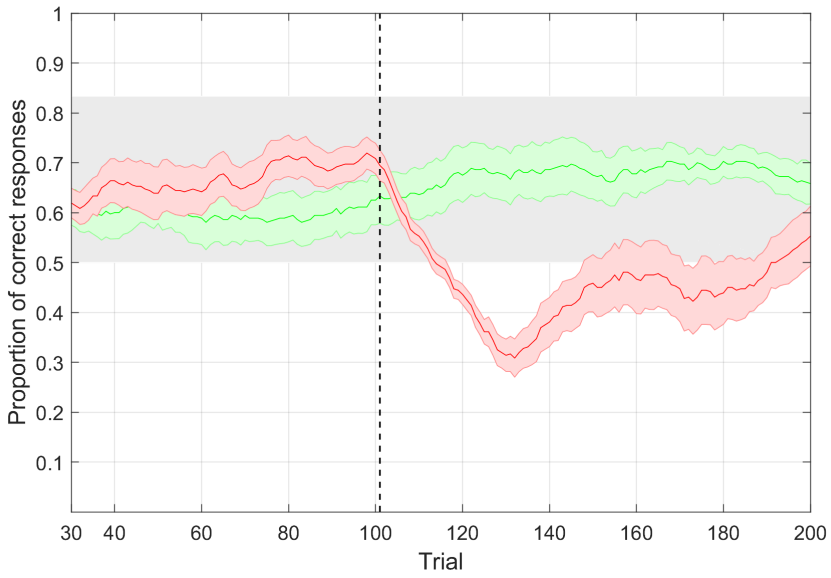

Figure 4: Running average score of the "average subject" in the control (solid green line) and test group (solid red line) with a window size of 30. The shading along each curve represents the standard error of the mean, while the light grey region represents an approximate 95% confidence interval (CI) of the moving average of a Bernoulli variable with probability of success equal to 0.65. The CI is constructed to give an indication of the normal range of performance in each window of 30 trials for a subject who has learned the task rules (it is a simple measure to assess learning/unlearning for each point in isolation). A point in the learning curve that is below the CI range suggests that it is unlikely that the (aggregated) participant was using the correct mapping during that window, while a point above the CI range suggests that the participant was performing at a higher level than the target level of uncertainty. The vertical dashed line indicates the trial where the switch occurs in the test condition.

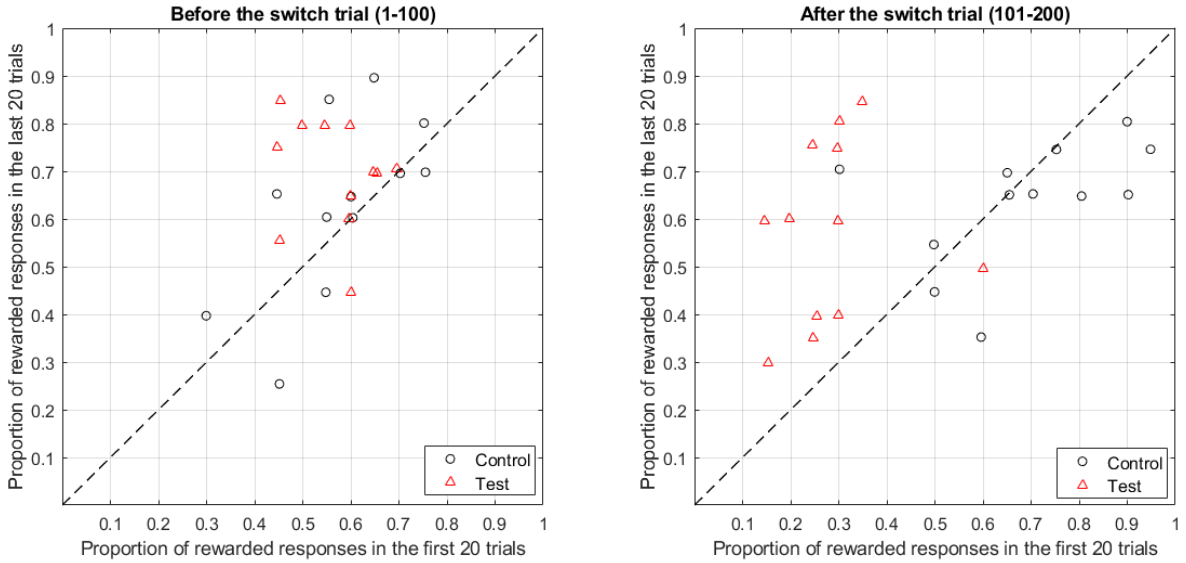

Figure 5: Comparison of the average proportion of rewarded responses in the first 20 trials and last 20 trials before the "switch trial" (left) and after it (right), for both the control and test group. A jitter of 0.005 was applied to enhance the visualisations.

most participants from the test group (11 out of 12) performed better in the last 20 trials in comparison with the first 20 trials following the switch; however, only 8 participants reached a performance above chance by the end of the testing sequence, and hence could be counted as successful in dealing with unexpected uncertainty.

Individual differences were evident in how participants adapted to the switch. For example, some participants adapted quickly to the switch (within about 30 trials), while other participants adapted late or did not adapt at all to the switch even though they managed to learn the task rules before the switch (see S.1 for a full analysis that constructs individual learning curves and groups them using hierarchical clustering). In addition, a few participants seemed to have performed at a higher level than the targeted level of uncertainty $\rho = 0.65$ for extended periods of time (see, for example, Participants 7 and 8 in Figure S.2).

2.4.3. Analysis of learning performance using statistical modelling

The GAMM model for reward acquisition that we retained after backward variable selection is presented in Table 1. The proportion of rewarded responses increased over time prior to the switch in the test condition, suggesting a learning trend as was expected (Figure 6A). A positive trend also occurred in the control condition but did not reach significance ($p = .104$), attributable to the relatively small sample size in each group. In fact, re-running the same model with data from both conditions restricted to the first 100 trials showed a significant increase with trial order ($edf = 1$, $p = .003$). Obtaining reward in the previous trial increased the likelihood of obtaining reward in the current trial, also suggesting that (recency-based) learning took place ($\beta = 0.16$, $SE = 0.07$, $z = 2.40$, $p = .016$). Reward likelihood was significantly lower in the test condition than in the control condition for most of the second block of trials (between the 102th and 188th trials as highlighted with the red segment in Figure 6B), confirming that the introduction of the switch negatively impacted participants' learning performance, and that, overall, the recovery from the switch effect occurred gradually and slowly. There was a very strong positive effect of the state match factor ($\beta = 0.71$, $SE = 0.06$, $z = -11.09$, $p < .001$), indicating that participants were more likely to get rewarded—in other words, respond correctly—when they were presented with the same state as the previous trial.

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	0.19	0.14	1.35	.178
Condition = test	−0.40	0.18	−2.22	.026
State match = same	0.71	0.06	−11.09	< .001
Reward acquisition at t-1 = yes	0.16	0.07	2.40	.016
B. Smooth terms	**edf**	**Rel. df**	**F-value**	**p-value**
s(Trial): Condition = control	1.00	1.00	2.64	.104
s(Trial): Condition = test	6.03	7.30	101.02	< .001
fs(Trial,Participant)	63.81	236.00	271.94	< .001

Table 1: Outputs of the generalised additive mixed model of reward acquisition in Experiment 1. **s**: adaptive thin plate regression spline. **fs**: factor smooth. Estimated degrees of freedom (edf) larger than 1 indicate non-linearity of the smooth. AIC = 5861; fREML = 6808; R-sq (adj) = .137; n = 4776

2.5. Discussion

Participants were, overall, able to learn successfully under expected state uncertainty due to the ambiguous nature of the psychophysical stimuli, and the unexpected uncertainty due to the non-signalled reversal. The speed of adaptation to the switch in the state-action mapping varied

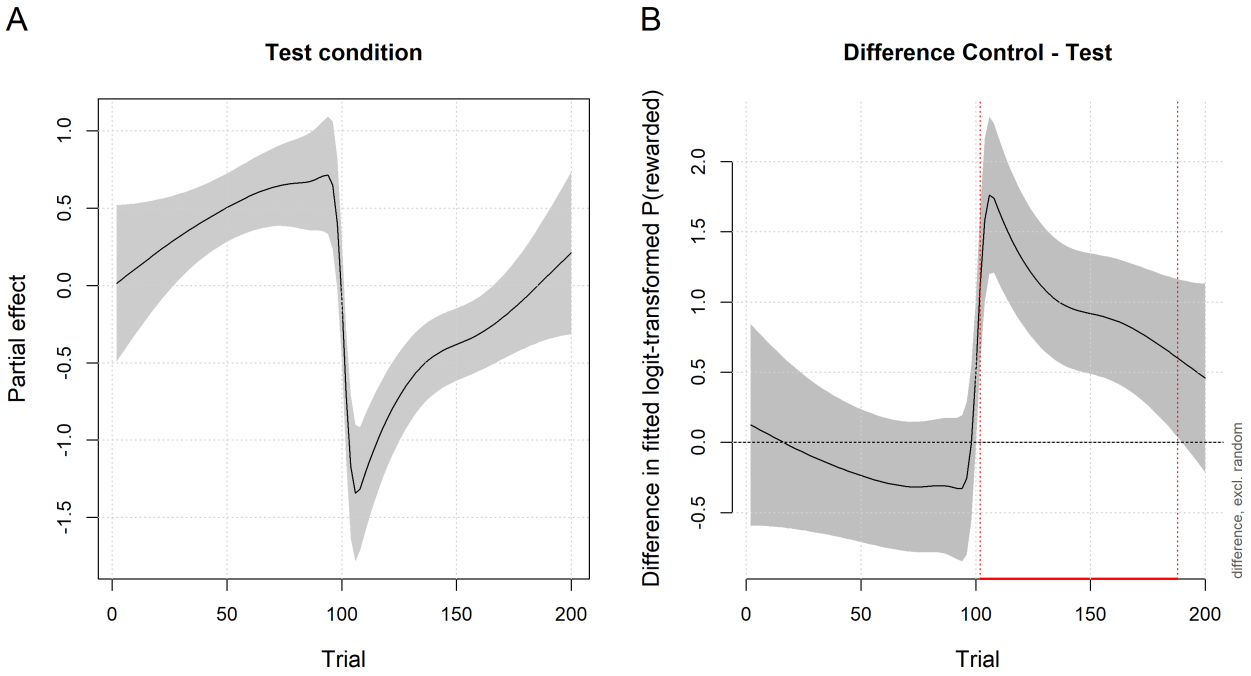

Figure 6: Partial effect of the significant predictor (corresponds to the contribution of Trial order $f(Trial)$ that would be used in Equation 1) along with the interaction effect in the generalised additive mixed effects model of reward acquisition in Experiment 1. A: Nonlinear regression lines for the test condition. The grey shadings along the curves represent pointwise 95% confidence intervals around the curve estimates. B: Difference in fitted logit-transformed proportion of rewarded responses between the control condition and test condition (here the difference excludes the random effects). The red segments indicate where the difference is significant, as assessed on the basis of the 95% CI.

considerably across participants. An additional finding was that participants' learning benefited from re-encountering the same state in two consecutive trials, which could be due to a number of reasons. As we explained when we introduced the state match factor, this could be a simple by-product of how participants keep track of the rules of the task, with value updates occurring at the level of state-action pairs as implicitly assumed in RL models. In such frameworks, getting reward feedback on a visit to a given state should promote selection of the action that will give reward on the next visit of that state, while if the state is not visited, the task knowledge is not updated, and hence no improvement should be observed. The effect of re-encountering a state is likely to be weaker under the Bayesian RL model than under the simple RL model because updates will take place for all states irrespective of which state the learner visits, and thus the effect of the update in the previous trial should be smaller. This prediction is tested later through computational modelling of the data (see S.6 in the supplementary material). Another explanation of the positive effect of state match is that participants learned to disambiguate the states better by using reward feedback; that is, getting feedback on trial t helped correctly identify the state on trial $t + 1$. This interpretation is partially supported by the observation that some participants performed at a higher level than the target uncertainty level, suggesting that they continued to improve in the perceptual discrimination during the learning phase. We address this shortcoming in experiment 2.

3. Experiment 2

Experiment 2 introduced a reward uncertainty condition, and also provided a close replication of Experiment 1. Moreover, we attempted to address one shortcoming in the first experiment, which is that some participants performed at a higher level than the targeted level of uncertainty, potentially due to either an imprecision in the estimation of the psychometric function in the calibration phase, or to an improvement in the detection of the noisy states in the learning phase using reward information. Accordingly, we made three changes: we introduced feedback in the calibration phase; increased the number of contrast levels used for the estimation of the ρ-thresholds from 9 to 10; and increased ρ from 0.65 to 0.7 to reduced the state uncertainty (see Section 3.2 for more details).

3.1. Participants

Experiment 2 included 61 participants: 31 were randomly allocated to the state uncertainty condition, while the remaining 30 participants were allocated to the reward uncertainty condition. Participants were recruited from the University of Western Australia. As in Experiment 1, participants' payment included a fixed amount ($5) and an extra bonus payment that depended on their performance (up to $8). The study was approved by the Human Research Ethics Committee at the University of Western Australia. The number of participants was chosen based on previous studies of reward-based learning under uncertainty (e.g., Behrens et al., 2007; Nassar et al., 2010; Wilson & Niv, 2011). In addition, model simulations indicated that 30 participants would be enough to differentiate between the two RL models used in our study in the state uncertainty condition (see S.2).

3.2. Material and procedure

3.2.1. State uncertainty condition

The structure and the design of the task in the state uncertainty condition was similar to that of the first experiment with three major modifications, mainly in the calibration phase, as follows.

First, to assist participants in learning to disambiguate the states in the learning phase, we introduced feedback in the calibration phase. More specifically, after each response, participants were told whether their identification of the patch state (light or dark) was correct or not by displaying on the screen either the word 'Correct' in green or 'Incorrect' in red. Participants were told to use the feedback to improve their discrimination ability in later trials.

To improve the precision of the estimation of the ρ-thresholds for both the light and dark states in the calibration phase, we increased the number of contrast levels from 9 to 10. We also added 100 trials at the beginning of the calibration to let participants "warm up". In addition, these warm up trials were used to produce an initial, rough estimate of the slope of the psychometric function, which then determined the exact contrast levels to use in the actual calibration phase (250 trials: 25 repetitions for each of the 10 contrast levels). Both the feedback and the increased number of trials served a common goal: to ensure that most of perceptual learning for state classification is complete by the time the state-action learning phase started.

We increased ρ from 0.65 to 0.7, to reduce the extent of individual differences. We expected that the increase in ρ would enable most participants to adapt to the switch, allowing a cleaner comparison of performance between the state uncertainty and reward uncertainty conditions. Nevertheless, a value of $\rho = 0.7$ still represents a considerable amount of state uncertainty, so that the models can be differentiated. Indeed, initial simulations (see S.2) showed that a level of 0.7 made it easier to differentiate between models when model fitting artificial behavioural data, but also led to more accurate model recovery than the lower value of ρ.

3.2.2. Reward uncertainty condition

In the new reward uncertainty condition, the states were unambiguous while rewards were stochastic. The patches used were either completely white (light state) or completely black (dark state), and of the same size as the patches used in the state uncertainty condition. As in the state uncertainty case, participants were required to learn the correct mapping between the two states and two response options by using the reward feedback given to them after each of their responses. To match the uncertainty level used in the state uncertainty condition, we generated rewards following a $Bernoulli(0.7)$ distribution for correct responses and a $Bernoulli(0.3)$ for incorrect responses. The task was otherwise similar to the learning part of the state uncertainty condition in all other aspects, except that obviously there was no calibration phase to construct the light and dark patches.

3.3. Results

The structure of this section largely follows that of the results section of Experiment 1. We start by visualising the average learning curves under the different forms of uncertainty. The patterns of learning will then be tested statistically using the GAMM approach, focusing in particular on the comparison between state and reward uncertainty.

3.3.1. Learning performance under expected uncertainty

In addition to the previously used reward-based learning curves, we also included learning curves that are based on accuracy (see Figure 7). Both types of learning curves were again constructed using running averages with a window size of 30 trials. The main reason for introducing accuracy-based learning curves was that the learning curves in the reward uncertainty condition provide a different perspective on learning performance than the reward-based learning curves, given the stochasticity in the reward function but not in the accuracy function. That is, a response can be accurate but not rewarded, which can never be the case in the state uncertainty condition (hence the match between the two curves under state uncertainty in the left figure panel). In the reward uncertainty condition, a participant's accuracy-based learning curve should reach 1 when the participant has learned and has been using the correct rules. The reward-based learning curve should fluctuate around 0.7 in the case of successful rule acquisition, assuming participants exploit this rule with limited or no exploration. In what follows, we will primarily analyse reward-based performance (proportion of rewarded responses) but will make reference to accuracy-based performance when needed during the analysis of the reward uncertainty condition data.

First, consider the case of (expected) state uncertainty in the first 100 trials before the switch. The aggregated learning curve in Figure 7 (left panel) suggests that, overall, participants were able to learn the task within the first 100 trials and performed at the target level of state uncertainty $\rho = 0.7$. Looking at participants individually in Figure 8 (left), the majority of participants (26 out of 31) reached a reward acquisition rate above chance (0.5) in the last 20-trial block before the switch. Twenty out of 31 participants showed improved performance relative to the first 20-trial block (a further 5 participants did not show performance improvement even with a score above 65% in the last block because they had reached an even higher score in the first block of 20 trials). These results confirm our findings in the first experiment, i.e., the majority of participants were able to deal with expected uncertainty when the uncertainty comes from the states.

In the case of reward uncertainty, participants performed well below those from the state uncertainty group in the first 100 trials, and also below our uncertainty threshold (red curve in the right panel of figure 7). The finding that the aggregated accuracy-based curve (in green

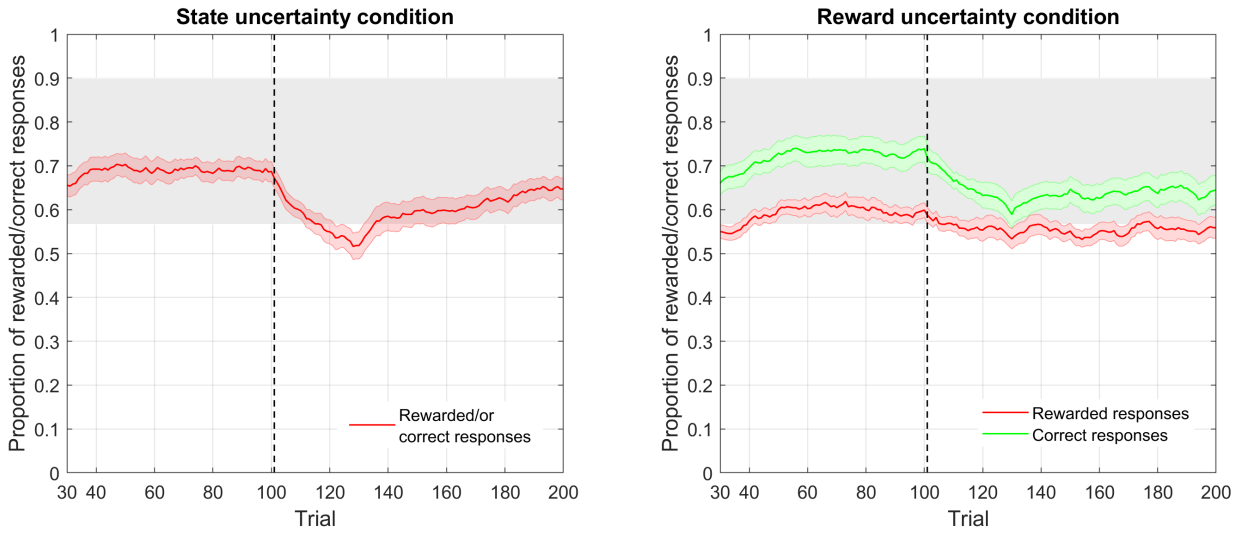

Figure 7: Overall running average scores of the proportion of rewarded responses (red) and correct responses (green) in the state uncertainty (left) and reward uncertainty (right) condition. The running averages were computed using a window size of 30. The shading along each curve represents the standard error of the mean, while the light grey region represents an approximate 95% confidence interval of the moving average of a Bernoulli variable with probability of success equal to 0.7. The vertical dashed line indicates the trial where the switch occurs. Note that the reward-based running average matches the accuracy-based running average in the reward uncertainty condition.

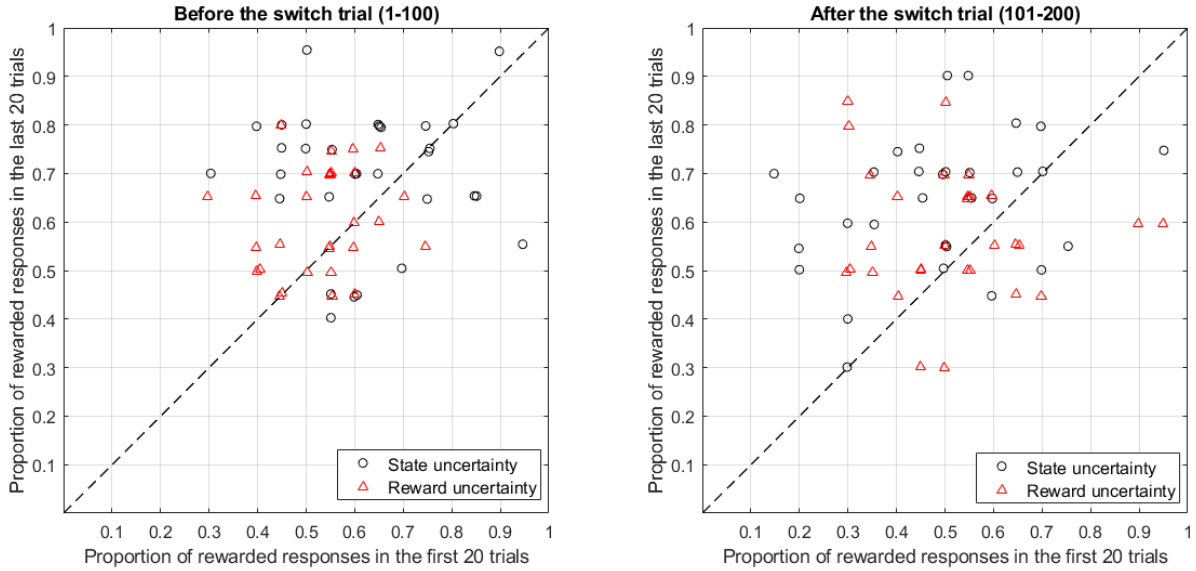

Figure 8: Comparison of the average proportion of rewarded responses in the first 20 trials and last 20 trials before the "switch trial" (left) and after it (right), for both the control and test group. A jitter of 0.005 was applied to enhance the visualisations.

colour) is well below 1 confirms that participants, in general, were not exclusively employing the optimal response mapping. However, the upward trends in the learning curves provide evidence that some learning took place. The individual data presented in Figure 8 (left) shows that 22 participants out of 30 reached a performance score above the chance level (points above 0.5), but only half of the participants showed improved performance (points above the equality line in the figure). The results are therefore more mixed in the reward uncertainty condition in comparison with state uncertainty, with no clear majority for those who learned the mapping rules. It is possible that some participants may have learned the task rules but continued to explore because they may think that the rules might change at some point or because they were behaving in accordance with probability matching (Vulkan, 2000).

3.3.2. Learning performance under unexpected uncertainty

Responses to the switch (in trials 101–200) in the state uncertainty condition were similar to those for Experiment 1 (see the left panel in Figure 7 and the right panel in Figure 8). The data suggest that the majority of our participants managed to adapt to the switch (20 out of 31 participants performed above chance in the last 20-block post-switch and showed improved performance in comparison with the first block post-switch; a further 3 participants did not show improvement because their performance score was already high in the first 20 trials).

For reward uncertainty, the impact of the unexpected switch on overall performance was less pronounced than in state uncertainty because performance was already low before the switch. However, performance remained around the chance level for most of the trials post-switch (right panel in Figure 7). The accuracy-based learning curve in green also shows that performance did not recover back to the level it was at prior to the switch. The average proportion of rewarded responses in reward uncertainty was lower than that in the state uncertainty condition except in the first 30 trials post-switch. Figure 8 (right panel) shows that, again, only half of our participants improved their performance relative to the first 20-trial block post-switch and ended up with a reward acquisition score above chance. These results suggest that the switch had a negative impact on participants' performance in the reward uncertainty condition as well, with even more participants failing to deal with the unexpected switch in comparison with state uncertainty.

3.3.3. Analysis of learning performance using statistical modelling

The saturated model was selected as the best model based on backward model selection. This model (see Table 2) included the three-way non-linear interaction between trial order, uncertainty condition and state match (recall that this was coded as a two-way non-linear interaction by combining uncertainty and state match as a single factor), as well as reward acquisition in the previous trial (used mainly to remove temporal residual autocorrelation). Figure 9 presents the significant partial effects involving trial order along with the interaction effects. First, in the state uncertainty condition (SU, Panel A), the effect of trial order on the likelihood of receiving a reward, when the states do not match, followed the same pattern encountered in the previous exploratory analyses: an increase in the likelihood of being rewarded with trial order until the occurrence of the switch, then an abrupt decrease in performance before the performance recovers after a few trials. The likelihood of receiving a reward remained significantly higher for matched states in comparison with unmatched states for most of the trials, can be seen in Panel C.

In the reward uncertainty condition, the effect of trial order on reward likelihood was non-significant for the unmatched state ($p = .274$), but was significant for matched states ($p = .015$), most likely due to the initial large dip in the non-linear regression line; the large dip is due to the high proportion of rewarded responses in the second trial where the state matched that of

the first trial; see Figure 9B. The interaction graph in Figure 9D shows that the effect of trial order on reward likelihood was almost the same for both matched and unmatched states.

Figure 9E-F compares the state and reward uncertainties. For matched states (Panel E), reward likelihood was significantly higher in state uncertainty than in reward uncertainty for all but the first three trials. When the states at t and $t-1$ were different (Panel F), the reward likelihood was significantly higher in the state uncertainty condition for 28 trials in the middle of the first block of 100 trials (i.e. under expected uncertainty), but the reward likelihood was significantly lower in state uncertainty during the first 40 trials in the second 100-trial block (i.e. after unexpected uncertainty has been introduced). The difference between the state and reward uncertainty conditions was otherwise non-significant.

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	.14	.07	2.11	.035
Condition & State match = RU.diff	0.02	0.09	0.265	0.791
Condition & State match = SU.same	0.61	0.06	10.95	$< .001$
Condition & State match = RU.same	-0.01	0.09	-0.09	0.930
Reward acquisition at t-1	0.23	.04	5.84	$< .001$
B. Smooth terms	**edf**	**Rel. df**	**F-value**	**p-value**
s(Trial): Condition & State match = SU.diff	8.59	10.34	95.91	$< .001$
s(Trial): Condition & State match = RU.diff	2.53	3.13	3.78	.274
s(Trial): Condition & State match = SU.same	1.46	1.73	1.28	.576
s(Trial): Condition & State match = RU.same	3.85	4.71	13.17	.015
fs(Trial,Participant)	107.49	606.00	294.87	$< .001$

Table 2: Outputs of the generalised additive mixed model of reward acquisition in Experiment 2. **s**: adaptive thin plate regression spline. **fs**: factor smooth. Estimated degrees of freedom (edf) larger than 1 indicate non-linearity of the smooth. AIC = 15770; fREML = 17251; R-sq (adj) = .055; n = 12139

3.4. Discussion

Participants' performance was better in state uncertainty when measured by reward received. The difference between the two conditions, however, was mainly observed in the first block of 40 trials after the switch when participants encountered a different state than in the previous trial (reward uncertainty was associated with higher reward likelihood during the first few trials post switch, mainly because performance was already low prior to the switch and, therefore, did not get affected much by the change in response mapping). Participants also displayed different switch adaptation patterns in reward uncertainty in comparison with state uncertainty (and often less obvious; see Figure S.6 in S.1). Moreover, participants' behaviour appeared to be more exploratory in the reward uncertainty case. These combined effects present a challenge to the computational models, notably the interaction between uncertainty condition and state match as well as the observed the pattern of adaptation to the switch for unmatched states in the state uncertainty condition.

4. Computational models

To explain participants' performance in the two experiments, we compared three computational models. One is a Bayesian temporal-difference-based model, which assumes the learner can use the reward feedback to resolve state uncertainty (Larsen et al., 2010). The second is a simple temporal-difference-based model in which the learner ignores the state uncertainty, and

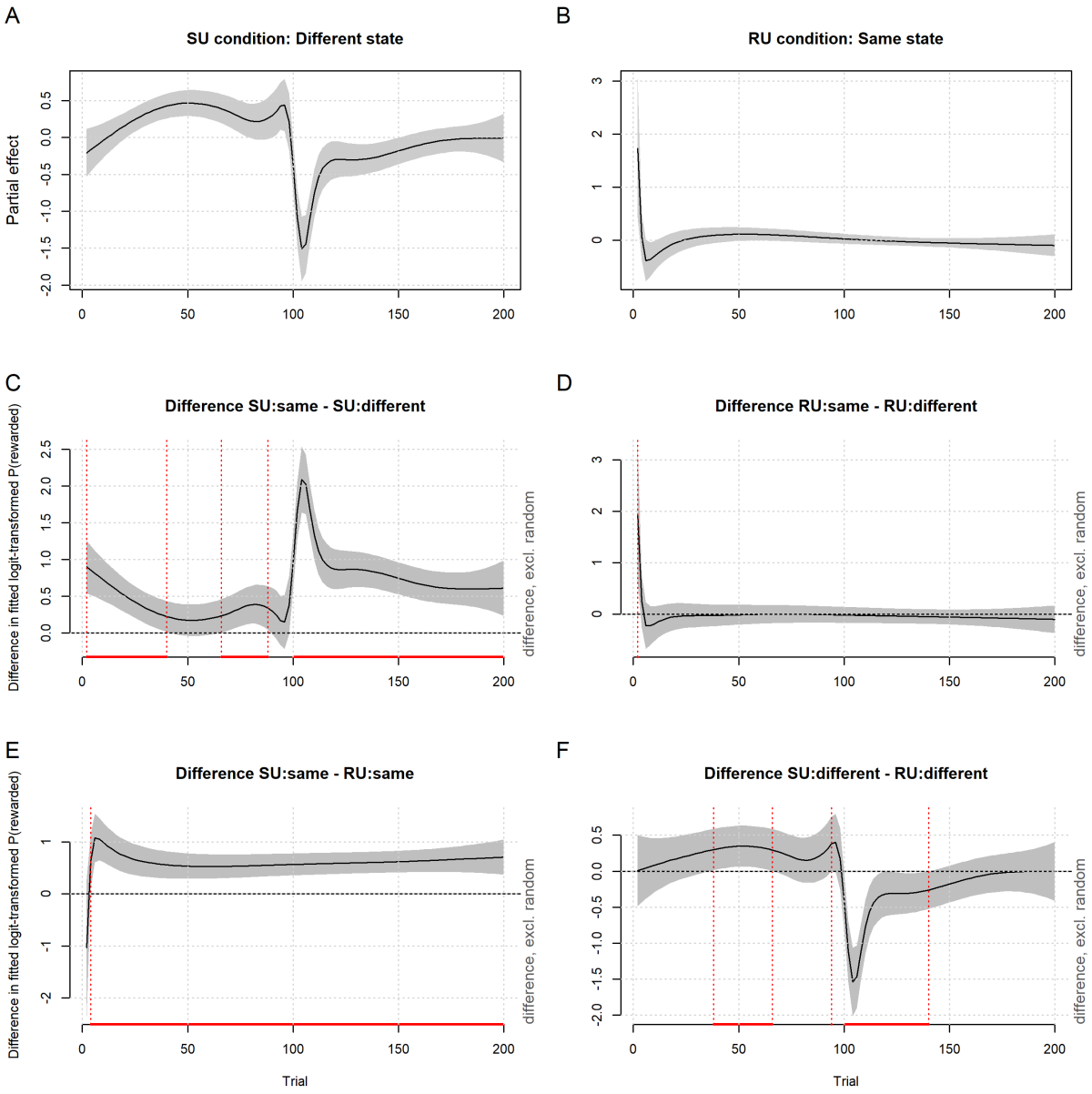

Figure 9: Partial effects (fixed effects only) of the significant predictors along with the interaction effects in the generalised additive mixed effects model of reward acquisition in Experiment 2. A-B: Each panel depicts the nonlinear regression line for a combination of uncertainty condition and state match level, with pointwise 95% confidence intervals (grey shading). C: Difference in fitted logit-transformed values between matching and non-matching states in the state uncertainty condition. D: Same as C but the difference is calculated for responses from the reward uncertainty condition. E: Difference in fitted logit-transformed proportion of rewarded responses between state uncertainty and reward uncertainty when the state on the present trial matches that on the previous trial. F: Same as E but the difference is calculated for trials where the states do not match. All difference terms were calculated with random effect terms set to 0. The red segments indicate where the differences are significant.

updates only the Q-value (predicted reward) of the state that they believe is the true state (this corresponds to "the winner takes all" model described in Larsen et al., 2010). The third is a simplified version of the learning-by-sampling model of Chen et al. (2011). We aimed to determine whether or not the same computational mechanism underlies learning under both state and reward uncertainty. Below, we describe each model together with its mathematical formulation under both uncertainty conditions.

4.1. Description of models and modelling procedure

4.1.1. The PWRL model

Formulation under state uncertainty

PWRL (Posterior Weighted Reinforcement Learning; Larsen et al., 2010) shows how to augment temporal difference learning to deal with state uncertainty. The model allocates reward to each state according to the posterior probability of each state being the true state having observed the reward. The basic idea behind this scheme is that rewards carry information about the true state as do observations, and hence should be used to estimate state probabilities. More precisely, the state-action values $Q_t(s,a)$, which here estimate the expected reward for action a (circle or square) in state s (light or dark) at time t, are updated as follows:

$$Q_{t+1}(s,a_t) = Q_t(s,a_t) + \alpha \mathbb{P}(S_t = s | r_t, a_t, i_t, \boldsymbol{Q_t})(r_t - Q_t(s,a_t)) \tag{3}$$

where updates are for all possible states and only the chosen action a_t, α is the learning rate, i_t is the state that is identified by the learner (it might be different from the true state s_t), $\boldsymbol{Q_t}$ is the vector of all state action values, and $\mathbb{P}(S_t = s | r_t, a_t, i_t, \boldsymbol{Q_t})$ is the posterior probability that the true state is s having observed the reward r_t.

Expanding $\mathbb{P}(S_t = s | r_t, a_t, i_t, \boldsymbol{Q_t})$ using Bayes rule produces the following updating rule for the Q-values (for more detail, see S.3):

$$Q_{t+1}(s,a_t) = \begin{cases} Q_t(s,a_t) + \alpha \frac{Q_t(s,a_t)\rho_t(s_t,s)}{\sum_{s'} Q_t(s',a_t)\rho_t(s_t,s')}(2.5 - Q_t(s,a_t)) & \text{if } r_t = 2.5, \\ Q_t(s,a_t) + \alpha \frac{(2.5 - Q_t(s,a_t))\rho_t(s_t,s)}{\sum_{s'} (2.5 - Q_t(s',a_t))\rho_t(s_t,s')}(0 - Q_t(s,a_t)) & \text{if } r_t = 0. \end{cases} \tag{4}$$

where $\rho_t(s,i)$ is the probability with which the learner identifies i as the true state, such that their identification i is correct with probability ρ:

$$\rho_t(s,i) = \mathbb{P}(I_t = i | S_t = s) = \begin{cases} \rho & \text{if } i = s, \\ 1 - \rho & \text{otherwise.} \end{cases} \tag{5}$$

For action selection, we assume that the learner uses the softmax selection rule to compute the probability of each action given the current (true) state, that is:

$$\mathbb{P}(A_t = a | S_t = s_t) = \sum_{i'} \frac{\exp(Q(i',a)/T)}{\sum_{a'} \exp(Q(i',a')/T)} \rho_t(s_t, i') \tag{6}$$

where where T is a temperature parameter, which determines the degree of stochasticity in action selection.

To sum up, the steps followed by the learner in each trial under this model are as follows: First, they observe a noisy input from the environment (whose state can be correctly identified with probability ρ). Second, they select an action according to the softmax choice function (see Equation 6). Third, the reward feedback is used to calculate the posterior probability of each

state being the correct state (i.e. $\mathbb{P}(S_t = s|r_t, a_t, i_t, \boldsymbol{Q_t})$). Fourth, they update the state-action values using Equation 4. This cycle is repeated with each new stimulus.

Although this model makes optimal use of reward feedback to get information about the current state before updating the state-action values, it can suffer from difficulties when a change in mapping rules is introduced, as shown in Larsen et al. (2010). In fact, consider the following example, which is taken from the same paper and adapted to our case. Let us first denote by s_L and s_D the light and dark state respectively, and by a_L and a_D the correct "light" and "dark" response before the switch. Prior to the switch—in our experimental setting—the reward for choosing a_L in state s_L is 2.5, whereas it is 0 if a_D is selected in s_L. In trial 101, the rewards switch, so that $R_t(s_L, a_L) = 0$ and $R_t(s_L, a_D) = 2.5$, for $t > 100$. Assume that by $t = 100$, the model has learned the correct state-action values, so that $Q_{100}(s_L, a_L) = 2.5$ and $Q_{100}(s_L, a_D) = 0$. Suppose also that the true state at $t = 101$ is s_L, so that the probabilities of detecting each state as the true state are: $\rho_t(s_L, s_L) = 0.65$ and $\rho_t(s_L, s_D) = 0.35$ (from now on, we consider t to be equal to 101). Let's say the learner chooses at t what they believe to be the best action, a_L. Because the mapping rules have switched, their action would result in a null reward. Thus, the posterior probability that the state is s_L is given by (from Equations S.2):

$$\begin{aligned} \frac{\rho_t(s_L, s_L)\,(2.5 - Q_t(s_L, a_L))}{\rho_t(s_L, s_L)\,(2.5 - Q_t(s_L, a_L)) + \rho_t(s_L, s_D)\,(2.5 - Q_t(s_L, a_D))} &= \frac{0.65 \times 0}{0.65 \times 0 + 0.35 \times 2.5} \\ &= 0 \end{aligned} \tag{7}$$

This means that the learner would believe with certainty, after observing the null reward, that the true state is s_D rather than s_L, and hence would not change $Q_t(s_L, a_D)$, thereby continuing to use the outdated state-action mapping.

Note that, in this example, we have assumed that the estimated state-action values have converged to the correct ones to illustrate why the PWRL model might get stuck with using the same mapping after the switch. The model can potentially overcome this, for example, by using a very low value for the learning rate to avoid convergence. However, a low learning rate slows down adaptation to the switch, because too much of the irrelevant (pre-switch) reward history is integrated in the value estimates. Note that the actual rate of adaptation is not just determined by the learning rate, but also by the rate of exploration (for a discussion of the intercorrelation between learning and exploration rates in RL models, see Daw, 2011). Accordingly, we expect adaptation in the PWRL model to be absent or slow if the model successfully switches after trial 100.

Formulation under reward uncertainty

In the reward uncertainty condition, the PWRL model becomes equivalent to a standard SARSA (State-Action-Reward-State-Action) model (Rummery & Niranjan, 1994), with state-action values updated on each trial as follows:

$$Q_{t+1}(s, a) = \begin{cases} Q_t(s, a) + \alpha\,(r_t - Q_t(s, a)) & \text{if } s = s_t, a = a_t, \\ Q_t(s, a) & \text{otherwise,} \end{cases} \tag{8}$$

where α is the learning rate, r_t is the Bernoulli generated reward received at t, s_t and a_t refer respectively to the current state and current action. Equation 8 follows from Equation 3 given that the posterior probability that the true state is s_t given that the reward is r_t is $\mathbb{P}(S_t = s|r_t, a_t, i_t, \boldsymbol{Q_t}) = 1$.

4.1.2. The WTA-RL model

Formulation under state uncertainty

One simple approach to solve reward-based learning tasks with state uncertainty is to ignore the uncertainty and simply update the Q-value of the identified state in each trial. This is the basic idea behind the "winner takes all" reinforcement learning (WTA-RL) model presented by Larsen et al. (2010). More specifically, the model assumes that, in trial t, the learner identifies the true state as i_t (via a sampling process that correctly identifies the true state with probability ρ), then updates the Q-value for only that identified state. Thus, in our experimental setting, the learner would update the right state-action value, on average, with frequency ρ. Under this scheme, the updating equation for the state-action values is given by:

$$Q_{t+1}(s,a) = \begin{cases} Q_t(s,a) + \alpha\left(r_t - Q_t(s,a)\right) & \text{if } s = i_t, a = a_t, \\ Q_t(s,a) & \text{otherwise.} \end{cases} \tag{9}$$

Using the state-action values, the learner selects actions according to the softmax selection rule defined in Equation S.6. This model is hence similar to a SARSA model with the difference being that it updates $Q_t(i_t, a_t)$ instead of the Q-value for the current state s_t, $Q_t(s_t, a_t)$.

One advantage of this model is that it adapts well to a switch of contingencies, as it does not suffer from the problem of self-confirming beliefs encountered with the PWRL model. However, the WTA-RL model is not assured to converge to the correct state-action values (Larsen et al., 2010). Nevertheless, learning incorrect Q-values does not necessary result in incorrect actions, and hence the model can still be applied to solve learning tasks with state uncertainty.

Formulation under reward uncertainty

Just like the PWRL model, WTA also becomes equivalent to a SARSA model under reward uncertainty. In fact, the SARSA's updating rule of state-action value (Equation 8) is obtained from the that of the WTA model (Equation 9) by replacing the identified state i_t by the true state s_t since the states are unambiguously identified by the learner.

On this account, both the WTA and PWRL models predict that humans would follow a SARSA-type learning behaviour in the reward uncertainty condition, while predicting different behaviours in the state uncertainty condition. In this condition, under PWRL, either no or very slow adaptation would be observed, whereas the opposite is expected under WTA-RL.

4.1.3. The BI-SAW model

Formulation under state uncertainty

We extend the sampling-based BI-SAW model (Bounded memory, Inertia, Sampling and Weighting; Chen et al., 2011), which was originally developed to account for human reward-based learning in binary stateless bandit tasks, to the case of multiple-state environments. For consistency with the two RL models, we replace the exploration rule in the original model (ϵ-greedy followed by a Bernoulli-distributed choice between the two actions) with a softmax exploration rule. We also exclude the assumption of inertia (i.e. tendency to repeat the last choice).

The model assumes that the learner chooses actions based on small samples of experiences drawn from the recent past, while also considering the overall trend of rewards from all past trials. The extent to which the learner relies on a small sample of the most recent trials or the overall reward trend is modulated by a weight parameter that is subject-specific. This property of the model makes it well suited to capture different possible learning behaviours in tasks that include an unexpected change as the one we use here. For example, a fast adaptation to the

switch can be simulated by increasing the weight associated with the small sample component, while a slow adaption to the switch can be simulated by increasing the weight of the overall reward trend.

More specifically, each state-action pair is associated with an Estimated Subjective Value (ESV) that is computed on each trial $t > 1$ as follows:

$$ESV_t(i_t, a) = (1 - \omega)SampleM_t(i_t, a) + \omega GrandM_t(i_t, a), \tag{10}$$

where i_t is the identified state on trial t. Index a refers to each possible action (light or dark response in our case). Parameter ω determines the weight that is given to $GrandM_t(i_t, a)$, the grand average payoff choosing action a in state i_t in all past experience. $SampleM_t(i_t, a)$ is the average reward for choosing action a in state i_t within a sample of μ past payoffs drawn from the most recent b trials. The sampling process is assumed to be independent and with replacement such that, in each draw, the most recent trial $(t - 1)$ has a probability p_r to be chosen. All the remaining $(b - 1)$ trials have an equal probability to be selected. To simplify the model further, we restrict μ to be equal to b—that is, the learner samples with replacement exactly b payoffs from the b most recent ones. Finally, the learner is assumed to select an action according to the softmax rule using the computed ESVs.

The model is similar to the WTA-RL model in that both assume that the learner identifies one of the states as the true state, then calculates the value(s) associated with that state. The main difference between the two models is that the subjective values in the BI-SAW model are not updated based on previous values but are calculated afresh on each new trial.

Formulation under reward uncertainty

In the reward uncertainty case, the learner can identify the correct state with certainty, and will be positively rewarded for choosing the correct response with probability ρ and positively rewarded with probability $1 - \rho$ even if they select the incorrect response. Therefore, the sampling model predicts that people's learning behaviour should be identical in the two uncertainty conditions, since the distribution of the sequences of rewards is identical in the two conditions.

4.1.4. Model evaluation

To compare the models in their account of participants' choices, we estimated the free parameters of each of the three models separately for each participant using (approximate) maximum likelihood estimation. More specifically, we selected model parameters that maximised the log-likelihood of observed actions conditioned on the previously encountered observations and rewards (Daw, 2011). A detailed description of how the likelihood functions were computed and implemented is provided in the supplementary material in S.4.

To compare the goodness of fit of the models, we computed the Bayesian information criterion (BIC) for each subject and model using the maximised log-likelihood for that subject. To find out which model provides substantial evidence for the observed data from each participant, we converted the BIC scores of each pair of models into a Bayes factor using the following approximation

(Kass & Raftery, 1995):

$$BF_{12} = \exp\left(\frac{BIC(M_2) - BIC(M_1)}{2}\right) \tag{11}$$

where BF_{12} is the Bayes factor for model M_1 against M_2. We then considered a model as the winning model if its Bayes factor against the model with the second highest BIC score is

greater than 3.2, indicating substantial evidence for the model according to Jeffreys's (1961) scale of evidence. Bayes factors lower than 3.2 were taken as providing inconclusive evidence for differentiating between the models, and in this case no model was declared as the best fitting model.

The analysis of participants' learning curves reported earlier indicated that several participants performed at a significantly higher level than ρ in the state uncertainty conditions, suggesting that their real level of state uncertainty might deviate from the targeted level (i.e., $\rho = 0.65$ in Experiment 1 and $\rho = 0.7$ in Experiment 2). To accommodate such deviations, we fit the models with and without ρ as a free parameter, and found that considering ρ as a free parameter produced better fits for all models and conditions, as judged by the BIC scores. Hence in the main text, we only present the fitting results with ρ as free parameter (the fitting results with ρ fixed are presented in the supplementary material in S.9).

4.2. Results and discussion

Here we focus on the model fitting results for the second experiment, which are summarised in Table 3 (we report the results for the first experiment in S.7 since they were inconclusive). The estimated ρ in the state uncertainty condition was markedly higher than the targeted level of 0.7 and consistent across all three models. This suggests that despite the measures introduced in the second experiment to control the uncertainty level, some participants still managed to improve their capability of discriminating between the noisy states. Furthermore, the estimated exploration rate T provides a hint as to the reason for the lower performance in the reward uncertainty condition, as exploration for all three models was greater in the reward uncertainty condition than in the state uncertainty condition. For example, WTA-RL, the best-fitting model as will be shown below, had a significantly lower exploration rate under state uncertainty ($Mdn = 5.04$) in comparison with reward uncertainty ($Mdn = 8.12$), as indicated by a Wilcoxon rank-sum test, $W = 296, p = .015, r = -.31$. Learning rate α did not differ significantly between state uncertainty ($Mdn = 0.34$) and reward uncertainty ($Mdn = 0.33$); $W = 450, p = .834, r = -.03$.

Table 3: The mean ± S.D. of best-fitting parameter values and BIC scores by model and condition in Experiment 2. For the bound memory parameter b of the BI-SAW model, the median and the median absolute deviation were used instead of the mean and standard deviation since b can only take discrete values.

	State uncertainty			Reward uncertainty	
Parameter	PWRL	WTA-RL	BI-SAW	SARSA	BI-SAW
α	0.43 (0.18)	0.41 (0.36)		0.37 (0.17)	
T	9.96 (6.15)	5.75 (4.05)	6.50 (5.98)	12.17 (12.10)	10.51 (10.28)
ρ	0.89 (0.19)	0.86 (0.13)	0.85 (0.14)		
b			3 (2)		3 (2)
w			0.35 (0.29)		0.27 (0.31)
p_r			0.52 (0.30)		0.69 (0.31)
$-logML$	122.4 (12.4)	120.9 (13.1)	119.2 (13.5)	117.9 (23.6)	114.8 (18.5)
BIC	260.7 (24.8)	257.6 (26.3)	264.9 (27.1)	246.3 (47.3)	250.8 (37.0)

A comparison between observed and simulated learning curves in both uncertainty conditions (see Figure 10) shows that all computational models qualitatively captured the behavioural choices well, certainly in the state uncertainty condition. In this condition, the simple RL model (WTA-RL) provided a better qualitative account of participants' choices, most noticeably in the post-switch trials. In the reward uncertainty condition, the RL and sampling-based models were hard to differentiate. In addition, both the RL and sampling models seem to have underestimated

the proportion of rewarded responses for most trials, suggesting a lower quality of model fit in comparison with state uncertainty. This is partly explained by the complete misfit of several participants' learning curves as shown in Figures S.18 and S.19 (e.g. subj 13, 16, 18, 24, 27 and 30 where the models simply behaved randomly).

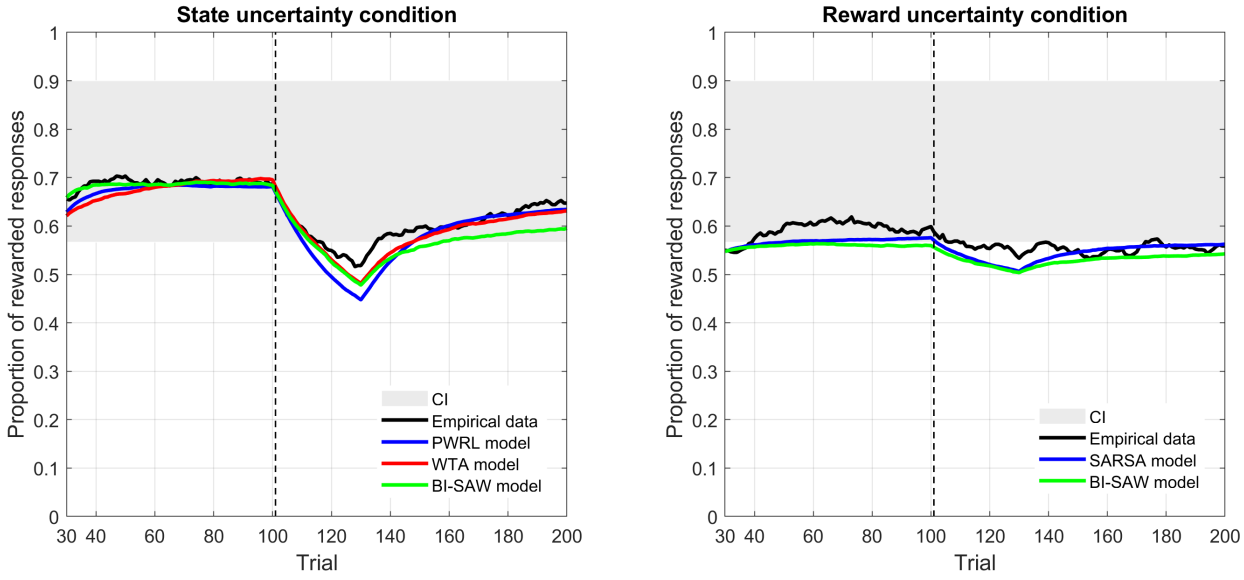

Figure 10: Comparison of the reward-based learning curves of participants and fitted models in both the state uncertainty (left) and reward uncertainty (right) conditions. To draw model learning curves, we first constructed a learning curve for each subject by averaging over 100 simulations with her fitted parameters, then averaged across participants to get the overall learning curve. The light grey region represents an approximate 95% confidence interval of the moving average of a Bernoulli variable with probability of success equal to 0.70. The vertical dashed line indicates the trial where the switch occurs.

To compare the model fits quantitatively, we evaluated the BIC values of our models and computed the number of participants best fitted by each model based on Bayes factors, as described in the model evaluation section. The results are summarised in Figure 11. In the state uncertainty condition, the WTA-RL model had a better overall BIC score and was the best-fitting model for twice as many participants as were best fit by PWRL (14 versus 7). None of the participants' data was best fit by the BI-SAW model. In the reward uncertainty condition, the SARSA model—the equivalent version of the Bayesian and simple RL models used in the state uncertainty condition—was the best fitting model for twice the number of participants as were best fit by the sampling model (16 versus 8), though the median BIC score slightly favoured the sampling model. These results demonstrate that the simple RL model explained participants' choices the best in both state uncertainty and reward uncertainty conditions. We link individual differences in the learning patterns to the model fits in section S.5 of the supplementary material. We also show in section S.6 that the simple RL model was able to recover the effects observed in the GAMM analysis of the behavioural data from Experiment 2 reported in Section 3.3.3.

5. General discussion

5.1. Summary of the results

We aimed to assess and compare how people learn from reward feedback under conditions of expected and unexpected uncertainty, and compared their learning behaviour under conditions

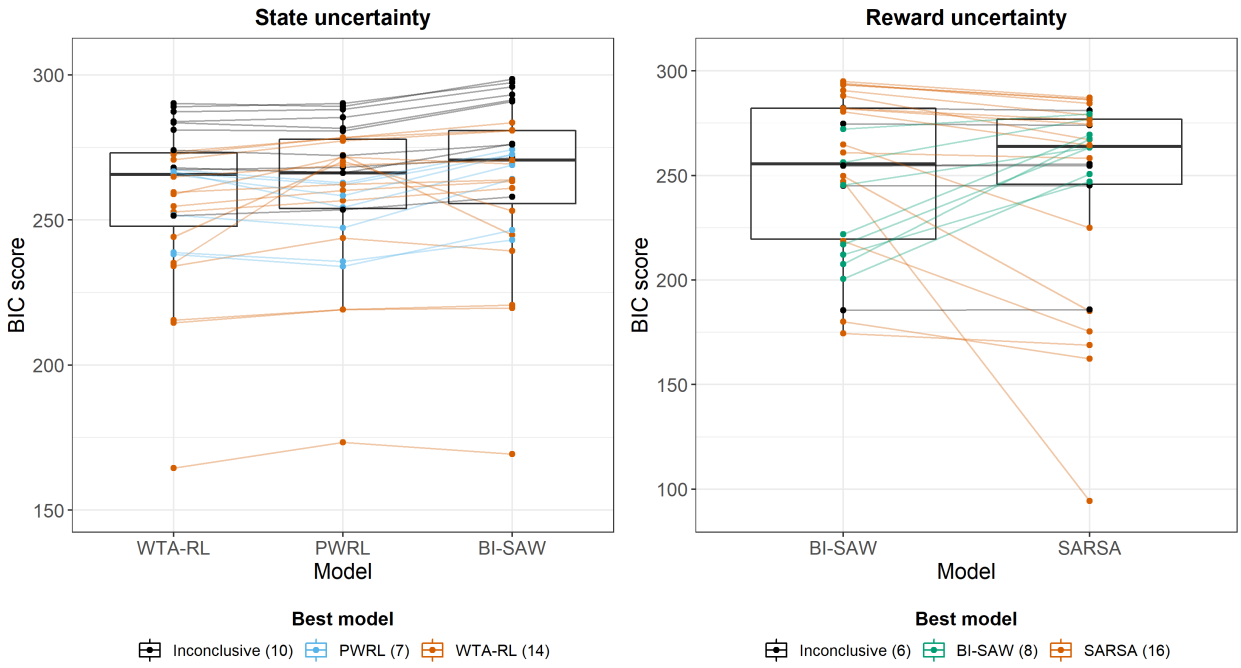

Figure 11: Distribution of BIC scores by model for both the state uncertainty (left) and reward uncertainty (right) conditions, with lower BIC scores indicating a better overall fit. Note that in the reward condition, the PWRL and WTA-RL models are equivalent to a SARSA model.

of state and reward uncertainty. Previous studies have investigated expected and unexpected uncertainty arising from reward and mostly focused on volatility (e.g., Yu & Dayan, 2005; Behrens et al., 2007; Nassar et al., 2010; Wilson & Niv, 2011; Payzan-LeNestour et al., 2013; Piray & Daw, 2020, 2021), but not when the source of uncertainty is the environment's state (Bach & Dolan, 2012; Ma & Jazayeri, 2014; Mathys et al., 2014). Importantly, we matched carefully the degree of reward uncertainty and state uncertainty, and subjected performance under both conditions to the same analyses so that it may be compared directly. We showed that most participants were able to adapt successfully to all three types of uncertainty; participants managed to learn the correct state-action mapping under expected state uncertainty and expected reward uncertainty, and managed to adapt to the unexpected uncertainty due to a non-signalled reversal in the state-action reward contingency. This is in line with and goes beyond previous work that looked at reward-based learning under expected reward uncertainty (Don et al., 2019), expected state uncertainty (Bruckner et al., 2020) and unexpected reward uncertainty (Brown & Steyvers, 2009). Surprisingly, participants collected more reward under state uncertainty than under reward uncertainty, partly due to their higher exploration rate under the reward uncertainty condition.

To identify the psychological mechanisms underlying participants' behaviour, we used two RL models for learning under state uncertainty that were described in Larsen et al. (2010). The first model, named WTA-RL (Winner-Takes-All), is a simple RL model that, on each trial, selects an action based on the state identified by the learner. Importantly, this model only updates the state-action value for the identified state even though the state may have been identified incorrectly. The second model is PWRL (Posterior Weighted RL), which is a Bayesian RL model that selects actions based on belief states, and updates on all states rather than the one state that is identified (Jaakkola et al., 1995). These two RL models become identical in the

reward uncertainty condition, where the state is known with certainty. We also included a third model—BISAW (Bounded memory, Inertia, Sampling and Weighting; Chen et al., 2011)—that selects actions based on a small sample of past rewards as well as the average reward over all trials. We adapted this model to the case of state uncertainty by ignoring the state uncertainty and using only the state that was identified by the learner on each trial, as assumed under the simple RL model. Our modelling results showed that participants' learning patterns were well-captured by the simple RL model in both state and reward uncertainty, suggesting a common mechanism for dealing with these two types of uncertainty.

More specifically, in state uncertainty, the simple RL was by far the model that accounted for participants' choices the best. However, the Bayesian RL model captured the behaviour of those who did not adapt to the switch and some participants who adapted slowly. None of the participants was best fit by the sampling model. In the reward uncertainty condition, the RL framework explained how participants learned to maximise reward better than the sampling model at the level of individual subjects, although there was some indication that for the sample as a whole, the BISAW model was competitive.

5.2. Relationship with belief-state based learning

Several recent studies have found that belief-state based models provide a good account of the dopaminergic response to reward prediction errors in animals (Starkweather et al., 2017; Lak et al., 2017; Babayan et al., 2018; Lak et al., 2020). However, most of these studies used prior rather than posterior beliefs over states and none compared their belief-state based models to a RL model that commits to a state during both the decision and update stages as the WTA-RL model does. Therefore, a direct comparison between a posterior belief state-based RL model and a state "committing" RL model has not been carried out in any of these studies. Starkweather et al. (2017) administrated a Pavlovian conditioning task to mice where state uncertainty arise from temporal (as opposed to perceptual) cues, as cue-reward interval lengths were varied and reward was occasionally omitted. They show that a (posterior) belief state temporal difference model explains mice' dopamine reward prediction errors better than a standard temporal difference model (with a complete serial compound representation; see, for example, Moore et al., 1998).

In a follow-up study, Babayan et al. (2018) proposed a different Pavlovian task where state uncertainty was due to ambiguous visual cues. Their belief-state model is quite similar to the PWRL model, but with values being computed over continuous belief vectors through linear approximation (their task was designed in such as way that the nature of state could be inferred from the reward signal). They found that this model fit the dopaminergic response in mice better than a stateless RL model (though a state-"committing" model could have been included for a more stringent test of the belief-state model). Interestingly, the values extracted from their fitted belief-state model also accounted well for the anticipatory licking patterns of mice, even though the licking data were not used in model fitting. Lak et al. (2017, 2020) provided support for belief-state models by linking measures of response confidence extracted from their models to the dopaminergic response in monkeys and mice, respectively.

A more recent study showed that people use belief states when learning under state uncertainty, but their learning is modulated by categorical perceptual commitments (Bruckner et al., 2020). As in our experiment, participants had to learn state-response mappings via reward feedback under perceptual uncertainty. However, unlike our state uncertainty task, they asked participants to first report the state of a noisy (Gabor) patch (perceptual decision), and only then to choose the most rewarding response (also, the degree of state uncertainty varied between trials in their task, though it was, on average, substantially lower than in our task). Bruckner

et al. (2020) tested different Q-learning and normative Bayes-optimal learning models, including hybrid versions of the Q-learning and Bayes models that combine predictions from both belief state and state committing models. They found that these hybrid versions accounted for participants' choices better than state-belief based models or models that make categorical perceptual choices (as the WTA-RL model does), thus providing support that perceptual commitment biases play a role in reward-based learning beyond belief-state tracking. Similar biases were also reported in several studies of human decision making, which show that people tend to commit to one perceptual interpretation and ignore the other potential interpretations (Stocker & Simoncelli, 2007; Fleming et al., 2013; Maniscalco et al., 2016). This could explain why the WTA-RL fitted participants' choices better than the Bayesian RL model, which assumes that people retrospectively update their state belief after observing reward signals.

It is also possible that the effectiveness of different learning modes will depend on the level of state uncertainty, with higher levels requiring belief-state based style of learning to avoid perceptual noise corrupting, or at least slowing down, learning (Bruckner et al., 2020). This is partially supported by our findings in the first experiment (see S.7), where the simple and Bayesian RL models were almost indiscernible with a greater degree of uncertainty ($\rho = 0.65$ instead of 0.7). It is also possible that the training that participants went through in the calibration stage primed participants to adopt a strategy where they commit to a state before making the Q-value updates and choosing an action.

5.3. Relationship with sampling based learning

Even though RL has been shown to be a viable proxy for human reward-based learning under expected reward uncertainty (e.g., Tanaka et al., 2004; Paulus et al., 2004), a recent body of work has found support for learning by sampling mechanism (Hochman & Erev, 2013; Plonsky et al., 2015; Bornstein et al., 2017; Bornstein & Norman, 2017; Hotaling et al., 2022). The finding that the RL models fit the data better than the sampling model in the reward uncertainty condition is thus quite surprising, especially given that the sampling model also has the most flexible mechanism of choices among the three models considered in this study. In fact, the sampling model can easily generate different patterns of learning under expected and unexpected uncertainty by increasing or decreasing the weight of the overall average payoff relative to the payoff over small samples. For example, increasing this weight would make the adaptation to the switch slower, whereas relying more on recent samples should speed up the adaptation to the switch. So how can we reconcile our findings, which support RL, with the emerging literature that supports sampling-based learning?

First, the two models fit the data in the reward uncertainty condition very closely, especially at the aggregate level (close match between the overall learning curves and even slightly better median BIC score for the sampling model). As Plonsky et al. (2015, p. 640) argued, the similarity between these two classes of models "are more important than the differences". One important similarity is relying on small samples of experiences to respond to similar contexts encountered in the past. In fact, in RL, a state-action value can be seen as a weighted average of the rewards obtained for the same state-action pair with weights decaying exponentially with the learning rate α, or more formally, $Q_t(s,a) = \alpha \sum_{i=1}^{t} (1-\alpha)^i r_{i-1}$. This means that reward information received further in the past will have a negligible impact on the present decision, especially when the learning rate is large (Bornstein et al., 2017). Other similarities include behaving near optimally when the similarity between contexts can be measured by the learner and underweighting (unpredictable) rare events.

Second, and as noted also by Plonsky et al. (2015), the differences between the two learning mechanisms might arise simply due to the specific experimental manipulations or the modelling

choices that one makes. For example, relying on decaying averages of past rewards as in RL models is likely to lead to poor results in comparison with a learning-by-sampling approach when the relevant past experiences are sparse or/and scarce (Bornstein et al., 2017). Our experimental paradigm does not fall in this category as reward is consistent and fairly stable (a fixed uncertainty level with only one changepoint). Bornstein and colleagues (Bornstein et al., 2017; Bornstein & Norman, 2017), however, used a reward function that changed continuously according to a random walk (either in terms of magnitude or probability of delivery). Another difference related to experimental design is that all studies supporting sampling-based learning over RL except that of Bornstein & Norman (2017) did not have the contextual component that we had in the present experiment (i.e., multiple states), and simply employed "multi-armed bandit" paradigms with stochastic rewards.

Other differences related to modelling choices could also explain our seemingly divergent findings. For example, Hotaling et al. (2022) compared four different versions of a sampling model against a basic RL model that does not have a mechanism for action exploration such as softmax or ϵ-greedy (compare this to sampling models, which have an inherent randomness allowing them to generate variability in choices) across five experiments. They reported that none of the four versions of their sampling model consistently offered the best quantitative fit to participants' data across all their five experiments (each of the four model versions was the best fitting model in at least one experiment). Their study, however, demonstrates that their sampling framework captures a wide range of behavioural patterns observed in their experimental paradigms such as people making riskier choices when rare rewards are highlighted and choices being affected by outcome order. Hochman & Erev (2013) found that a sampling-based model accounted for human choices in a two-armed bandit task better than an RL model, using a model comparison procedure focused on the predictive capacity of the models (they generated choice predictions from the sampling and RL models then compared the predicted choices from each model to participants' actual choices using mean squared errors).

Besides, the sampling frameworks compared to RL in the literature tend to be simple with one or two free parameters. We chose the BISAW model because, as we mentioned earlier, it has a very flexible mechanism of choices that seems suitable for our experimental paradigm, albeit with the cost of perhaps being overparametrised relative to the two other RL models. We, thus, do not rule out the possibility that other types of sampling models might fit the data presented in this study better. It is also possible that we are in a situation where the data are sufficiently straightforward for the simpler RL model to mimic human learning patterns. Although the sampling model may be able to capture these patterns just as well (or almost just as well), it may be penalised for being able to account for a greater variety of patterns that were never observed (but that may be observed in other experiments; e.g. Hotaling et al., 2022). That said, our model comparison findings, especially in the state uncertainty condition, and in particular the generalised additive modelling analysis that tested the ability of the models to capture the experimental patterns observed in our second experiment, provide strong support for the RL model in comparison with the sampling framework that we used.

5.4. Limitations and future directions

In our work, we considered only model-free RL models. Model-based RL (Daw et al., 2005, 2011; Doll et al., 2012), which assumes that the agent learns an internal model (or representation) of the task structure, is another framework that could be used to shed further light on the mechanisms involved in learning under uncertainty. One of the most heavily studied tasks for probing for model-based control is the 'two-step' task, proposed by Daw et al. (2011). In this task, a participant starts each trial by selecting between two choices, each of which then leads

probabilistically to one of two states (each choice has a 0.7 probability of transitioning to the "common" state and a 0.3 probability of transitioning to the "rare" state). The participant then has to make a second-stage choice between two options depending on the current state, and is rewarded with a particular probability. The original version of the task also employed dynamic rewards that changed according to a Gaussian random walk. Daw and colleagues showed that human participants engaged in both model-free and model-based RL while solving the 'two-step' task, and identified neural correlates of model-free and model-based valuations.

An interesting hypothesis to test in future investigations is whether participants in our state uncertainty task engage in a two-stage decision making as in the 'two-step' task: on each trial, participants would first decide whether the stimulus is light or dark (identification stage), then they would transition to a hidden state that either matches their identified state (with probability $\rho = 0.7$ as in Experiment 2) or not (with probability 0.3). The simple RL model also assumes an identification stage but lacks the ability to track the second-stage transitions. Incorporating a model-based approach that learns such a graph-like structure could be key to allowing faster and more robust adaptation to changes in the reward distribution (Daw et al., 2011).

Another open question from this study is why some participants performed at a higher level than the target level of state uncertainty ($\rho = 0.65$ in Experiment 1 and $\rho = 0.7$ in Experiment 2), despite the several measures we introduced in Experiment 2 to prevent state disambiguation learning. One compelling idea is that people rely on a complex learning mechanism that combines reward-based learning with state disambiguation learning. That is, there is a process of learning to identify the states, which goes hand in hand with reward-based learning, where reward feedback is used to improve state disambiguation. Under state uncertainty, having a deterministic distribution of reward provides valuable feedback to participants to resolve the state uncertainty. This idea forms the basis of the Bayesian model—rewards can be used to (a posteriori) re-evaluate state identification—but the model does not have a mechanism for using this reevaluation to improve state identification. Lak et al.'s (2017) study supports this interpretation (i.e. reward feedback can be used to concurrently learn how to take actions and how to better identify the uncertain states) since they used a belief-state based RL model to estimate perceptual confidence in monkeys, and found this estimate to correlate with monkeys' dopamine response. Given that dopamine is strongly linked to reward-based learning, this might suggest that there are two dopamine-dependent learning processes taking place simultaneously: one for learning action-state values, and one for learning to disambiguate uncertain states. One way to test whether participants' perceptual categorisation improved during the learning task is to ask participants to explicitly report the state they perceive before choosing a motor response, as done in Bruckner et al. (2020), although this additional requirement may artefactually influence the primary learning task.

One related question is why exploration rates were higher in the reward uncertainty task (vs state uncertainty), which ultimately led to people receiving less reward because they were not maximising. One suggestion for the phenomenon of probability matching in choice tasks is that people treat tasks such as ours as problem solving tasks and effortfully search for patterns, even in random sequences (e.g., Wolford et al., 2004; Gaissmaier & Schooler, 2008). It has been found that deploying a secondary task can move participants away from probability matching to maximising (Wolford et al., 2004). The greater exploration under reward uncertainty in our study might reflect that this task was less demanding and thus allowed more cognitive resources for the use of an effortful problem solving approach that resulted in probability matching. Such questions could be addressed by increasing the challenge of the task by including simultaneously both state and reward uncertainty.

Conclusion

To conclude, this study goes beyond previous studies on reward-based learning under uncertainty by contrasting state uncertainty with reward uncertainty and combining them with unexpected uncertainty due to a non-signalled reversal of stimulus-response contingencies. We have shown that human participants are capable of successfully dealing with all these different types of uncertainties and that they perform better under state uncertainty in comparison with reward uncertainty, due to a higher exploration under reward uncertainty and their ability to improve their perceptual categorisation under state uncertainty by using the reward feedback. Our findings also support a common reinforcement learning mechanism in both uncertainty conditions, whereby learners commit to a state and only update state-action value of the identified state.

Supplementary Information

The article has accompanying supplementary material.

Acknowledgements

We would like to thank Charles Hanich for his help with the data collection of the second experiment, Andreas Jarvstad for his help with Psychtoolbox while programming the first experiment and the members of the 'Decision making in an unstable world' project for their insightful discussions and for their help with piloting the first experiment. We also thank Scott Brown and three anonymous reviewers for their insightful comments on this work during the review process.

Declarations

Availability of Data and Materials

The datasets generated and analysed during the current study are available from the Open Science Framework at https://osf.io/43jx8/?view_only=8ccae6aaabd74faeb61be1a82989174f.

Code availability

The code that was used to produce the findings of this study is openly available from the Open Science Framework at https://osf.io/43jx8/?view_only=8ccae6aaabd74faeb61be1a82989174f.

Author contributions

AE, SF, DL, GM and CL conceptualised the study aims and design. DL, CL and SF acquired the funding. CL and SF obtained ethical approval. AE programmed the tasks, recruited participants and collected data. AE curated and visualised the data, implemented the computational models and conducted all statistical analyses. SF, DL, GM and CL provided insights in how to pose questions and analyse the data. AE wrote the original draft of the manuscript. All authors reviewed the final manuscript.

Funding

This research was carried out as part of the project 'Decision making in an unstable world', supported by the Engineering and Physical Sciences Research Council (EPSRC; EP/1032622/1). CL received additional support from a Leverhulme Research Fellowship (RF-2021-168). SF received additional support from the Australian Research Council (FT1301000149 and DP160101752).

Ethics Approval

The first experiment was approved by the Faculty of Science Human Research Ethics Committee at the University of Bristol. The second experiment was approved by the Human Research Ethics Committee at the University of Western Australia.

Consent to Participate and Consent for Publication

Informed consent for participation and publication was obtained from all individual participants included in the two experiments.

Conflict of Interests

The authors declare no competing interests.

References

Babayan, B. M., Uchida, N., & Gershman, S. J. (2018). Belief state representation in the dopamine system. *Nature communications*, *9*, 1–10.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature reviews neuroscience*, *13*, 572–586.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*, 1214–1221.

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience*, *6*.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*, 15958.

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*, 997.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Bruckner, R., Heekeren, H. R., & Nassar, M. R. (2022). Understanding learning through uncertainty and bias, .

Bruckner, R., Heekeren, H. R., & Ostwald, D. (2020). Belief states and categorical-choice biases determine reward-based learning under perceptual uncertainty. *bioRxiv*, .

Chen, W., Liu, S.-Y., Chen, C.-H., & Lee, Y.-S. (2011). Bounded memory, inertia, sampling and weighting model for market entry games. *Games*, *2*, 187–199.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, *23*, 3–38.

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*, 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*, 1704–1711.

D'Errico, J. (2005). fminsearchbnd an extension to fminsearch. http://www.mathworks.com/matlabcentral/fileexchange/8277-fminsearchbnd–fminsearchcon. Latest update 2012.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, *22*, 1075–1081.

Don, H. J., Otto, A. R., Cornwall, A. C., Davis, T., & Worthy, D. A. (2019). Learning reward frequency over reward probability: A tale of two learning rules. *Cognition*, *193*, 104042.

Farhi, E., Debab, Y., & Willendrup, P. (2014). ifit: A new data analysis framework. applications for data reduction and optimization of neutron scattering instrument simulations with mcstas. *Journal of Neutron Research*, *17*, 5–18.

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *Journal of Neuroscience*, *33*, 19060–19070.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*, 416–422.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* volume 43. CRC press.

Hochman, G., & Erev, I. (2013). The partial-reinforcement extinction effect and the contingent-sampling hypothesis. *Psychonomic bulletin & review*, *20*, 1336–1342.

Hotaling, J. M., Donkin, C., Jarvstad, A., & Newell, B. R. (2022). Mem-ex: An exemplar memory model of decisions from experience. *Cognitive Psychology*, *138*, 101517.

Jaakkola, T., Singh, S. P., & Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, (pp. 345–352).

Jeffreys, H. (1961). *The theory of probability*. (3rd ed.). OUP Oxford.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*, 773–795.

Kingdom, F., & Prins, N. (2009). *Psychophysics: A Practical Introduction*. Academic Press.

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Current Biology*, *27*, 821–832.

Lak, A., Okun, M., Moss, M. M., Gurnani, H., Farrell, K., Wells, M. J., Reddy, C. B., Kepecs, A., Harris, K. D., & Carandini, M. (2020). Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron*, *105*, 700–711.

Larsen, T., Leslie, D., Collins, E., & Bogacz, R. (2010). Posterior weighted reinforcement learning with state uncertainty. *Neural Computation*, *22*, 1149–1179.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*, 923–937.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in Human Neuroscience*, *8*, 825.

Mella, V. S. A., Possell, M., Troxell-Smith, S. M., & McArthur, C. (2018). Visit, consume and quit: Patch quality affects the three stages of foraging. *Journal of Animal Ecology*, *87*, 1615–1626.

Moore, J. W., Choi, J.-S., & Brunzell, D. H. (1998). Predictive Timing under Temporal Uncertainty: The Time Derivative Model of the Conditioned Response. In *Timing of Behavior: Neural, Psychological, and Computational Perspectives* (pp. 3–34). The MIT Press.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*, 12366–12378.

Paulus, M. P., Feinstein, J. S., Tapert, S. F., & Liu, T. T. (2004). Trend detection via temporal difference model predicts inferior prefrontal cortex activation during acquisition of advantageous action selection. *Neuroimage*, *21*, 733–743.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology*, *7*, e1001048.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, *79*, 191–201.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*, 437–442.

Piray, P., & Daw, N. D. (2020). A simple model for learning in volatile environments. *PLoS computational biology*, *16*, e1007963.

Piray, P., & Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature communications*, *12*, 1–16.

Platt, M. L., & Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nature neuroscience*, *11*, 398–403.

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological review*, *122*, 621.

Prins, N., & Kingdom, F. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the palamedes toolbox. *Frontiers in psychology*, *9*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. Technical Report CUED/F-INFENG/TR166 Department of Engineering, University of Cambridge.

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*, 635–644.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature neuroscience*, *20*, 581–589.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*, 1–26.

Stocker, A. A., & Simoncelli, E. (2007). A bayesian model of conditioned perception. *Advances in neural information processing systems*, *20*, 1409–1416.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*, 887–893.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2020). itsadug: Interpreting time series and autocorrelated data using gamms. R package version 2.4.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of economic surveys*, *14*, 101–118.

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*, 189.

Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *58*, 221.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692.

Supplementary material

S.1. Clustering participants' patterns of responses

To better understand the patterns of responses of our participants, especially how they adapted to the switch, we grouped their learning curves using hierarchical clustering with a weighted Euclidean distance and average linkage method. Each participant was represented with a vector of dimension 171. To better detect the different patterns of switch adaptation, we weighted the post-switch trials (i.e. between 101–171) in the calculation of the Euclidean distance by a factor of two. We selected the number of clusters to use visually, upon inspection of the hierarchical tree, where we chose a cut that separates the different groups well both in the hierarchical tree and in the resulting groups of learning curves. For the implementation of the procedure, we used Matlab's clustering function, *cluster()*.

S.1.1. Experiment 1: Test group

Figure S.2 shows four different response patterns (the resulting hierarchical tree is depicted in Figure S.1): (1) there were participants who adapted quickly to the switch (i.e., within about 30 trials; see panel labelled as Cluster 1), those who learned well pre-switch, but could not recover completely from the switch (Cluster 4), (3) two participants who managed to learn the rules before the switch and either adapted late to the switch or had an oscillating learning behaviour after the switch (Cluster 2), and (4) one participant who discovered the correct rules late in the first phase and did not recover from the switch (Cluster 3).

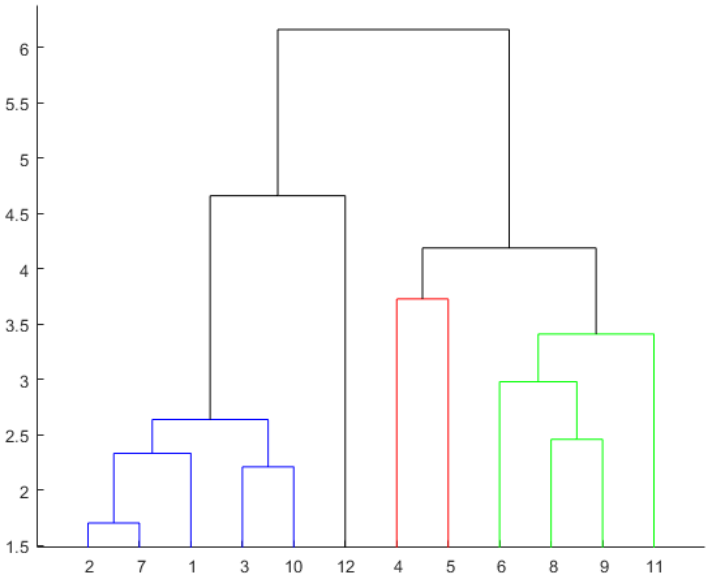

Figure S.1: Hierarchical clustering tree in the test condition. Different colours represent different clusters.

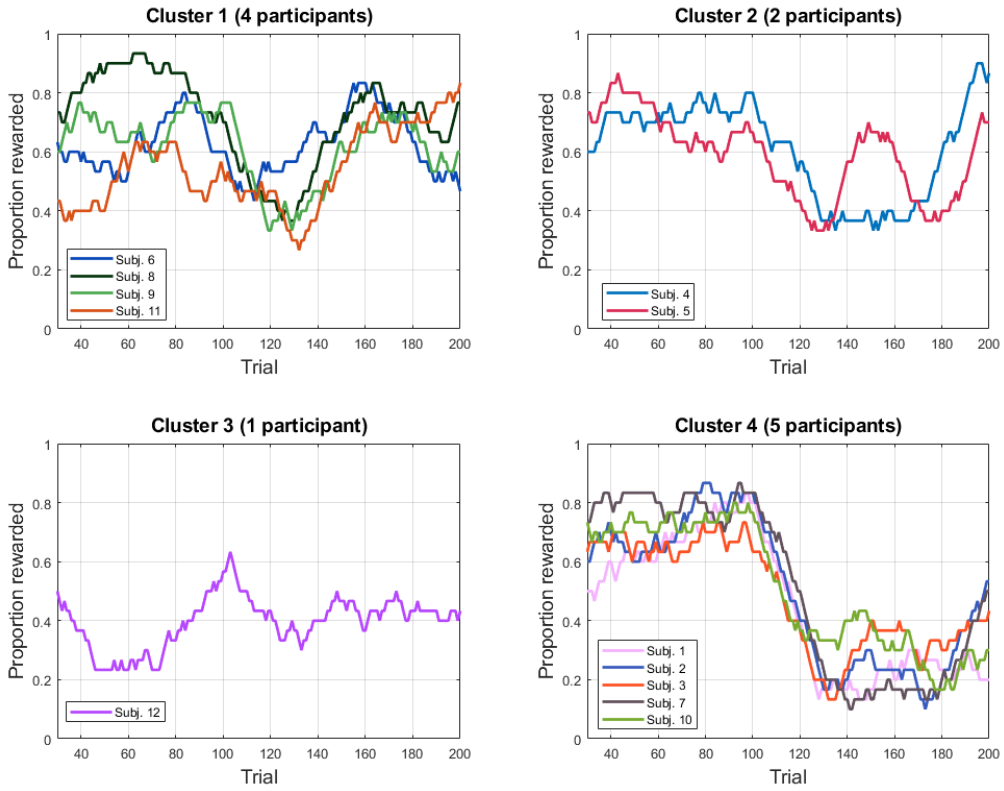

Figure S.2: Individual learning curves categorised by response pattern in the test condition.

S.1.2. Experiment 2: State uncertainty condition

As in the test phase of Experiment 1, participants' data suggests different adaptation patterns in the state uncertainty condition. Based on the hierarchical clustering tree (Figure S.3), we identified eight clusters. For example, as shown in Figure S.4, some participants adapted slowly (Cluster labelled as 2 in Figure S.4) or did not adapt at all (Cluster 8). Many participants adapted to the switch quickly (clusters 4 and 5), with a few showing a markedly larger dip in performance (Cluster 5). The larger dip for the participants in Cluster 5 indicates that they persisted with the old mapping for relatively longer. In contrast, participants in Custer 4 seem to have had a period of unlearning where their performance dropped to the chance level, akin to how they performed at the start of the task. Cluster 3 is characterised by learning patterns where there is little or no drop in performance after the switch. Overall most of the clusters' patterns of responses are easily identifiable from the learning curves.

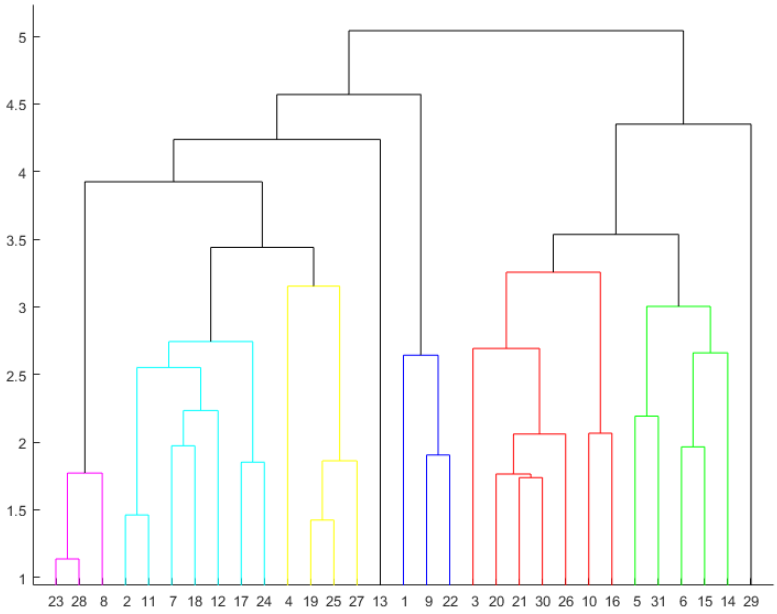

Figure S.3: Hierarchical clustering tree in the state uncertainty condition. Different colours represent different clusters.

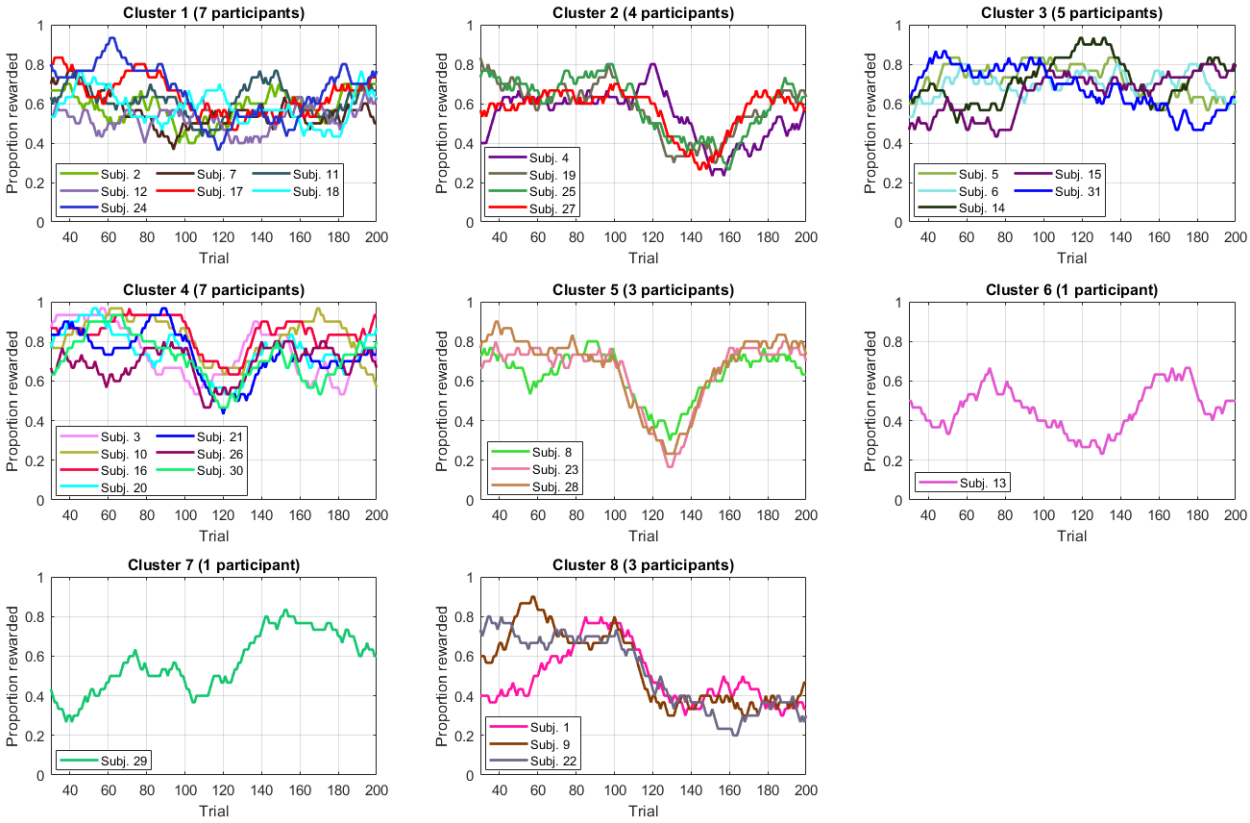

Figure S.4: Individual learning curves categorised by response pattern in the state uncertainty condition.

S.1.3. Experiment 2: Reward uncertainty condition

In the reward uncertainty, the data seems noisier and the seven identified clusters look largely different from the ones identified in the state uncertainty condition (see Figure S.6). In particular, there are no easily identifiable groups based on the patterns of adaptation as before. This is probably a reflection of the higher exploration in this condition as shown in section 4.1.4.

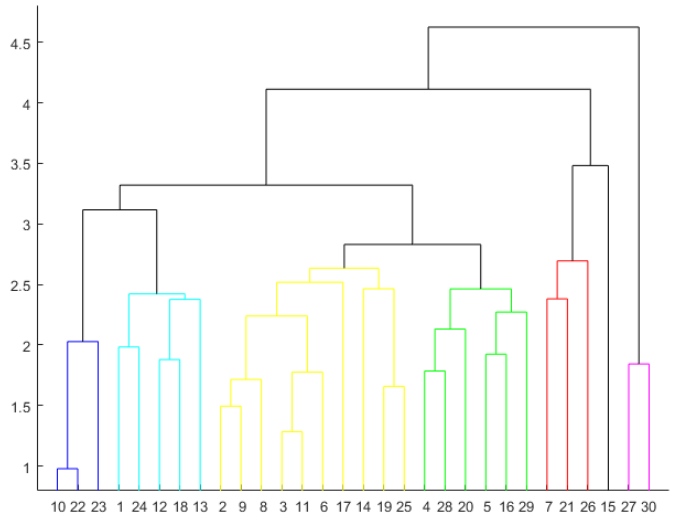

Figure S.5: Hierarchical clustering tree in the reward uncertainty condition. Different colours represent different clusters.

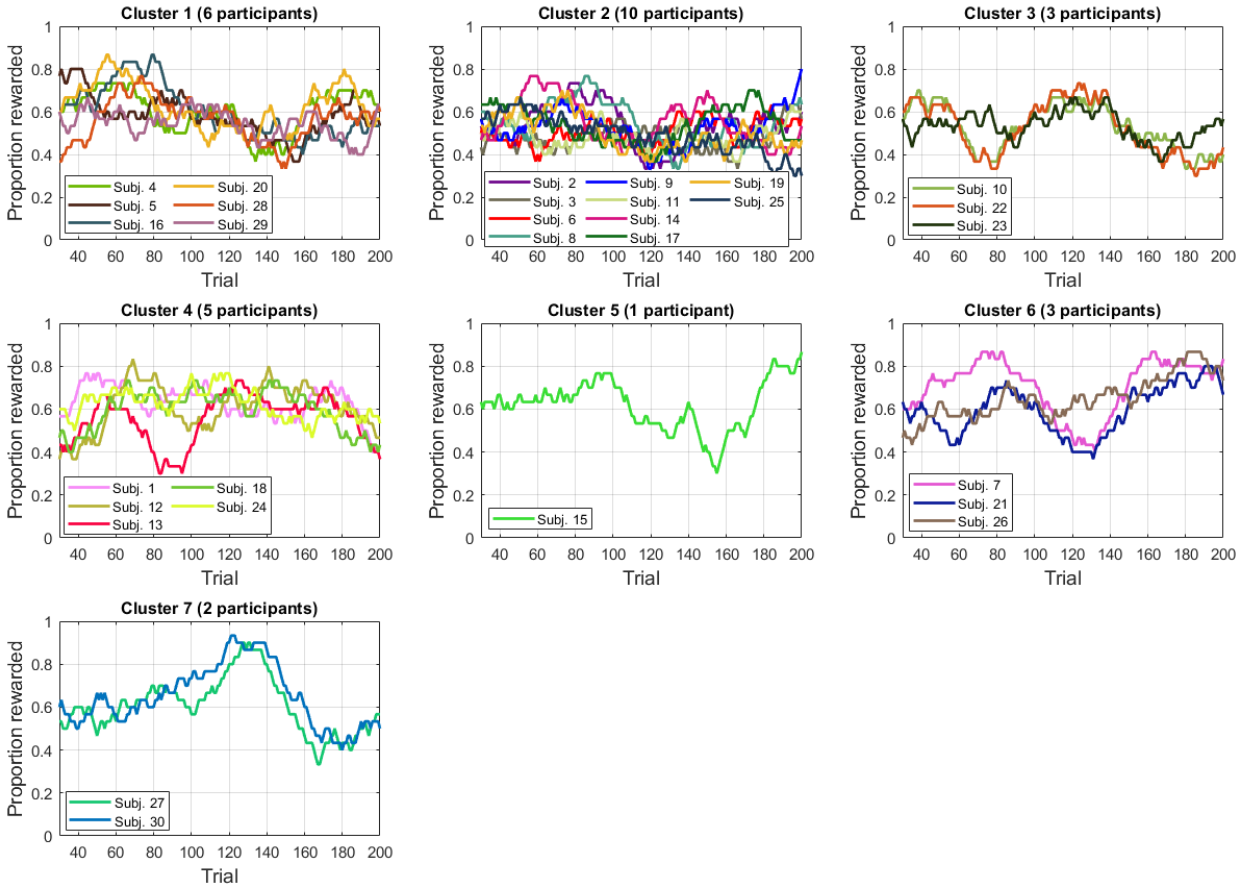

Figure S.6: Individual learning curves categorised by response pattern in the reward uncertainty condition.

S.2. Simulation results for validating the uncertainty level and sample size in Experiment 2

The data from the first experiment indicated large differences among participants in terms of the pattern of adaptation to the switch. To reduce the extent of individual differences, and thus focus on the comparison of the performance between the state uncertainty and reward uncertainty conditions, we decided to increase the level of uncertainty, ρ, slightly and opted for two possible values: 0.7 or 0.75. We did not include higher values to keep the task challenging enough for participants and avoid a ceiling effect. To select between the two ρ values, we conducted a parameter/model recovery analysis to see which value leads to the highest accuracy of model recovery. Another aim of the recovery analysis was to validate the sample size for Experiment 2. More specifically, we wanted to check whether a sample size of 30 would be enough to differentiate between our models (the selected sample size was based on previous studies of reward-based learning under uncertainty, which had tested around 30 to 40 participants; e.g., Behrens et al., 2007; Nassar et al., 2010; Wilson & Niv, 2011).

The model recovery analysis was based on the two RL models only (the sampling model was not included). We generated 30 artificial participants from each of the two models, then fit them back by both models and assessed the quality of model fit using BIC scores. This procedure was run separately with $\rho = 0.7$ (Figure S.7) and $\rho = 0.75$ (Figure S.8). In each case, we generated 10 artificial participants using parameters from the top, mid and low range performance with Gaussian noise. For both values of ρ, there was a good separation between the two models as judged by the BIC scores (right panels of Figure S.7 and S.8). The accuracy of recovery in the

case of $\rho = 0.7$ was 91.7%, while it was equal to 90% in the case of $\rho = 0.75$. We thus used 0.7 as the level of uncertainty in Experiment 2.

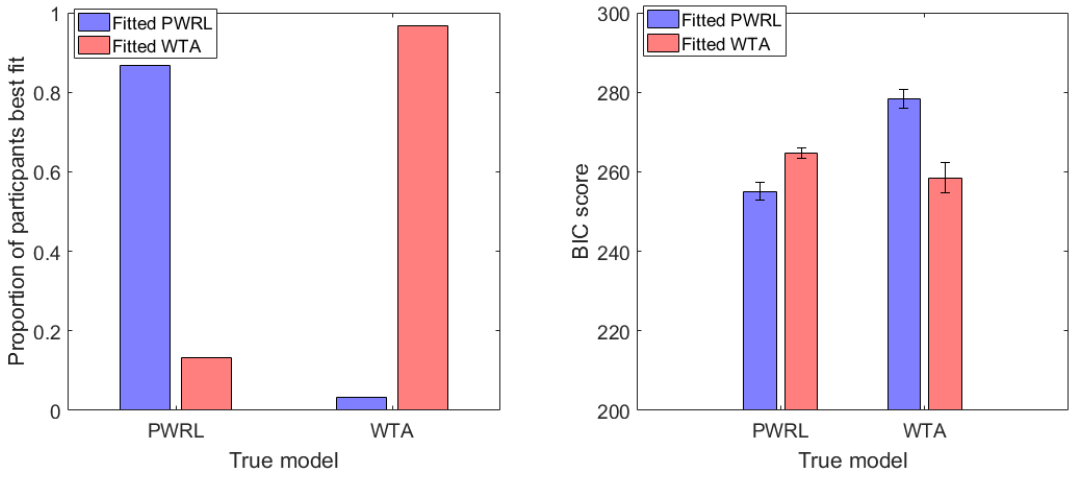

Figure S.7: Model recovery results for $\rho = 0.7$. **Left:** Proportion of artificial participants that each model best fit for each generating model. **Right:** BIC scores by model for each generating model.

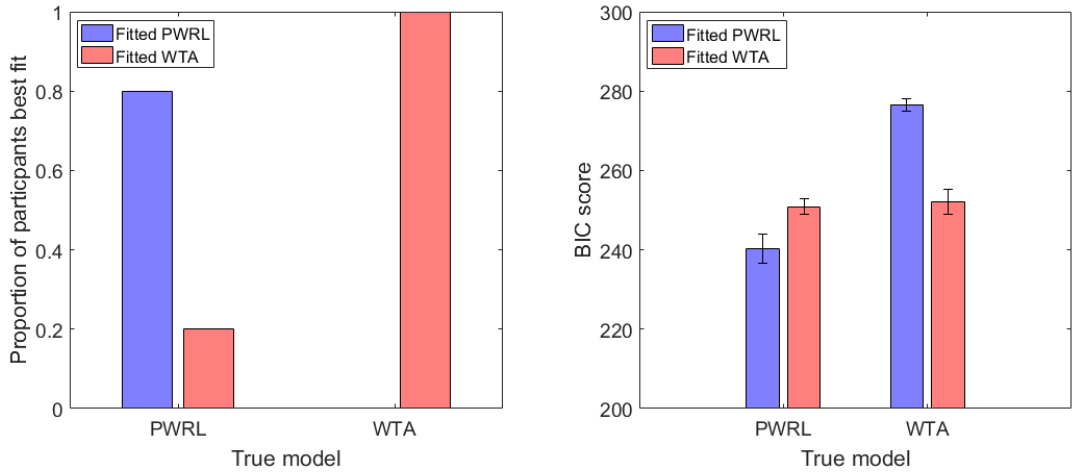

Figure S.8: Model recovery results for $\rho = 0.75$. **Left:** Proportion of artificial participants that each model best fit for each generating model. **Right:** BIC scores by model for each generating model.

S.3. Detailed formulation of the PWRL model under state uncertainty

As presented in the main text (see Section 4.1.1), the PWRL model updates the state-action values $Q_t(s, a)$, for each state s and current action a_t as follows:

$$Q_{t+1}(s, a_t) = Q_t(s, a_t) + \alpha \mathbb{P}(S_t = s | r_t, a_t, i_t, \boldsymbol{Q_t})(r_t - Q_t(s, a_t)) \tag{S.1}$$

where α is the learning rate, i_t is the state that is identified by the learner, $\boldsymbol{Q_t}$ is the vector of all state action values and $\mathbb{P}(S_t = s | r_t, a_t, i_t, \boldsymbol{Q_t})$ is the posterior probability that the true state is s once the reward r_t has been observed, which using Bayes rule, can be re-written as:

$$\mathbb{P}(S_t = s | R_t = r_t, a_t, i_t, \boldsymbol{Q_t}) = \frac{\mathbb{P}(R_t = r_t | s, a_t, \boldsymbol{Q_t}) \rho_t(s, i_t)}{\sum_{s'} \mathbb{P}(R_t = r_t | s', a_t, \boldsymbol{Q_t}) \rho_t(s', i_t)} \tag{S.2}$$

where $\rho_t(s,i)$ is the probability with which the learner identifies i as the true state, such that her identification i is correct with probability ρ. More precisely:

$$\rho_t(s,i) = \mathbb{P}(I_t = i | S_t = s) = \begin{cases} \rho & \text{if } i = s, \\ 1-\rho & \text{otherwise.} \end{cases} \tag{S.3}$$

Since we are treating the simple case of a deterministic reward R_t (wining 2.5 pennies or nothing), R_t can be considered as a Bernoulli random variable, and hence the probability of each outcome can be estimated using $Q_t(s,a_t)$, the expected value of R_t, as follows:

$$\mathbb{P}(R_t = r_t | s, a_t, \boldsymbol{Q_t}) = \begin{cases} Q_t(s,a_t)/2.5 & \text{if } r_t = 2.5, \\ 1 - Q_t(s,a_t)/2.5 & \text{if } r_t = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{S.4}$$

Using this expression, we can re-write Equation 3 in a more explicit way as:

$$Q_{t+1}(s,a_t) = \begin{cases} Q_t(s,a_t) + \alpha \frac{Q_t(s,a_t)\rho_t(s_t,s)}{\sum_{s'} Q_t(s',a_t)\rho_t(s_t,s')}(2.5 - Q_t(s,a_t)) & \text{if } r_t = 2.5, \\ Q_t(s,a_t) + \alpha \frac{(2.5-Q_t(s,a_t))\rho_t(s_t,s)}{\sum_{s'}(2.5-Q_t(s',a_t))\rho_t(s_t,s')}(0 - Q_t(s,a_t)) & \text{if } r_t = 0. \end{cases} \tag{S.5}$$

Until now, we have only shown how the PWRL model updates the state-action values using a modified version of temporal difference learning that takes into account the uncertainty in the state identification. There remains the question of how the learner selects her actions given the estimated Q-values. Here, we assume that they do it according to the softmax selection rule:

$$\mathbb{P}(A_t = a | I_t = i_t) = \frac{\exp(Q(i_t,a)/T)}{\sum_{a'} \exp(Q(i_t,a')/T)} \tag{S.6}$$

where T is a temperature parameter, which determines the degree of stochasticity in action selection. Note that Equation S.6 shows how to compute the distribution of action selection given an identified state i_t, and not the true state s_t. In fact, the true state s_t is not known to the learner with certainty, but can only be correctly identified with probability ρ. In order to get the probabilities of selecting each action a marginalising out the true state s_t, we average out Equation S.6 according to the probability of detecting each state given the true state, $\rho_t(s_t,i') = \mathbb{P}(I_t = i' | S_t = s_t)$. This gives:

$$\mathbb{P}(A_t = a | S_t = s_t) = \sum_{i'} \frac{\exp(Q(i',a)/T)}{\sum_{a'} \exp(Q(i',a')/T)} \rho_t(s_t,i') \tag{S.7}$$

where we assume that this is the softmax distribution that the learner uses to select actions.

S.4. Model evaluation

The estimation of each model's parameters was carried out using maximum likelihood estimation. For the WTA-RL and BI-SAW models, which include at least one sampling step (e.g. when the learner identifies the current state in the state uncertainty condition), we did not have an explicit formula for computing the likelihoods. Instead, we estimated the probability of each observed response by simulating a one-step ahead model prediction of that choice a number of times, then recording the proportion of times the model accurately predicted that choice. More precisely, to approximate the likelihood of a participant's choice at time t, we provided the model with the sequence of the participant's responses, rewards and observations until $t-1$ along with

the observation at time t, and used that to generate a model response. We repeated this 1000 times, then computed the proportion of times the observed choice was accurately predicted by the model. Although we could compute choice likelihoods directly for the PWRL model (and SARSA, its equivalent in the reward uncertainty condition), we used the same stochastic likelihood estimation approach that we used for the WTA-RL and BI-SAW models. This was to avoid any bias or extra variability that could arise from intermixing exact and stochastic likelihood estimation when comparing our models.

For model comparison, we converted Bayesian information criterion (BIC) scores to Bayes factor as described in Section 4.1.4. In terms of implementation, we programmed the fitting procedure in Matlab using the simulated annealing-based optimisation function fminsimpsa (iFit library; Farhi et al., 2014), which was found to produce more stable and reliable results than the classical constrained optimisation function fminsearchbnd (D'Errico, 2005), given the noise generated by the sampling steps in our models. We constrained all parameters as outlined in Table S.1. To avoid finding local optima, we run the fitting procedure for each model with 30 randomly generated initial parameter values.

Condition	Model	Parameter	Constraint
State uncertainty	PWRL	α	$0 \leq \alpha \leq 1$
		T	$0 \leq T < 100$
		ρ	$0.5 \leq \rho < 1$
	WTA-RL	α	$0 \leq \alpha \leq 1$
		T	$0 \leq T < 100$
		ρ	$0.5 \leq \rho < 1$
	BI-SAW	ω	$0 \leq \omega \leq 1$
		p_r	$0 \leq p_r \leq 1$
		b	$0 \leq b \leq 10$
		T	$0 \leq T < 100$
		ρ	$0.5 \leq \rho < 1$
Reward uncertainty	SARSA	α	$0 \leq \alpha \leq 1$
		T	$0 \leq T < 100$
	BI-SAW	ω	$0 \leq \omega \leq 1$
		p_r	$0 \leq p_r \leq 1$
		b	$0 \leq b \leq 10$
		T	$0 \leq T < 100$

Table S.1: Constraint used for each parameter in each of the three learning models.

S.5. Categorising learning curves by the best fitting model

We examined participants' learning curves according to the model by which they were best fit, in the state uncertainty condition (Figure S.9). This revealed that the PWRL model best fit mainly participants who did not adapt (participants 9 and 22 in Panel C) or those who had a large dip in performance after the switch (participants 8, 23 and 28 in Panel C). These two categories of learners were also picked up by the hierarchical clustering analysis of learning curves reported in Figure S.4 (see panels labelled as clusters 5 and 8). The WTA-RL model seemed to capture participants who had a better overall performance across the task trials (see Panels B and D), most noticeably those who adapted quickly to the task (see panels labelled as clusters 3 and 4 in Figure S.4). These results suggest that, although most of our participants' learning

patterns in the state uncertainty condition were best captured by the simple RL model, different types of learning patterns for different participants characterise each of the two RL mechanisms.

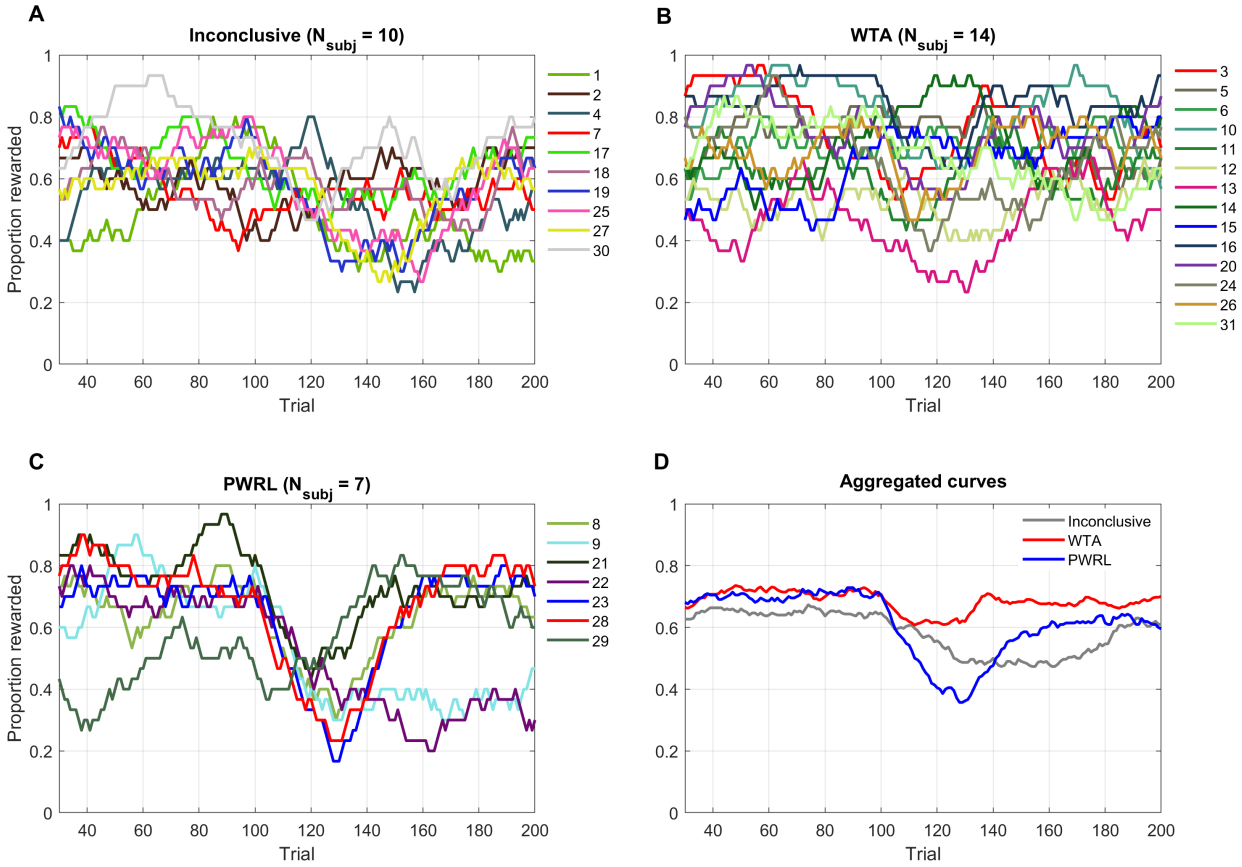

Figure S.9: Fitted learned curves in the state uncertainty condition categorised by the best fitting model (A-C). Panel D depicts the aggregated learning curve in each category.

In the reward uncertainty condition (Figure S.10), differentiating between the learning patterns in each best-fitting model group is difficult due to the high exploration rate as noted previously. The aggregated learning curves displayed in Panel D suggests quite similar overall performance levels within the three groups, with the main difference coming from the group of participants who were best fit by the sampling model, which had a higher performance in the 40-50 trials post switch. Linking the groups of patterns best fit by each model to the clusters identified through the hierarchical clustering analysis reported in Figure S.6, we can see that SARSA was the best fitting model for two clusters (see panels labelled as clusters 1 and 2).

S.6. Statistical effect recovery analysis using the three computational learning models

Here we asked how well the three models could recover the effects observed in the GAMM analysis of the behavioural data from Experiment 2 reported in Section 3.3.3. To answer this, we fit binomial GAMMs to data simulated from each model. The simulated data consisted of the proportion of rewarded instances (corresponds to the success rate) and non-rewarded instances (failure rate) in each trial, calculated based on 100 simulation runs using the best fitting parameters for each participant (these are the same 100 simulation runs we carried out to construct the learning curves reported in Figure 10). Each of the three GAMMs that were fit to the simulated data (henceforth referred to as the "simulated" GAMMs as opposed to the

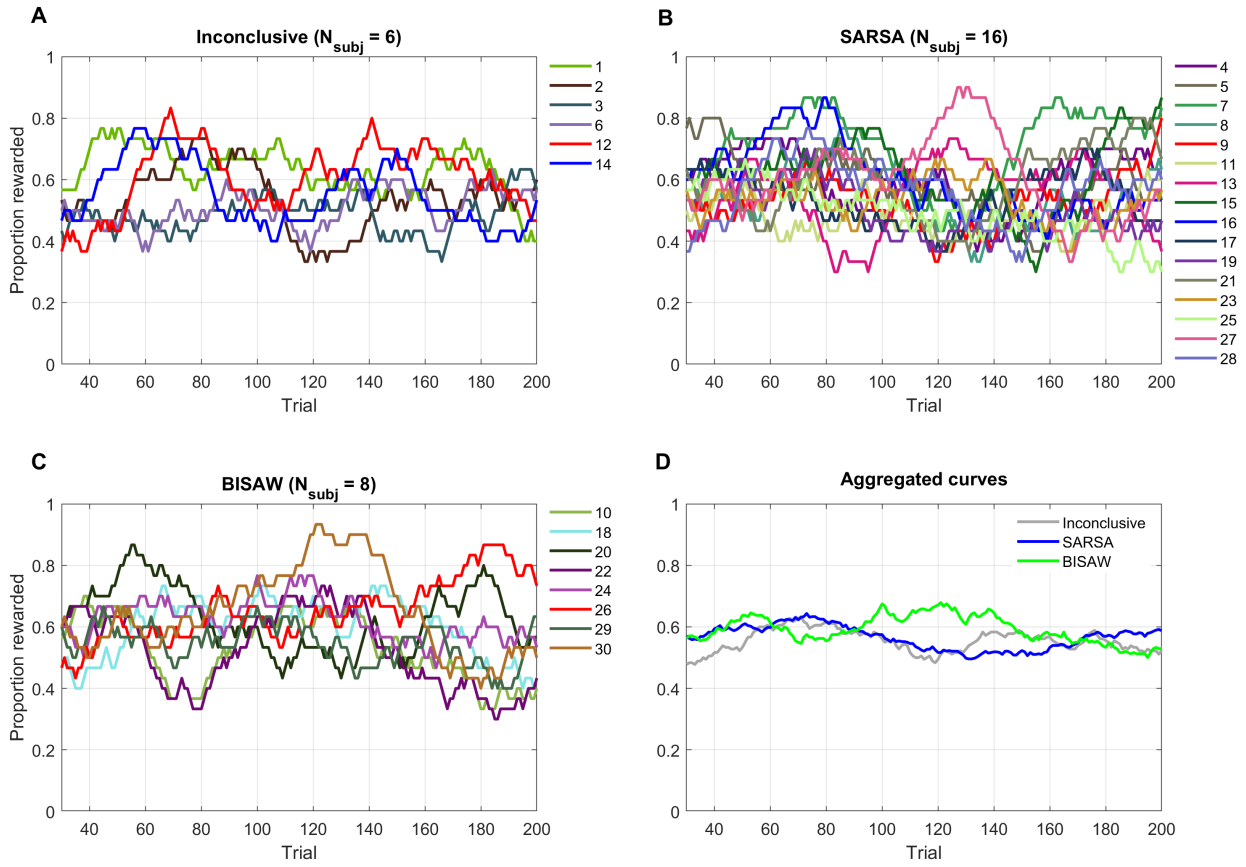

Figure S.10: Fitted learned curves in the reward uncertainty condition categorised by the best fitting model (A-C). Panel D depicts the aggregated learning curve in each category.

"behavioural" GAMMs reported in Section 3.3.3) included exactly the same fixed and random structures as the GAMM that we previously constructed to analyse the behavioural data. Table S.2 shows the p-values of all predictors for the "behavioural" and each of the three "simulated" GAMMs. The simple RL reproduced all but one predictor's significance more than any of the two other computational models (the model missed the trial-smooth corresponding to the state uncertainty condition and matched states; the trial-smooth corresponding to the reward uncertainty condition and matched states shows as significant in the behavioural GAMM but this is most likely due to the large initial dip as highlighted previously). The largest number of mismatches in effects' significance was found for the BISAW-based GAMM, with four predictors coming out significant, while they were reported non-significant in the behavioural GAMM.

The simple RL also reproduced the fixed effects observed in the behavioural data reasonably well, as can be seen in Figure S.11 (compare with Figure 9). Perhaps the most noticeable mismatch is that of the effect difference between matched and unmatched states in the reward uncertainty condition, where the difference came out significant in about two-third of the first block of 100 trials, while the difference is non-significant throughout the task in the behavioural data (see panel D in Figure 9).

These results provide further support for the simple RL model as the best model for how our participants behaved in both uncertainty conditions, and rule out the sampling model as a potential best-fitting model.

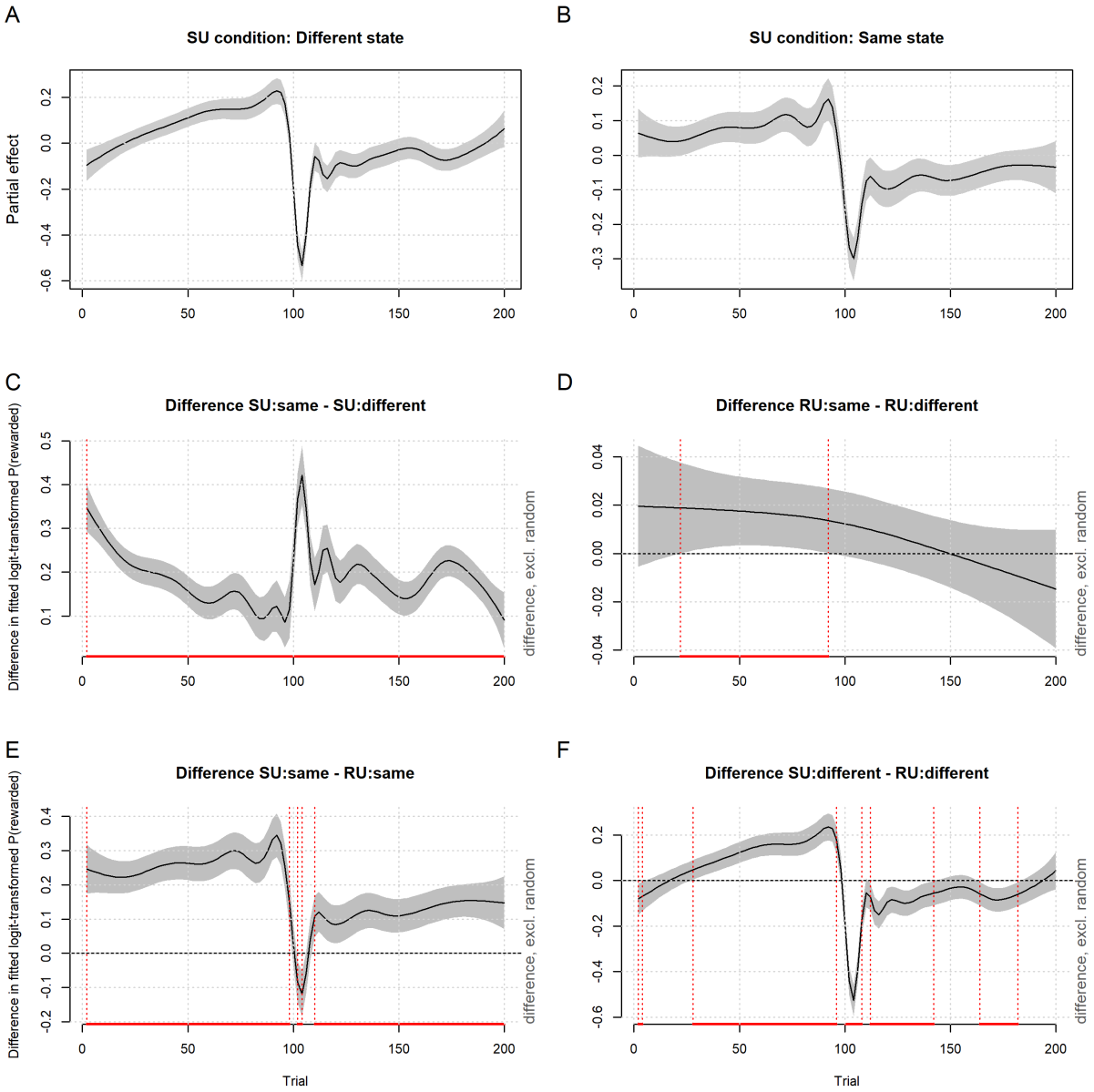

Figure S.11: Partial and interaction effects of the generalised additive mixed effects model of simulated reward acquisition in Experiment 2 using the simple RL model (corresponding to WTA-RL in the state uncertainty condition and to SARSA in the reward uncertainty condition). A-B: Nonlinear regression lines for the control and test conditions, respectively. The grey shadings along the curves represent pointwise 95% confidence intervals around the curve estimates. C: Difference in fitted logit-transformed proportion of rewarded responses between the control condition and test condition. The red segments indicate where the difference is significant.

A. Parametric coefficients	**Human**	**Simple RL**	**Bayesian RL**	**Sampling**
Intercept	.035	$< .001$	$< .001$	$< .001$
Condition & State match = RU.diff	.791	.778	$< .001$	$< .001$
Condition & State match = SU.same	$< .001$	$< .001$	$< .001$	$< .001$
Condition & State match = RU.same	.930	.645	$< .001$	$< .001$
Reward acquisition at t-1	$< .001$	$< .001$	$< .001$	$< .001$
B. Smooth terms				
s(Trial): Condition & State match = SU.diff	$< .001$	$< .001$	$< .001$	$< .001$
s(Trial): Condition & State match = RU.diff	.274	.104	.100	$< .001$
s(Trial): Condition & State match = SU.same	.576	$< .001$	$< .001$	$< .001$
s(Trial): Condition & State match = RU.same	$.015^*$	.991	.992	$< .001$
fs(Trial,Participant)	$< .001$	$< .001$	$< .001$	$< .001$

Table S.2: Significance of the effects in the generalised additive mixed models of the observed and simulated reward acquisitions based on the three fitted computational models.

S.7. Computational modelling results for Experiment 1

Table S.3 summarises the model fitting results for the first experiment. Across both conditions and all three models, the level of uncertainty ρ was estimated to be higher than the targeted level of uncertainty of 0.65, in line with our observation that some participants performed better than expected (see Section 2.4.2). The mean BIC scores suggest that the RL models fit the data better than the sampling model, and that the difference between the two RL models was negligible.

Table S.3: The mean ± S.D. of best-fitting parameter values and BIC scores by model and condition in Experiment 1. For the bound memory parameter b of the BI-SAW model, the median and the median absolute deviation were used instead of the mean and standard deviation since b can only take discrete values.

	Control condition			Test condition		
Parameter	PWRL	WTA-RL	BI-SAW	PWRL	WTA-RL	BI-SAW
α	0.35 (0.26)	0.32 (0.33)		0.32 (0.27)	0.18 (0.29)	
T	1.05 (0.93)	1.09 (1.10)	1.54 (1.23)	1.67 (1.56)	1.12 (1.63)	2.00 (1.59)
rho	0.74 (0.20)	0.77 (0.18)	0.80 (0.19)	0.86 (0.15)	0.80 (0.13)	0.86 (0.14)
b			4.5 (1.5)			7 (2)
w			0.56 (0.29)			0.52 (0.37)
pr			0.55 (0.28)			0.38 (0.41)
BIC	257.0 (20.8)	255.8 (20.5)	264.4 (19.0)	257.2 (22.6)	259.3 (20.7)	268.4 (19.6)

The comparison between the observed and fitted learning curves in Figure S.12 shows that the RL models qualitatively captured the behavioural data well in both conditions, and that they produced hardly discernible fitted learning curves in the control condition. The sampling model clearly produced the worst fit to the data in both conditions.

The comparison of the distribution of the BIC scores (Figure S.13) confirmed our previous observations: the RL models offered a substantially better fit in both conditions, and their fit was very close in the control condition. In the test condition, the PWRL was slightly better as it had a slightly lower median BIC score ($Mdn = 257$ vs $Mdn = 262$) and best fit more participants (5 versus 2).

Taken together, These results suggest that the RL models fit participants' data better but the data was inconclusive to infer which of the two RL models offered the best fit.

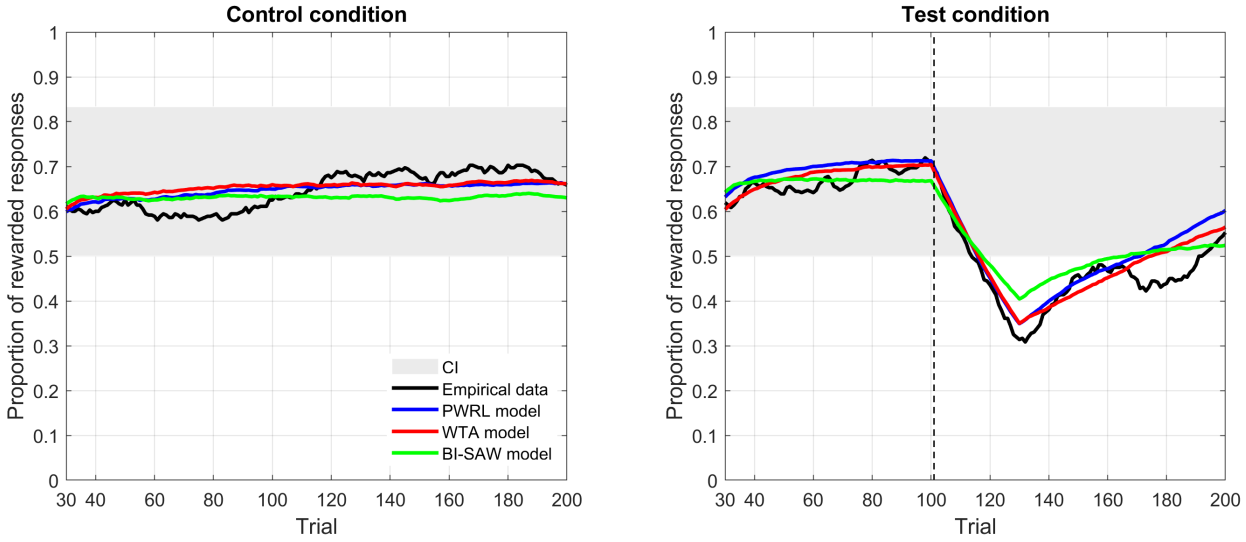

Figure S.12: Comparison of the reward-based learning curves of participants and fitted models in both the control (left) and test (right) conditions. To draw model learning curves, we first constructed a learning curve for each subject by averaging over 100 simulations with her fitted parameters, then averaged across participants to get the overall learning curve. The light grey region represents an approximate 95% confidence interval of the moving average of a Bernoulli variable with probability of success equal to 0.65. The vertical dashed line indicates the trial where the switch occurs in the test condition.

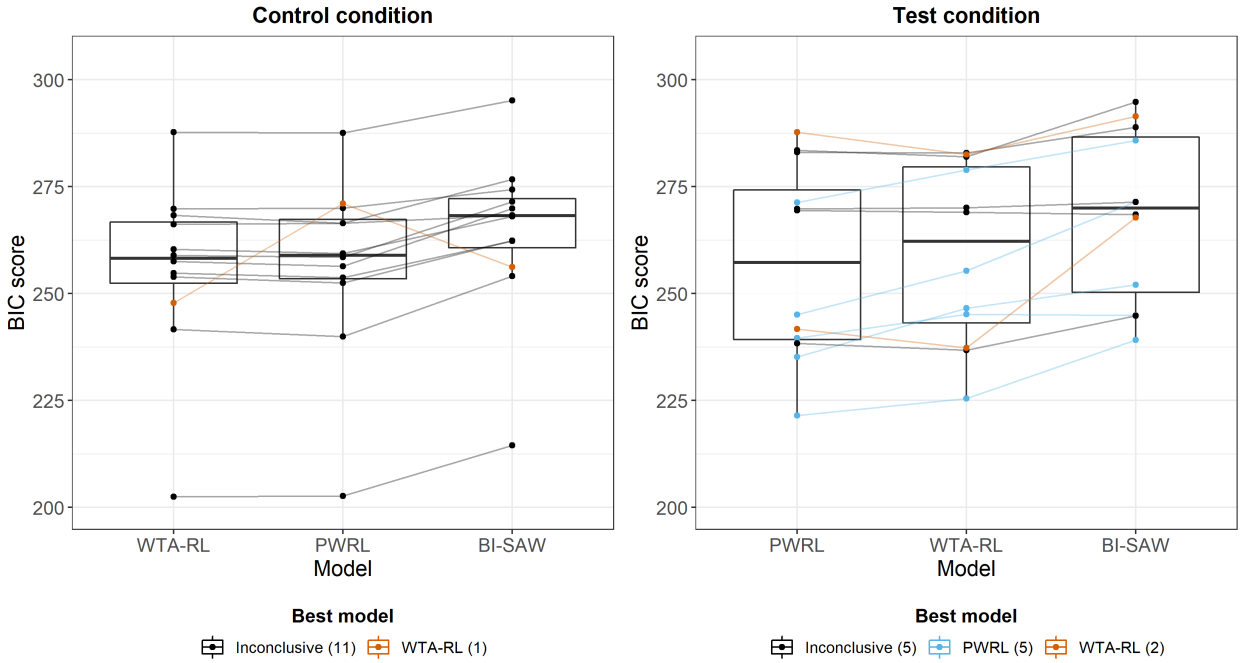

Figure S.13: Distribution of BIC scores by model for both the control (left) and test (right) conditions, with lower BIC scores indicating a better overall fit.

S.8. Individual learning curves with model fits

This section presents the learning curves of participants in each experiment (Experiments 1 and 2) and each condition (control/test in Experiment 1 and state/reward in Experiment 2) along with their fitted curves.

S.8.1. Experiment 1: Control condition

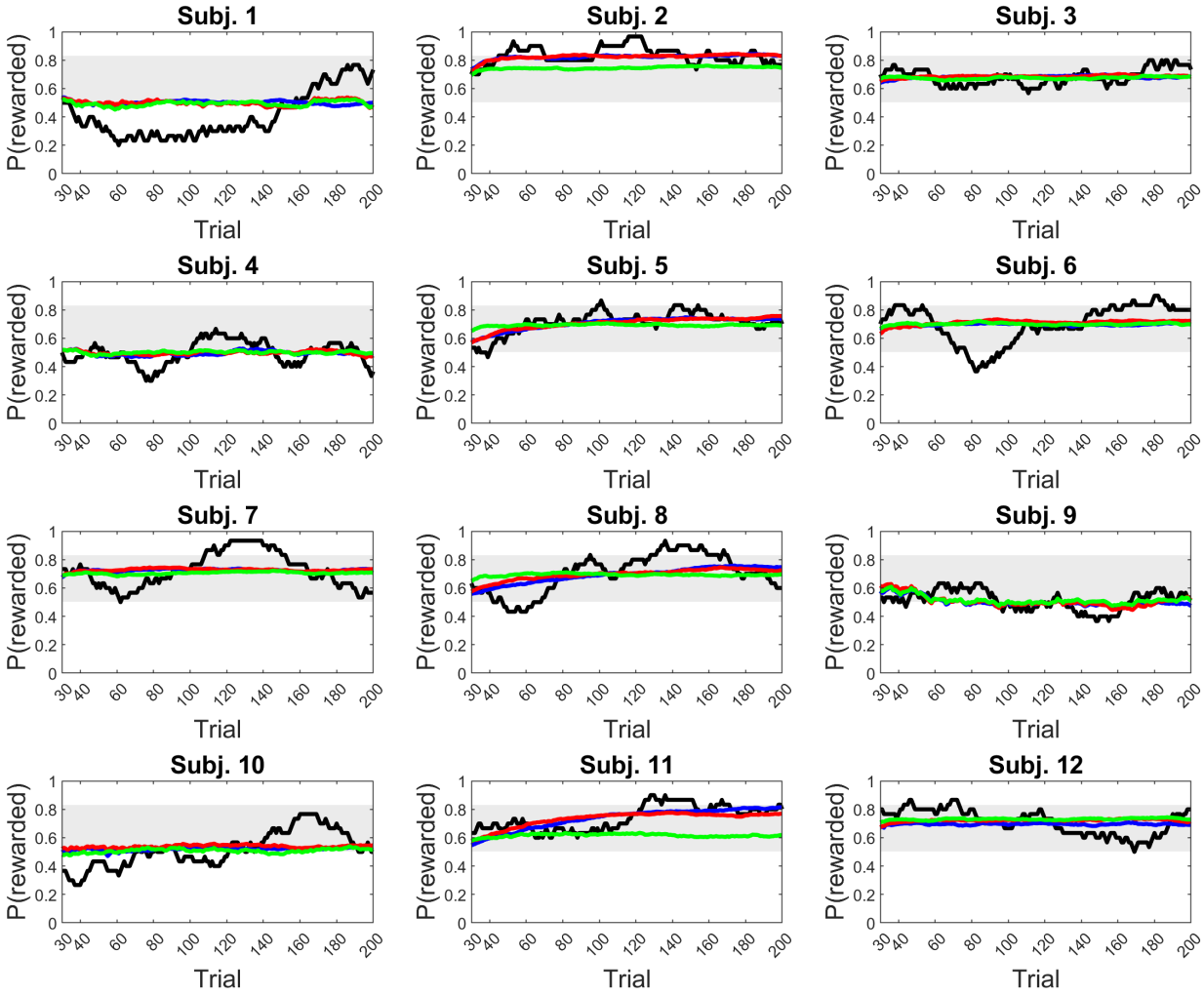

Figure S.14: Comparison of individual learning curves of participants and fitted models in the control condition. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.65.

S.8.2. Experiment 1: Test condition

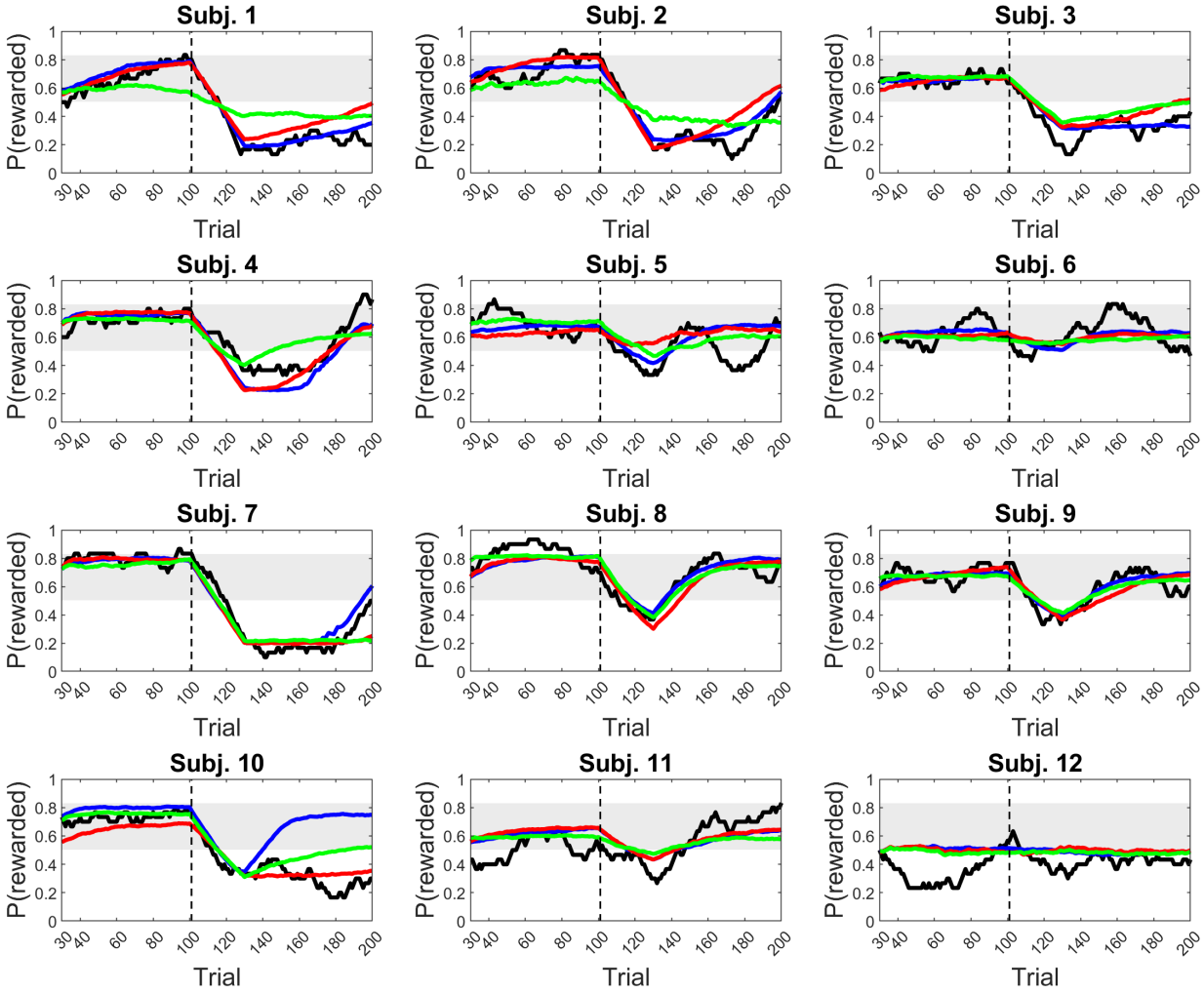

Figure S.15: Comparison of individual learning curves of participants and fitted models in the test condition. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.65.

S.8.3. Experiment 2: State uncertainty condition

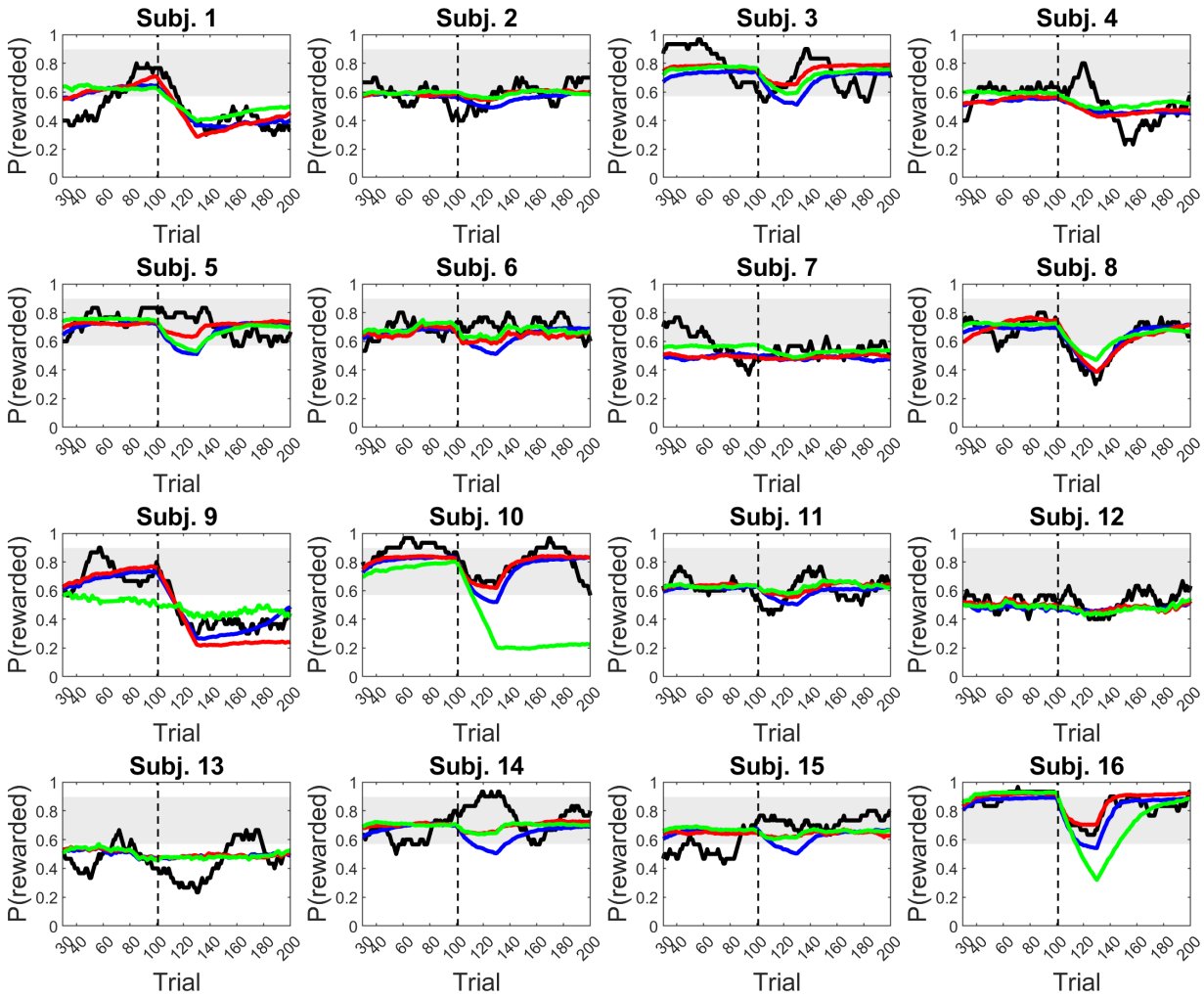

Figure S.16: Comparison of individual learning curves of participants and fitted models in the state uncertainty condition (Participants 1–16). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.

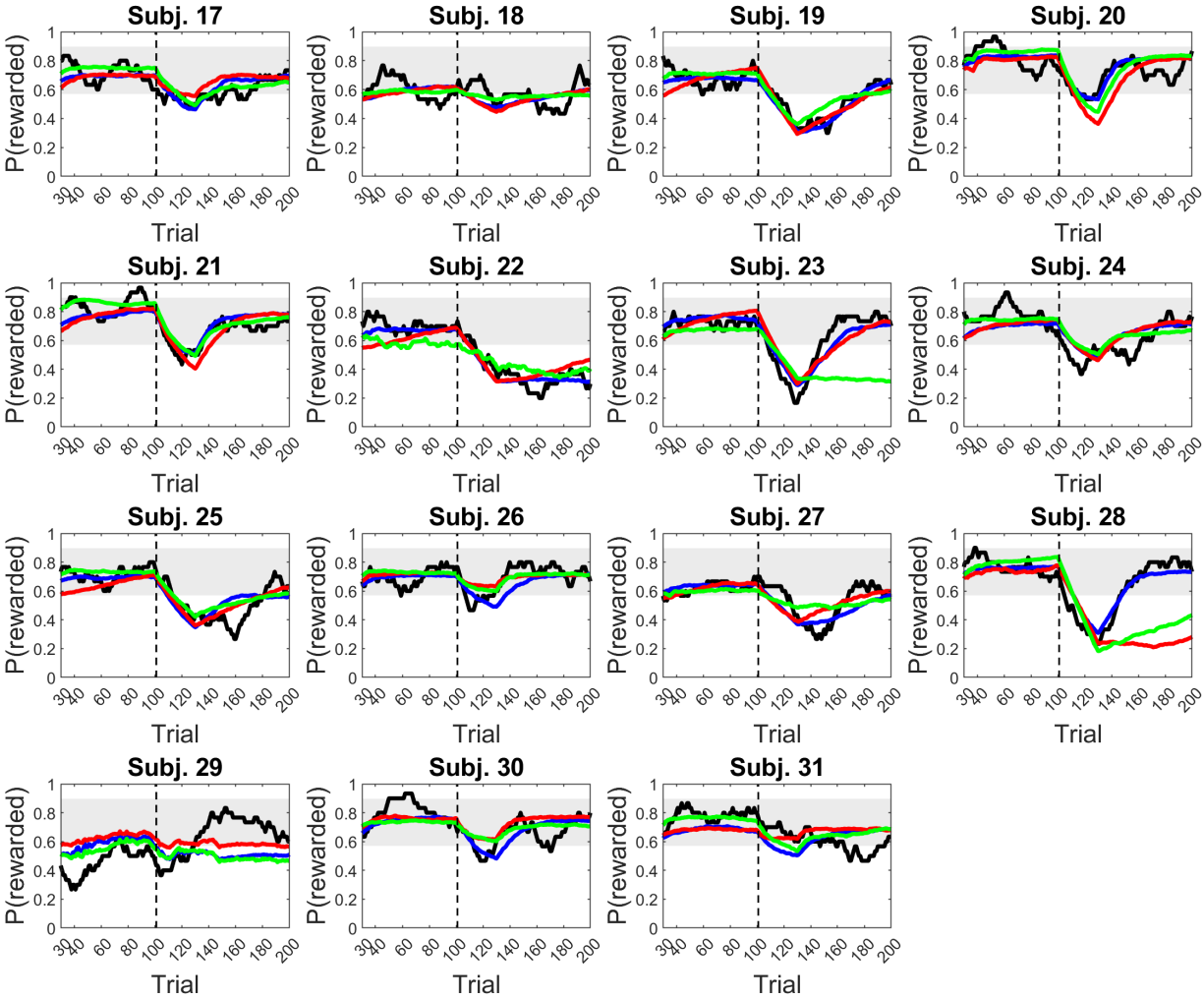

Figure S.17: Comparison of individual learning curves of participants and fitted models in the state uncertainty condition (Participants 17–31). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.

S.8.4. Experiment 2: Reward uncertainty condition

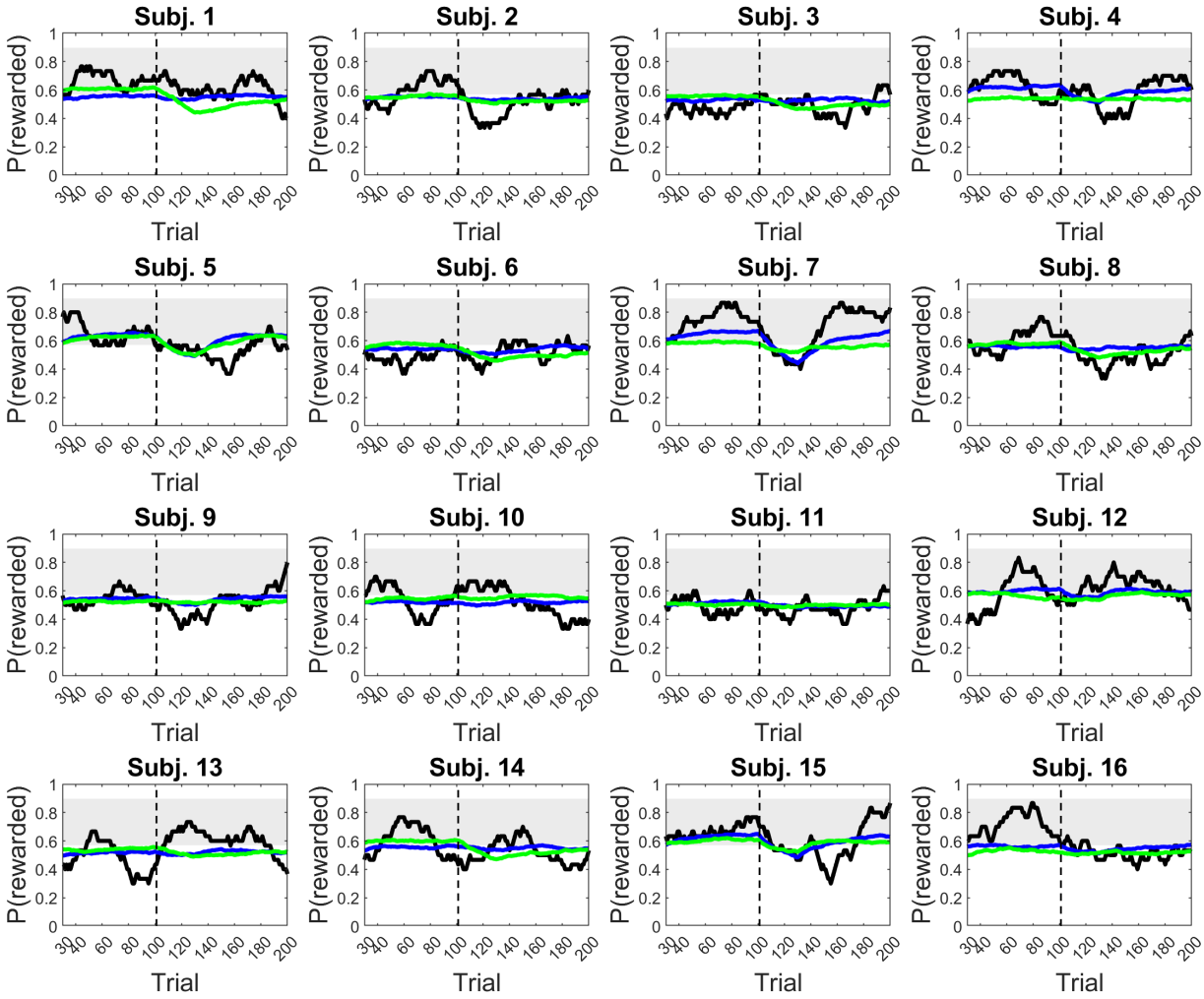

Figure S.18: Comparison of individual learning curves of participants and fitted models in the reward uncertainty condition (Participants 1–16). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.

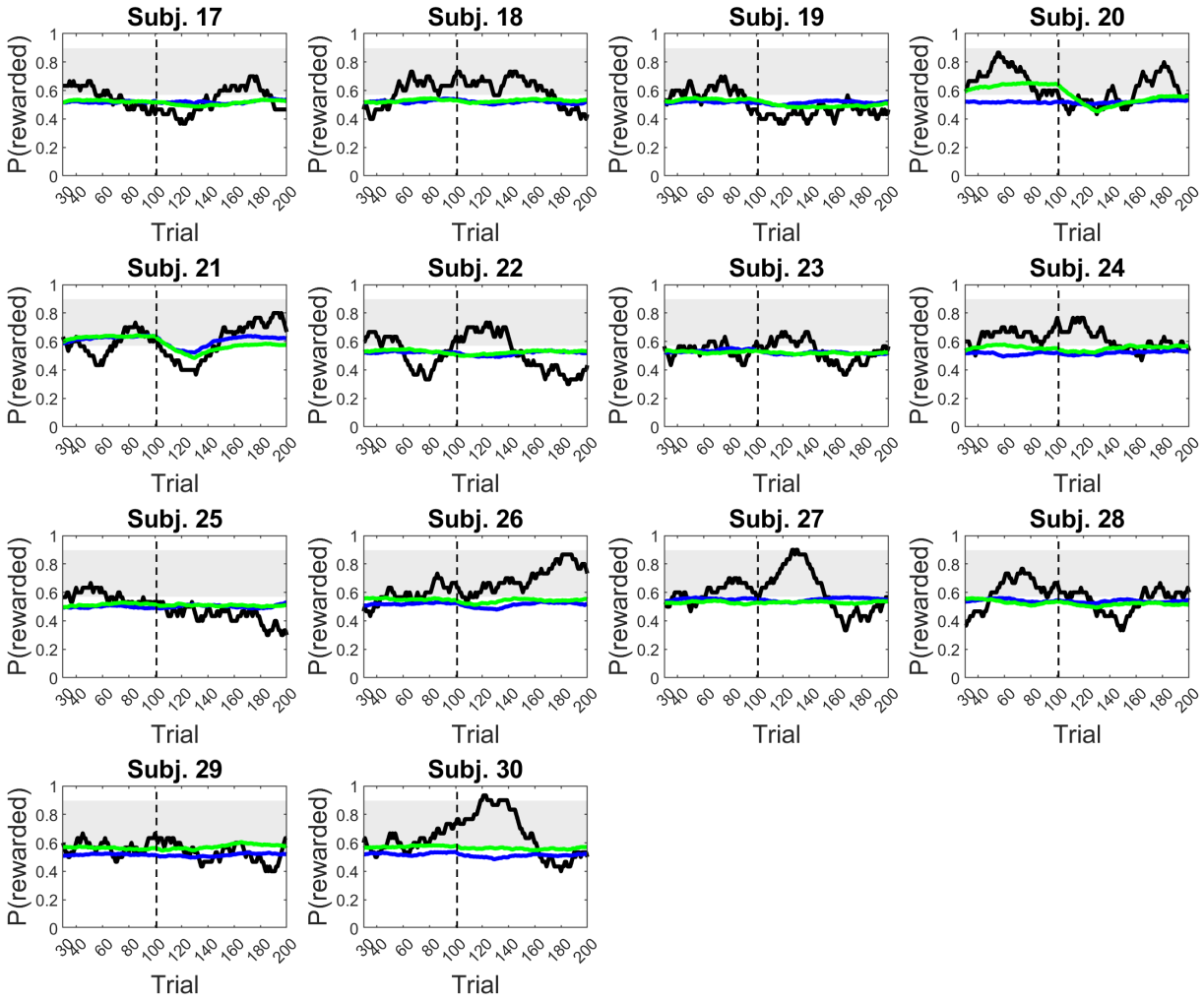

Figure S.19: Comparison of individual learning curves of participants and fitted models in the reward uncertainty condition (Participants 17–30). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.

S.9. Model fit results with the level of uncertainty ρ as a fixed parameter

S.9.1. Best-fitting parameters: Experiment 1

Table S.4: The mean ± S.D. of best-fitting parameter values and BIC scores by model and condition in Experiment 1 with ρ fixed. For the bound memory parameter b of the BI-SAW model, the median and the median absolute deviation were used instead of the mean and standard deviation since b can only take discrete values.

	Control condition			Test condition		
Parameter	PWRL	WTA-RL	BISAW	PWRL	WTA-RL	BISAW
α	0.28 (0.32)	0.22 (0.37)		0.10 (0.09)	0.18 (0.32)	
T	6.28 (19.00)	0.31 (0.58)	0.91 (1.34)	0.80 (1.05)	0.35 (0.67)	1.27 (1.24)
b			4.5 (1.5)			6.5 (2.5)
w			0.63 (0.40)			0.42 (0.37)
pr			0.71 (0.25)			0.57 (0.37)
$-logML$	125.9 (9.1)	125.0 (7.2)	125.9 (8.2)	128.8 (9.8)	126.4 (8.1)	130.3 (7.2)
BIC	262.4 (18.17)	260.6 (14.4)	272.9 (16.4)	268.1 (19.5)	263.4 (16.2)	281.8 (14.5)

S.9.2. Best-fitting parameters: Experiment 2

Table S.5: The mean ± S.D. of best-fitting parameter values and BIC scores by model and condition in Experiment 2 with ρ fixed. For the bound memory parameter b of the BI-SAW model, the median and the median absolute deviation were used instead of the mean and standard deviation since b can only take discrete values.

	State uncertainty			Reward uncertainty	
Parameter	PWRL	WTA-RL	BISAW	SARSA	BISAW
α	0.19 (0.20)	0.26 (0.32)		0.37 (0.17)	
T	13.78 (19.71)	2.72 (5.02)	1.09 (3.53)	12.17 (12.10)	10.51 (10.28)
b			9 (1)		3 (2)
w			0.50 (0.31)		0.27 (0.31)
pr			0.27 (0.24)		0.69 (0.31)
$-logML$	134.9 (4.4)	127.2 (8.2)	125.7 (8.2)	117.9 (23.6)	114.8 (18.5)
BIC	280.4 (8.8)	265.1 (16.4)	272.7 (16.4)	246.3 (47.3)	250.8 (37.0)

S.9.3. Aggregated learning curves with model fits: Experiment 1

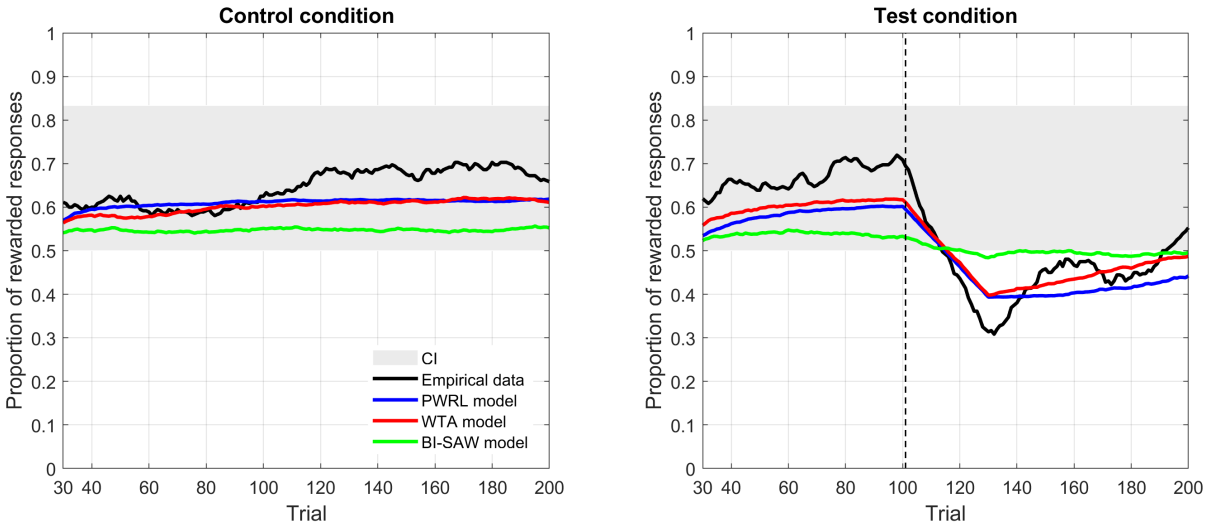

Figure S.20: Comparison of the reward-based learning curves of participants and fitted models in both the control (left) and test (right) conditions. To draw model learning curves, we first constructed a learning curve for each subject by averaging over 100 simulations with her fitted parameters, then averaged across participants to get the overall learning curve. The light grey region represents an approximate 95% confidence interval of the moving average of a Bernoulli variable with probability of success equal to 0.65. The vertical dashed line indicates the trial where the switch occurs in the test condition.

S.9.4. Aggregated learning curves with model fits: Experiment 2 (state uncertainty condition)

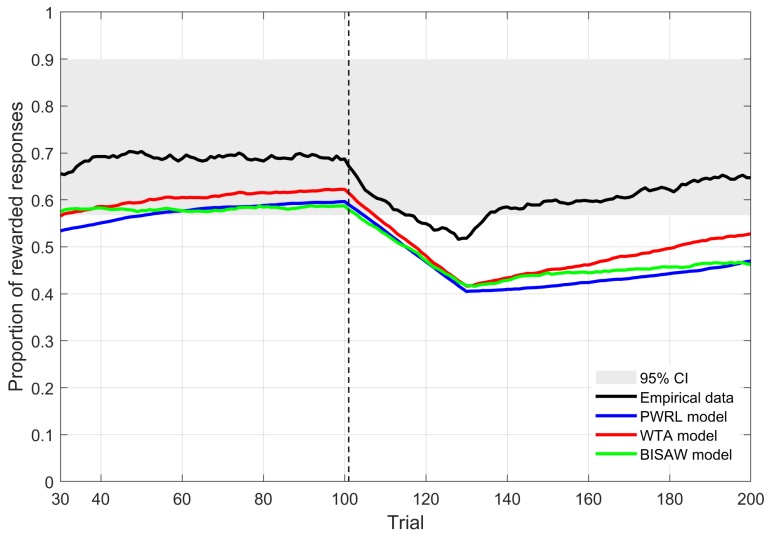

Figure S.21: Comparison of the reward-based learning curves of participants and fitted models in the state uncertainty. To draw model learning curves, we first constructed a learning curve for each subject by averaging over 100 simulations with her fitted parameters, then averaged across participants to get the overall learning curve. The light grey region represents an approximate 95% confidence interval of the moving average of a Bernoulli variable with probability of success equal to 0.70. The vertical dashed line indicates the trial where the switch occurs.

S.9.5. Comparison of model fit quality: Experiment 1

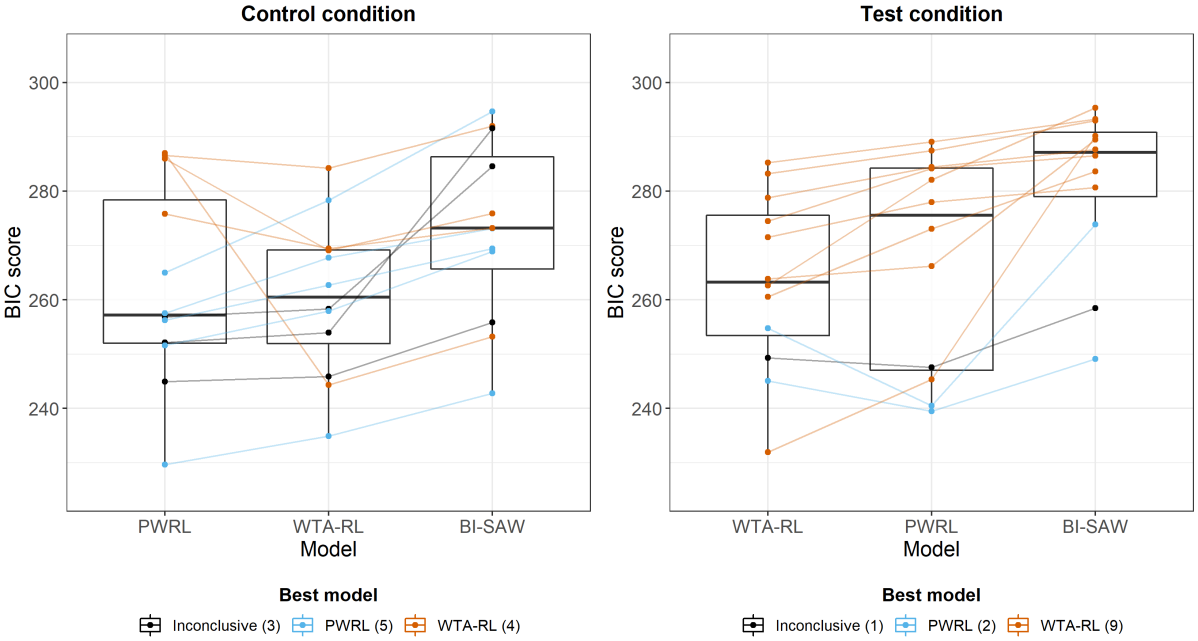

Figure S.22: Distribution of BIC scores by model for both the control (left) and test (right) conditions, with lower BIC scores indicating a better overall fit.

S.9.6. Comparison of model fit quality: Experiment 2 (state uncertainty condition)

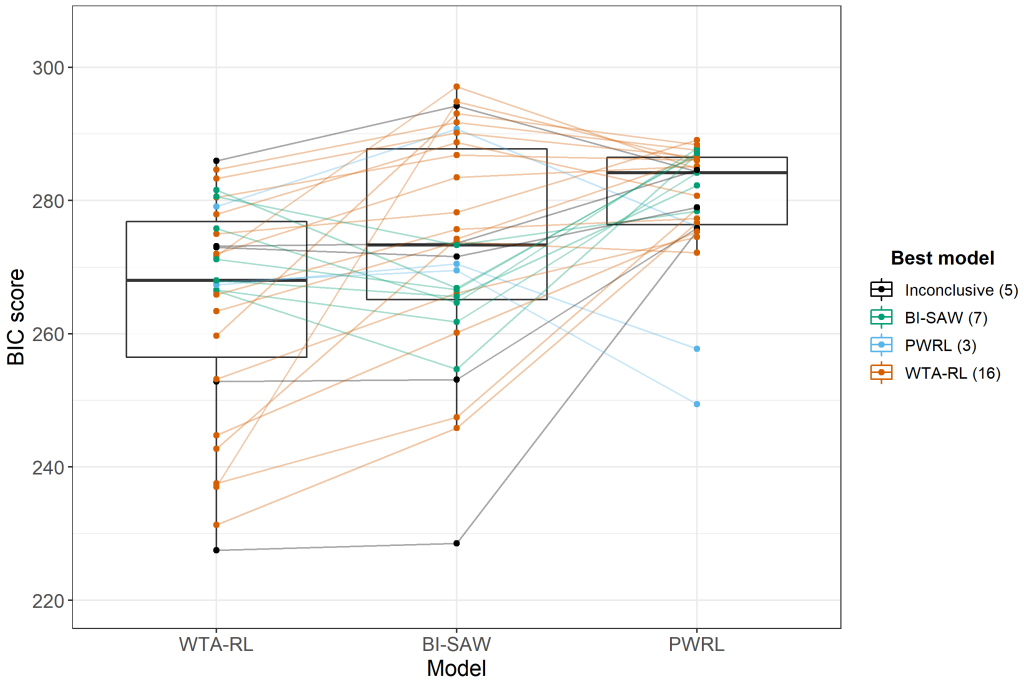

Figure S.23: Distribution of BIC scores by model for the state uncertainty condition, with lower BIC scores indicating a better overall fit.

S.9.7. Individual learning curves with model fits: Experiment 1 (control condition)

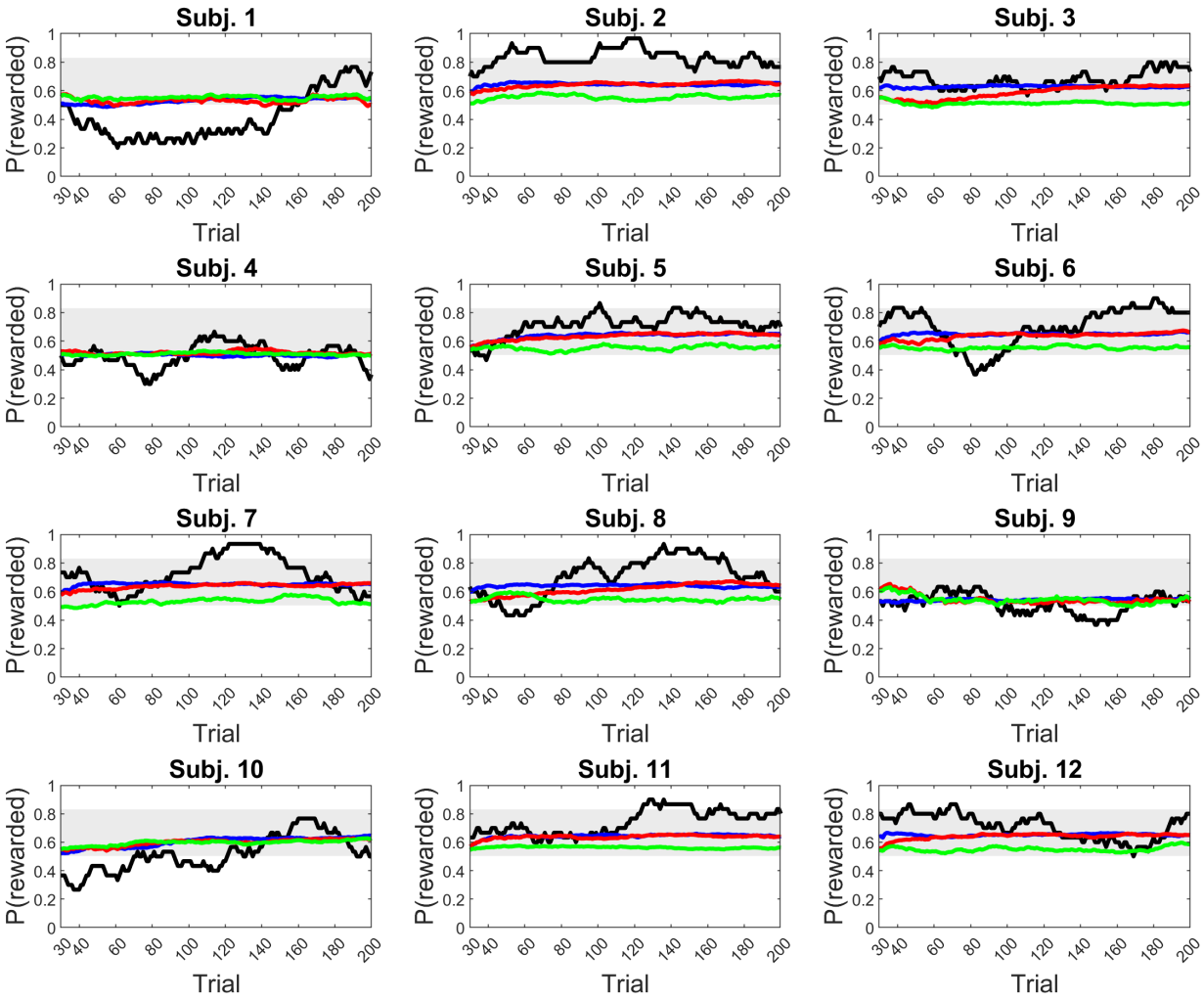

Figure S.24: Comparison of individual learning curves of participants and fitted models in the control condition. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.65.

S.9.8. Individual learning curves with model fits: Experiment 1 (test condition)

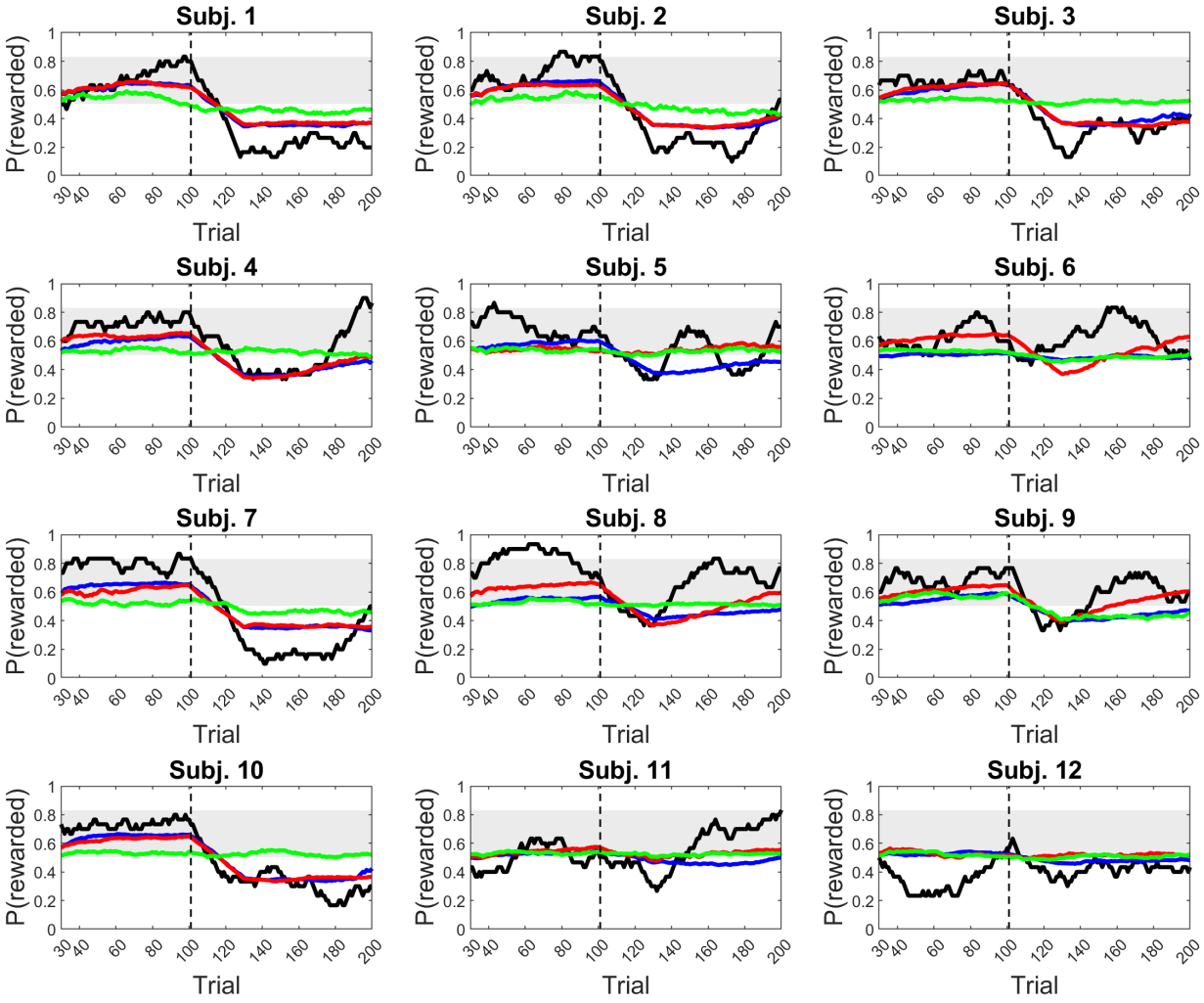

Figure S.25: Comparison of individual learning curves of participants and fitted models in the test condition. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.65.

S.9.9. Individual learning curves with model fits: Experiment 2 (state uncertainty condition)

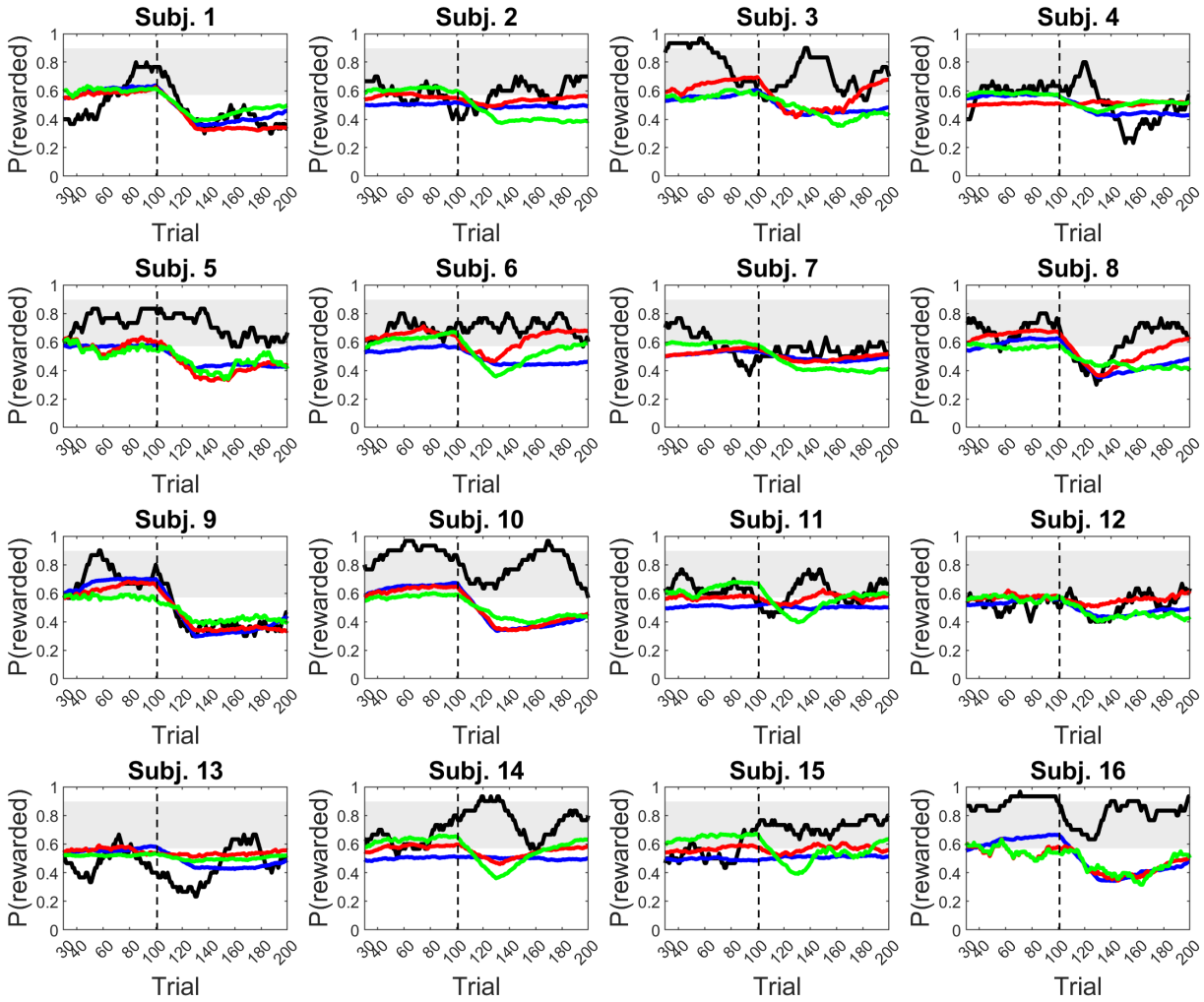

Figure S.26: Comparison of individual learning curves of participants and fitted models in the state uncertainty condition (Participants 1–16). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.

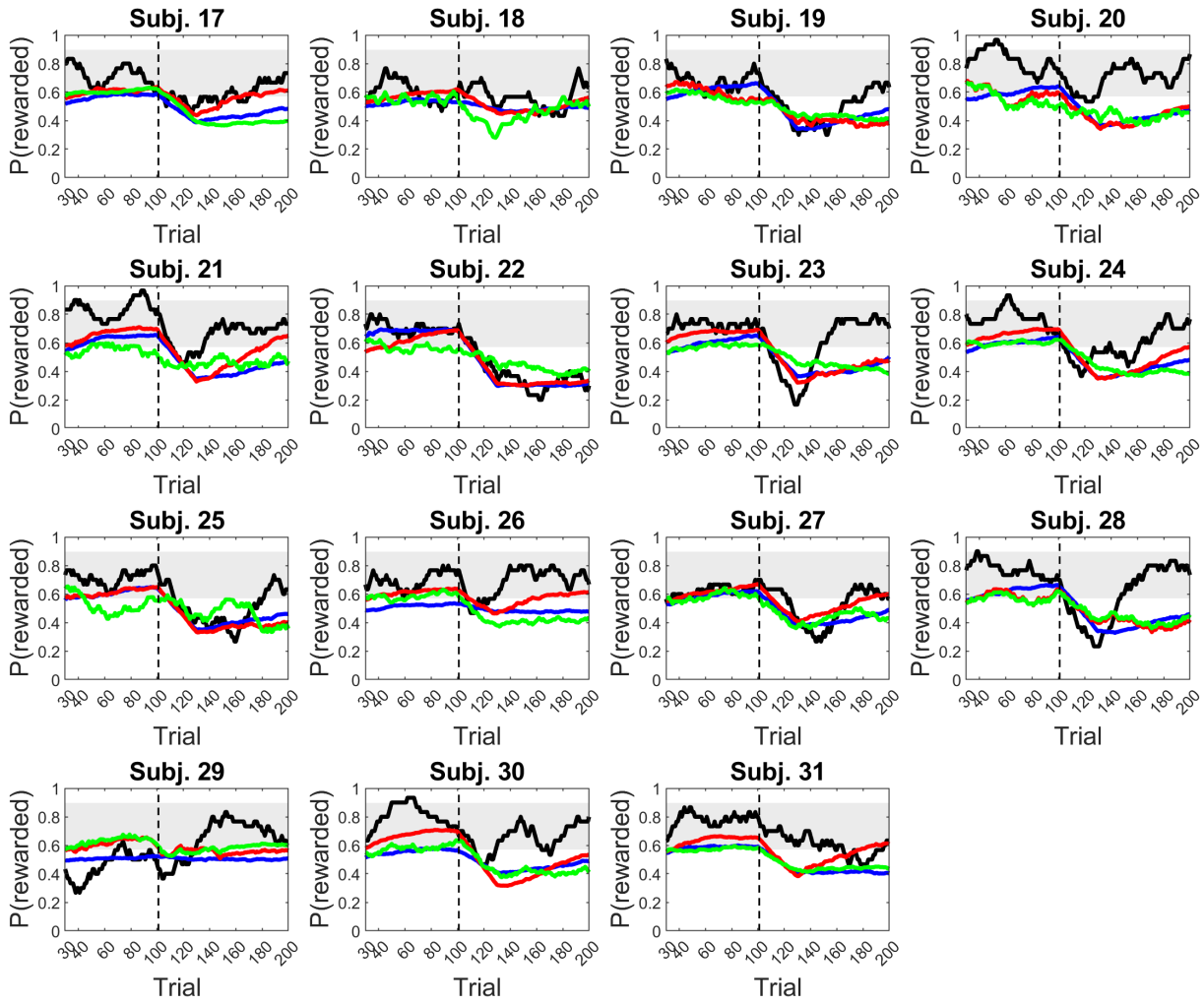

Figure S.27: Comparison of individual learning curves of participants and fitted models in the state uncertainty condition (Participants 17–31). The curves of the PWRL model are in blue, WTA-RL in red and BI-SAW in green. We constructed each participant's fitted learning curve by averaging over 100 simulations with her fitted parameters. The light grey region represents an approximate 95% confidence interval of the moving average with a probability of accuracy of 0.7.