Uncertainty in learning, choice and visual fixation

Hrvoje Stojić[a,1], **Jacob L. Orquin**[b,e], **Peter Dayan**[c], **Raymond Dolan**[a,2], **and Maarten Speekenbrink**[d,2]

[a]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square, London, WC1B 5EH, United Kingdom; [b]Department of Management, Aarhus University, Fuglesangs Alle 4, Aarhus, 8210, Denmark; [c]Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8-14, Tübingen, 72076, Germany; [d]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, United Kingdom; [e]Centre for Research in Marketing and Consumer Psychology, Reykjavik University, Menntavegur 1, 101 Reykjavik, Iceland

This manuscript was compiled on January 8, 2020

Uncertainty plays a critical role in reinforcement learning and decision making. However, exactly how it influences behaviour remains unclear. Multi-armed bandit tasks offer an ideal test-bed, since computational tools such as approximate Kalman filters can closely characterize the interplay between trial-by-trial values, uncertainty, learning, and choice. To gain additional insight into learning and choice processes we obtained data from subjects' overt allocation of gaze. The estimated value and estimation uncertainty of options influenced what subjects looked at before choosing; these same quantities also influenced choice, as additionally did fixation itself. A momentary measure of uncertainty in the form of absolute prediction errors determined how long participants looked at the obtained outcomes. These findings affirm the importance of uncertainty in multiple facets of behaviour, and help delineate its effects on decision making.

Reinforcement learning | Decision making | Uncertainty | Visual fixation | Exploration-exploitation

We often need to decide between alternative courses of action about whose outcome we are uncertain. Common examples include choosing a dish in a restaurant, a holiday trip, or financial investment. Uncertainty, which derives from initial ignorance and sometimes ongoing change, has two characteristic statistical and computational facets. One is straightforward: if we try an option, then the amount of learning, i.e., the extent to which we should update our beliefs, depends on our current uncertainty relative to the noise in the observation (1). The greater our uncertainty, the greater the impact an observation inconsistent with our current beliefs should have on our subsequent beliefs. There is good evidence that humans and other animals adapt their rate of learning to various factors in the environment which increase, or reduce, uncertainty (2–6).

The second facet concerns choice. Here, it is the options that we are uncertain about and that we need to learn about through sampling. This is more complicated, as our ignorance about their beneficial or malign consequences implies that we need to take a sampling risk. This is the notorious exploration/exploitation dilemma. Although there are elegant computational solutions for important special cases (Gittins indices; 7), a general solution is intractable. There is evidence that when choosing options, people explore in a directed manner, by integrating values with uncertainty about these values (8–12), particularly when these are carefully dissociated (9, 10). However, there is also evidence for a simpler form of random, undirected, exploration, which is sensitive to value but not to its uncertainty (5, 13). Integration of value and the uncertainty in its estimation is sensible. Estimation uncertainty serves as a proxy for how informative a choice is, or what the potential for improvement in value is (14, 15). The distinction from irreducible uncertainty is important. Irreducible uncertainty stems from the inherent stochastic nature of the environment that generates rewards and can not be reduced through learning.

Most studies only admit indirect inferences about the processes of learning and decision-making, exploiting the trajectory of choices alone. However, when options are presented visually and are spatially distinct, we have an opportunity to gain a window onto these processes by examining what people choose to look at, that is, their visual fixations (16–25). In typical tasks, including the one we employ in our experiment, we can expect two sorts of revealing fixation behaviour; namely, the relative time spent on each option when deciding (which bears on choice); and the absolute fixation time when receiving feedback about the consequences of choices (which bears on learning).

Fixation time might be correlated not only with subjects' internal states relevant to learning and choice, but might actually affect those states directly (18, 21). This also allows factors other than value and estimation uncertainty, including stimulus salience, momentary lapses of attention, or unrelated cognitive processes to influence fixation (26–28), and exert statistically untoward effects on behaviour.

In the case of choice, a prominent view is that the process leading up to a decision involves accumulating information about the options until one is judged to be sufficiently good or sufficiently better than the alternatives (29, 30). Under this framework, looking at an option facilitates accumulating information specifically about that option (18, 21). This would provide a mechanism through which relative fixation time before making a choice can have a direct influence on the decision itself. In this case, for choices to be approximately optimal (7, 8, 10, 11), the relative fixation time before a choice

Significance Statement

Humans cannot help but turn their gaze to objects that catch their attention. Our knowledge of the factors that govern this seizure, or of its effects in the context of learned decision-making, is currently rather incomplete. We therefore monitored the gaze of human subjects as they learned to choose between multiple options whose value was initially unknown. We found evidence that attention was influenced by uncertainty; and that the use of, and reduction in, uncertainty were in turn influenced by attention. Our findings provide evidence for approximately optimal models of learning and choice and uncover an intricate interplay between learning, choice and attentional processes.

The authors declare no conflict of interest.

[2]R.D. and M.S. contributed equally to this work.

[1]To whom correspondence should be addressed. E-mail: h.stojic@ucl.ac.uk

would have to reflect the learning history, with respect to both the value and estimation uncertainty. Our focus on directed exploration and estimation uncertainty distinguishes the present study from previous ones on reinforcement learning and attention, which focused on effects of value (22) and irreducible uncertainty (24), or did not in any case involve exploration (25).

In the case of learning, absolute fixation time might have a direct influence on the magnitude of belief change in response to a prediction error, which amounts to the learning rate. For instance, visual fixations facilitate working memory and memory retrieval operations (31–35). Based on this evidence, fixation time might influence how well a newly observed outcome is integrated with an old value retrieved from memory. Thus, to follow the precepts of Bayesian statistical learning, fixation should be related to an option's estimation uncertainty (3), allowing the latter to be observable from the former. While this prediction was made almost two decades ago, empirical evidence has been lacking (16).

To examine the role of estimation uncertainty and complex interactions between visual fixation, learning, and choice, we administered a multi-armed bandit task in which we also tracked subjects' gaze as they chose repeatedly between six, initially unknown, options. We varied the mean and variance of options' outcomes to motivate exploration and to ensure ample variability in value and estimation uncertainty. When ignoring fixation behaviour, we found that both value and estimation uncertainty play a role in learning and choice. As predicted, we found that over the course of decision making, estimation uncertainty and value jointly influenced relative fixation times. During feedback, when subjects could update their beliefs, uncertainty, in the form of the unsigned reward prediction error, guided the total fixation time on the chosen option. Even though relative fixation time during choice carried information about value and estimation uncertainty, fixation exerted a much stronger independent influence on choices than was warranted by that information. This indicates that an important fixation-specific component influenced choice. Finally, we show that a model including value, estimation uncertainty, and relative fixation time before choice, best explained actual choices. This suggests that the influence of the first two of these quantities is not completely mediated by their effect on the third, and that capturing an internal valuation process is therefore still important.

Results

Participants completed two games. In each game they repeatedly chose between six options, for a total of 60 trials (Fig. 1A, *Materials and Methods* and *SI Appendix, Methods*). Each game was a multi-armed bandit task in which rewards for each option were drawn from different Gaussian distributions (Fig. 1C). Participants were instructed to maximize the cumulative sum of rewards in each game. To attain this goal they needed to explore the options in the choice set in order to learn which option had the highest average reward, and subsequently exploit this knowledge.

To facilitate detecting whether estimation uncertainty guided participants' exploration, the variances of the reward distributions differed between each of the options. The rationale behind this manipulation is that choices that are guided by value alone would be less directly affected by such differences in variances. In a *Decreasing variances* game, variance decreased as the mean reward of the option decreased, so that, for instance, the option with the highest mean had the highest variance (Fig. 1C, left). In a *V-shaped variances* game, the variance was largest for the options with the highest and smallest means, and smaller for the middle options (Fig. 1C, right). Different games allow for better generalization of results and can serve as a further check for directed exploration, as again choices guided by value alone would be less sensitive to such differences.

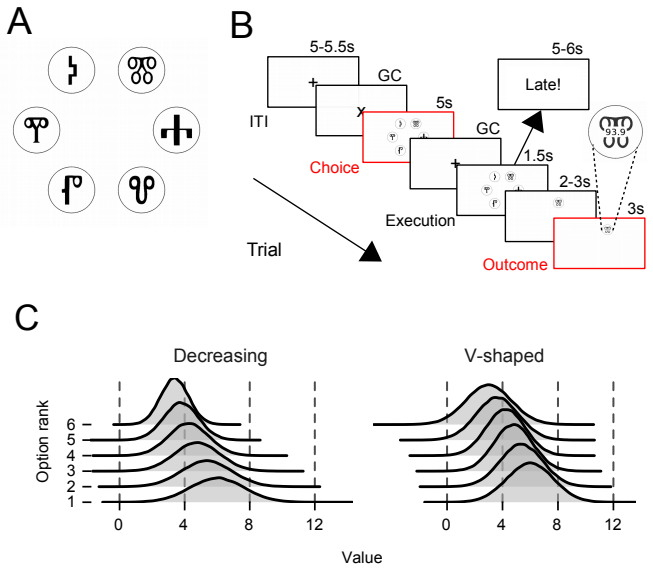

Fig. 1. Illustration of the six-armed bandit task. (A) Participants chose between six options on each of 60 trials. Each option was represented by a letter from the Gljagoljica alphabet. Options were displayed in a circle around the centre of the screen, always at the same location. (B) Time course of a single trial. Each box denotes a stage in a trial, with duration displayed above the boxes. For visual fixation analyses the main stages of interest were *Choice stage*, where participants considered which option to choose, and *Outcome stage*, where they observed a choice outcome (the inset displays reward outcome overlaid over the option). Two stages were gaze contingent (GC), where participants trigger an onset by fixating on a fixation cross. (C) To facilitate detecting whether estimation uncertainty guided participants' exploration, the variances of the reward distributions differed between each of the options. In *Decreasing variances* game distributions get narrower (more certain, easier to learn) going from the best (rank 1) to the worst (rank 6) option, while for *V-shaped variances* game they are the narrowest for the middle ranking and broader (more uncertain, taking more trials to learn) for the better and worse ranking options respectively.

Options' expected rewards were constant throughout the bandit task. In such a task any reasonable reinforcement learning agent that maximizes cumulative rewards would gradually allocate more and more choices to high value options as its estimates of options' rewards improve with experience. Indeed, choices improved from the first to the last block of 15 trials (Fig. 2A), as indicated by a clear negative block effect (mixed effects regression estimates: intercept = 2.50, 95% credible interval (CI) $[2.25, 2.75]$; block = −0.29, 95% CI $[-0.37, -0.21]$; game = 0.06, 95% CI $[-0.04, 0.16]$; block×game = 0, 95% CI $[-0.07, 0.08]$; see "Mixed effect regressions" in *Materials and Methods*). There was no strong difference in choice performance between the games, indicating that low ranking options did not attract more choices in the V-shaped game. While this could be due to choices not being guided by estimation uncertainty, an alternative explanation is that participants learned to ignore the low ranking options very quickly. This

would resulting in weak difference between the games since it was mainly these that distinguished the distributions between games. In most cases, choice performance did not reach ceiling by the last block of 15 trials (mean of 2.08, $SE = 0.10$), suggesting that the games were not trivial and participants were still exploring by the end of the task.

In the following section, we outline a computational model built to determine the extent to which estimation uncertainty influenced choice. We then use this model to examine the multi-way relationships between the visual fixation during the period preceding each choice, the values and uncertainties of all the options estimated by the model, and the actual decision made by participants. We repeat this analysis for the relationships among fixation statistics at the time of reward feedback, the prediction error and estimation uncertainty that the model estimated participants entertain about the chosen option, and the ensuing learning.

Estimation uncertainty and choice. To identify learning and choice processes underlying participants' behaviour, we fitted computational models to their decisions. These models consisted of a learning component, in which participants learn or estimate properties of each option, and a choice component where they rely on these estimates to decide between the options.

Along with four control models often used to capture learning and choice in these types of tasks (*SI Appendix, "Modelling learning and choices – control models" in Results*), we considered two more sophisticated learning models, each coupled with two forms of choice. The learning models were either a Kalman filter (8, 13, 36), or a "lazy" Kalman filter, both of which use a variant of the delta rule to update estimated values from a reward prediction error (*Materials and Methods*, Eq. 1 and 2). The Kalman filter is a Bayesian model that tracks the expected values of options, as well as the uncertainties in those expectations (i.e. estimation uncertainty). Moreover, it dynamically adjusts the learning rate according to its current estimation uncertainty and the relative noise in the observed rewards. At each point in time, the Kalman filter provides an estimate of the value of an option as a Normal distribution, whose mean reflects the expected value, and whose variance reflects estimation uncertainty (in the remainder of the text we will use the term uncertainty to refer to estimation uncertainty). These means and variances are the key quantities we subsequently use to examine the role of value and uncertainty in visual fixations. The lazy Kalman filter is similar to the regular Kalman filter but with one crucial difference: it uses a learning rate which is a fraction of that of the regular Kalman filter (hence its moniker). Both models take into account differences in variances of options' rewards in each game (i.e. irreducible uncertainty), leading to different learning rates for each option.

The choice component in the models consisted of either a softmax (SM; Eq. 3; 37) or an upper confidence bound (UCB; Eq. 4; 14) rule. The softmax choice rule only uses estimated value to determine choice. As such, exploration is not guided by uncertainty. By contrast, the UCB choice rule implements a form of directed exploration. It uses the uncertainty to approximate the information gained by choosing an option, and adds this as an "uncertainty bonus" to the estimated value (38), implying that exploration is driven by a form of expected information gain.

We used a Bayesian hierarchical approach to estimate the parameters of the models. This assumes the parameters at the individual participant level are drawn from common group-level distributions (39). Model evidence shows that models with the UCB choice rule fit the data better than models using the softmax choice rule that ignores uncertainty (Fig. 2B). The lazy Kalman filter model with a UCB choice rule described participants' choices best (KFL-UCB), with a posterior probability of approximately 0.99. Lazy versions of Kalman filter learning also outperformed the standard ones for the softmax choice rule. The Kalman filter models with the UCB choice rule convincingly outperformed all four control models (*SI Appendix, Results*). The probability of accurately predicting participants' choices with the KFL-UCB model increased steadily over the course of a game, reaching a mean of 0.46 ($SE = 0.08$) by trial 60 (Fig. 2C), well above the chance level ($1/6 = 0.17$) and above a simple non-learning model in which we estimate fixed probabilities of choosing each option (mean choice probability of 0.21). The overwhelming evidence in favour of the UCB choice rule shows that estimation uncertainty plays a clear role in choice. This shows that our model-based analysis is more sensitive than the model-free analysis predicated on the different variance patterns. The lack of a between-game effect in performance was likely due to participants quickly learning to ignore the low value options.

Since the only difference between the best fitting KFL-UCB model and its softmax counterpart (KFL-SM) is the β parameter that acts as a weight on uncertainty in the UCB choice rule, the strong evidence favouring the KFL-UCB model over the KFL-SM model indicates that the β parameter is reliably positive. Indeed, the posterior distribution of the β parameter of the KFL-UCB model has a mean of 0.37, and the 95% credible interval (CI) is $[0.16, 0.61]$ (Eq. 4; Fig. 2D). This "inflation of value" is a sizeable uncertainty bonus, given that the expected values of options ranged between 2.5 and 6 and their variances between 0.75 and 2.75. As a final check, we also fitted a variant of the KFL-UCB model where the β parameter is not constrained to be non-negative. The KFL-UCB model with the non-negative β parameter outperformed the unconstrained KFL-UCB model with a posterior probability of approximately 0.99 (see "KFL-UCB model with unconstrained β parameter" in *SI Appendix, Results*). This result further affirms that the β parameter is positive and that uncertainty guides choice together with value.

We can also examine the usefulness of the "laziness" parameter (η) that biases the learning rate in the KFL-UCB model. A value of $\eta = 1$ would make the lazy Kalman filter equivalent to the regular Kalman filter. The bias seems to be rather small, as evidenced by the group-level posterior mean (0.93, 95% CI $[0.80, 0.99]$; Eq. 1). However, the individual variability is substantial: for a sizeable number of games (and individuals) parameter values were much lower and closer to 0 (Fig. S5C). This suggests the laziness parameter captures significant variation in behaviour. Values of the remaining parameters are depicted in Fig. S5.

Interactions between choice and fixation process. We next sought to assess three-way interactions between fixation during the choice epoch, the choice itself, and the combination of value and uncertainty. We first report basic properties of fixation during the choice epoch. We then look at how value and uncertainty influence fixation. Finally, we ask whether

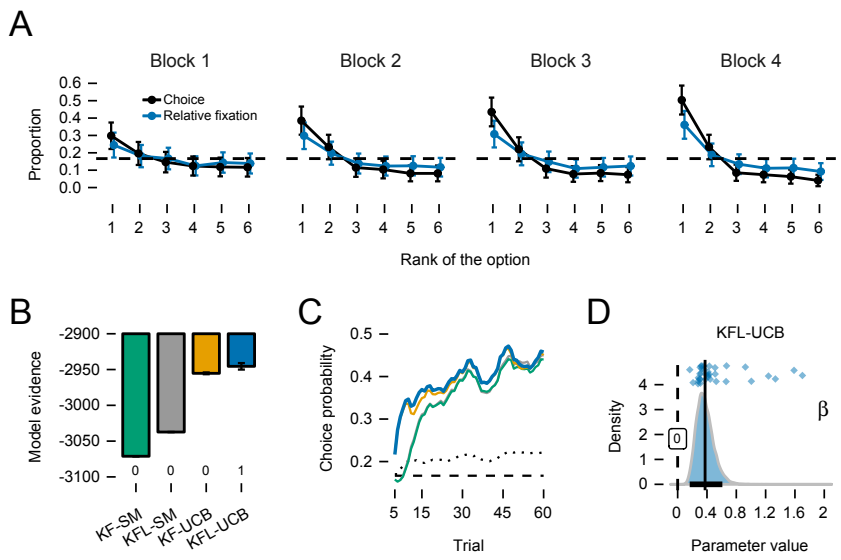

Fig. 2. (A) Proportions of choices allocated to options with the highest expected value (i.e. options with a rank equal to 1) increased from the first to the fourth block in a game. Relative fixation in the choice stage tracked the expected value of each option as well, but to a lesser extent, and shows learning effects. Error bars are SEM. (B) Model evidence (bars) and model comparison (numbers below bars) show that lazy Kalman filter learning with an upper confidence bound choice rule (KFL-UCB), captures participants' choices best. Error bars are interquartile ranges of bridge sampling repetitions (for some models too small to be visible; *SI Appendix, Methods*). (C) Mean probability with which each model accurately predicts participants' choices are well above chance level (dashed black line) and above a non-learning model that estimates fixed probabilities of choosing each option (dotted black line). The probability is highest for the KFL-UCB model (blue line). Means are computed over a rolling window of five trials. (D) Posterior of the group-level parameter for the KFL-UCB model that acts as a weight on uncertainty in the UCB choice rule (β). The posterior mean (vertical line) and 95% credible interval (black bar on the x-axis) shows the magnitude of uncertainty influence. Dots are posterior means of individual game level parameters.

and how fixation influences choice.

Properties of the fixation process in the choice stage. To analyse interactions between choice and fixation, we focus on the choice stage of a trial (Fig. 1B). Here participants had five seconds to consider which option to choose, before continuing to the next stage where they had to execute their choice, quickly. The fixation measure of interest in this section is the proportion of time spent fixating on each of the options. We computed the sum of the fixation durations received by each option, and divided this quantity by the sum total of fixation durations over all options. We refer to this measure of visual fixation as *relative fixation*.

Relative fixation resembled the allocation of choice, with increased allocation to high ranking options as learning progressed (Fig. 2A). This close correspondence to the choice distribution, including the gradual shift of fixation distribution toward high value options over time, is a first indication that relative fixation might be affected by the same learning process that is guiding choices, as we originally hypothesized. Importantly, relative fixation followed the expected value of each option (i.e. option rank) to a lesser extent than choice proportions (Fig. 2A). This could be due to a greater role of uncertainty in the trial-by-trial fixation dynamics, but could also be attributable to external, potentially independent, factors. Also as expected, and consistent with a reduction in uncertainty, the total time spent fixating on any of the options decreased over the course of learning (mixed effects regression estimates: intercept = 3.82, 95% CI [3.62, 4.02]; block = −0.20, 95% CI [−0.28, −0.13]; game = −0.05, 95% CI [−0.25, 0.15]; block×game = 0, 95% CI [−0.07, 0.07]; Fig. 3A). As for choice performance, there was no clear difference between the games. For analysis of other measures of the depth and breadth of the visual search process in the choice stage, see *SI Appendix, "Additional properties of visual fixation" in Results*.

Visual fixations in the choice stage are guided by both value and uncertainty. Given these suggestive results, we considered the conjoint influence of value and uncertainty on fixation in more detail. Previous studies that examined the relationship between choice and fixation (18, 21, 40) could not do this, since they used one-shot choices which precluded modelling of learning and thereby examining the role of uncertainty. To examine such influences, we regressed estimates of value and uncertainty from the KFL-UCB model fitting choices best on relative fixation in each trial (see "Modelling relative fixation in the choice stage" in *Materials and Methods*). Importantly, it was beliefs about values and uncertainty that were established at the end of the one trial that were used to explain variation in relative fixation in the next trial. We assumed that relative fixation followed a Dirichlet distribution whose shape was influenced by value, uncertainty and a game type indicator as a control variable, and whose scale was set by a separate parameter (Eq. 6 and 7).

As predicted, the results of Bayesian hierarchical estimation show a clear positive contribution of both value and uncertainty in explaining variability in relative fixation. The whole of the measurable posterior distribution of the value parameter (Val; Eq. 7) was on the positive side of zero (mean of 0.17, 95% CI $[0.12, 0.22]$; Fig. 3B) and the same holds for the uncertainty parameter (Unc; Eq. 7; mean of 0.12, 95% CI $[0.06, 0.17]$; Fig. 3C). Estimated game-type effects were negligible (mean of -0.002, 95% CI $[-9.66, 9.64]$; Eq. 7), while the estimated scale parameter mostly acted to flatten the predicted relative fixation further (mean κ parameter was 0.60, 95% CI $[0.50, 0.70]$; Eq. 6). We verified these results by additionally comparing the full model to two simpler models where we either regressed uncertainty alone or value alone on relative fixation, keeping the game type indicator as a control variable (Fig. 3D). The results of model comparison show that the model with both value and uncertainty clearly explains the relative fixation best (posterior probability of approximately 1), with simpler models lagging far behind. Hence, options with larger value and estimation uncertainty learned from previous trials attracted more relative fixation in the current trial. Thus, the same value and estimation uncertainty quantities that underlie block-wise changes in choice underlie block-wise changes in fixation allocation.

Visual fixations in the choice stage influence choice. Having established that value and uncertainty affect the fixation process

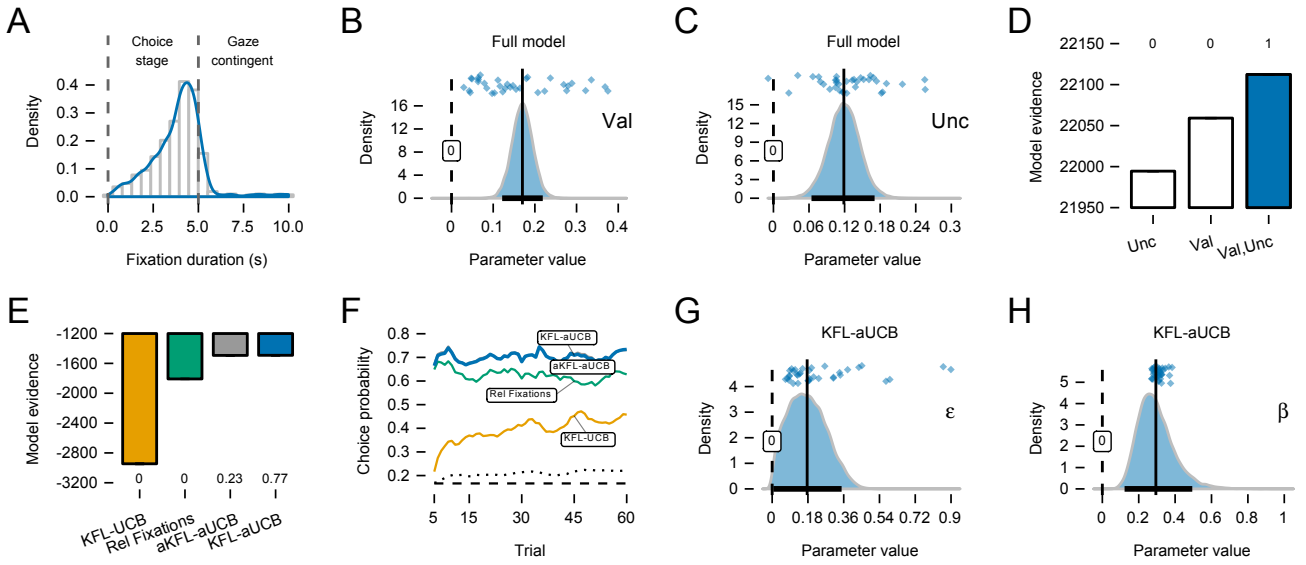

Fig. 3. Interactions between choice and relative fixation in the choice stage. (A) Density of total fixation duration for all games. Options disappear after 5 s, but participants sometimes kept fixating on the same location before triggering the execution stage. (B) Posteriors of the group-level value (Val) and (C) uncertainty parameter (Unc) in the full model regressing value and uncertainty on relative fixation. Both parameters are clearly positive, as evident from the mean (vertical line) and 95% credible intervals entirely above zero (CI, black bar on the x-axis). Dots are posterior means of individual game level parameters. (D) Model evidence (bars) and model comparison (numbers above bars) for the full model and simpler models that regressed either value or uncertainty alone. The full model fits the data best. Error bars are interquartile ranges of bridge sampling repetitions (too small to be fully visible; *SI Appendix, Methods*). (E) Choice model modulated by relative fixation (KFL-aUCB) outperforms the model that regressed relative fixation directly on choices. This indicates that modelling learning and the choice process is important even when relative fixation is taken into account. The KFL-aUCB model also outperforms the model where learning process is modulated as well (aKFL-aUCB), the KFL-UCB model was included for comparison. (F) The KFL-aUCB model predicts participants' choices with the highest mean probability. All three are well above the chance level (dashed line) and a non-learning model that estimates fixed probabilities of choosing options (dotted line). Means are computed in a rolling window of five trials. (G) The group-level ϵ parameter in the KFL-aUCB, which determines a pseudo relative fixation for options that were not fixated, is small and closer to zero, indicating that relative fixation was useful as is. (H) The group-level β parameter from the UCB choice rule in the KFL-aUCB model shows a decrease in the magnitude of the weight placed on uncertainty after accounting for relative fixation, but the weight is still substantial.

in the choice stage, we next examined whether visual fixation influenced choices. Such an influence has been shown in one-shot value-based choices (18, 40), but not yet for choices in a learning setting.

We first examined the effect of visual fixations on choices by regressing relative fixations in the choice stage directly on choices, using a simple multinomial logistic regression model (see "Modelling choices with visual fixations alone" in *Materials and Methods*). The results of Bayesian hierarchical estimation show that this simple model has a posterior probability of approximately 1 in comparison to the KFL-UCB model that fit choices best previously. What is surprising is the margin by which this simple model outperforms the KFL-UCB model, as shown clearly when examining the probability of accurately predicting participants' choices (Fig. 3F). Here it is evident that the ability of the simple regression model to predict choice is almost twice that of KFL-UCB, reaching a mean of 0.63 ($SE = 0.08$) by trial 60. This result establishes a strong effect of visual fixation on choice, suggesting the presence of a large choice-related but value- and uncertainty-independent component in visual fixations, which is not captured in our KFL-UCB model.

Values and uncertainty are not completely reflected in visual fixations. The excellent fit of choice using purely visual fixations prompts the question as to whether the effect of value and uncertainty on choice (KFL-UCB model; Fig. 2B) is mediated by their modest effect on fixation (Fig. 3D), or whether a part of the valuation process that enters choice is not reflected in visual fixation. To test this, we incorporated relative fixation into the best fitting KFL-UCB model (KFL-aUCB model – "a" prefix marks "attention-modulated"; see "Modelling learning and choices modulated by visual fixations" in *Materials and Methods*) and examined whether this variant describes choice better than a simple model regressing relative fixation on choice. There are various ways in which relative fixation might be included; here, we assumed that values and uncertainty of options are warped in proportion to the relative fixation that options capture (Eq. 12).

Bayesian hierarchical estimation showed that the KFL-aUCB model outperformed the simple regression model, describing participants' choices best with a posterior probability of approximately 0.77 (Fig. 3E; we included the KFL-UCB base model as well for comparison). The aKFL-aUCB model, in which learning process was modulated as well, followed suit with a posterior probability of approximately 0.23. Examining the models' probability of accurately predicting participants' choices again, we see a clear improvement over the simple regression model, with a constant advantage for the KFL-aUCB model throughout the game, reaching a mean of 0.73 ($SE = 0.07$) by trial 60 (Fig. 3F). This provides evidence that value and uncertainty are not completely reflected in visual fixation, and that explicitly modelling learning and choice processes provides additional predictive power. As a robustness check, we fitted additional attention-modulated models with a Softmax choice rule instead of UCB and a Kalman filter lacking the "laziness" parameter (*SI Appendix, "Comparison of learning and choice models modulated by visual fixation"in Results* and Fig. S6). The results showed that the UCB component

is important, as all models with it substantially outperform Softmax based models. The laziness parameter is important as well, but it has comparatively smaller impact.

We can compare the β parameter governing the strength of uncertainty-guidance in the UCB choice rule between the KFL-aUCB and KFL-UCB models. The posterior of β in KFL-aUCB is still clearly positive, but its magnitude was less once relative fixation is taken into account (posterior mean of 0.29, 95% CI $[0.12, 0.49]$; Eq. 4; Fig. 3H) – about 80% of the value for β in the KFL-UCB model without fixation modulation (Fig. 2D). Thus, some of the effect through which more uncertain options are more likely to be selected is sublimated when relative fixation is also taken into account.

In the KFL-aUCB model, the attention distribution over options was generated by squashing the relative fixation statistics according to a parameter ϵ (Eq. 11). The inferred value of this parameter can inform us about the importance of relative fixation. If ϵ is near 1, the distribution would be near uniform, independent of the relative fixation. If ϵ is near 0, then the distribution is dominated by the allocation of looking time. Consistent with the other analyses, the posterior distribution of ϵ parameter was small, with a mean value of 0.18 and 95% CI $[0.01, 0.35]$ (Fig. 3G).

Interactions between learning and fixation process. For analysing interactions between learning and fixation process we focus on the outcome stage of a trial (Fig. 1B), the three-second period during which participants could observe the reward outcome of their choice. The fixation measure of interest in this section is the total time fixating on the reward feedback in each trial. We will refer to this measure as *absolute fixation*. As for choice, we first examine the statistics of this measure, and then consider successively the effect of value and uncertainty on it and finally its potentially additional effect on learning.

Properties of the fixation process in the outcome stage. We first considered trial-by-trial variability in absolute fixation. Mean absolute fixation decreased over the course of learning and there are some, albeit weak, differences between the games (mixed effects regression estimates: intercept = 2.36, 95% CI $[2.20, 2.52]$; block = -0.10, 95% CI $[-0.14, -0.05]$; game = -0.06, 95% CI $[-0.22, 0.10]$; block×game = 0.06, 95% CI $[0.01, 0.11]$). The negative effect of the block is circumstantial evidence that uncertainty, which also decreases over the course of learning, is related to absolute fixation (Fig. 4B). There was a ceiling effect due to the three-second outcome presentation time and this led to a left skewed distribution of absolute fixation (Fig. 4A), but a mean of 2.36 s indicates that the effect was not particularly strong. Participants often continued looking at the feedback location for a few seconds more during the inter-trial interval (Fig. 4A). We assumed that these fixations were also associated with processing the reward feedback and included last fixations that ended within two seconds of the inter-trial interval. Most importantly for our subsequent considerations, when we repeat the same analysis on the standard deviations of absolute fixation, we observe considerable variability in absolute fixation (mixed effects regression estimates: intercept = 0.85, 95% CI $[0.74, 0.96]$; block = 0.07, 95% CI $[0.02, 0.07]$; game = 0.03, 95% CI $[-0.08, 0.14]$; block×game = -0.02, 95% CI $[-0.06, 0.03]$), as evidenced by the intercept estimate. For analysis of other measures of the visual search process, see *SI Appendix, "Additional properties of visual fixation" in Results*.

Unsigned reward prediction error guides fixation in the outcome stage. We next examined interactions between learning and fixation, focusing first on the theory-driven expectation that time spent looking at the reward feedback is guided by uncertainty, as is the case for the learning rate (3). There are two measures of uncertainty of interest here. One is the estimation uncertainty derived from the Kalman filter learning model (S variable, Eq. 2), the same quantity used in the UCB choice rule. The other is based on the prediction error and reflects both estimation and irreducible uncertainty (41). As predictions improve and estimation uncertainty decreases, unsigned (i.e. absolute) prediction error should generally decrease as well. However, because unsigned prediction error contains irreducible uncertainty (i.e. the variance of options' reward distributions), it will have continuing fluctuations as well, giving it a momentary character. Prediction errors play no role in uncertainty computations in the Kalman filter (Eq. 2), so these two measures should be largely decoupled. Indeed, the correlation between the two measures is negligible, with an average correlation across participants of 0.02 ($SE = 0.11$).

We regressed trial-by-trial uncertainty, prediction error, unsigned prediction error and value obtained from the KFL-UCB model on absolute fixation (see "Modelling absolute fixation in the outcome stage" in *Materials and Methods*, Eq. 8 and 9). We assumed absolute fixation follows a skew normal distribution constrained to the $(0, 5)$ interval (Fig. 4A) and we included a game type indicator as a control variable (Eq. 9). We compared the full model with all four predictors to simpler models that excluded particular predictors (*Materials and Methods*). The results of these model comparisons (Fig. S7), which naturally take into account model complexity, show that a model including only unsigned prediction error (uPE model) explained absolute fixation best ($P = 0.58$), with a model including unsigned prediction error and value (uPE, Unc model) following suit ($P = 0.28$). In the uPE model, the effect of unsigned prediction error was clearly positive (Fig. 4C), with almost the entire posterior distribution on the positive side (mean of 0.05; 95% CI $[0.02, 0.07]$). This means that reward outcomes accompanied with large unsigned prediction error tended to attract longer absolute fixation.

These results suggest that unsigned prediction error could in principle be a more important form of uncertainty, than estimation uncertainty, for guiding choice. On this basis we re-examined whether a class of models that uses unsigned prediction error, instead of estimation uncertainty, in the UCB choice rule might explain choices better than the KFL-UCB model. We implemented two models. The KFL-UPE model used a simple delta-rule to learn slow-moving estimates of unsigned prediction errors coming from the lazy Kalman filter learning model. These estimates were then used in the UCB rule. The K2-UPE model uses instead the K2 learning model which computes estimates of unsigned prediction errors in a more principled manner, following (41). However, the KFL-UCB model outperformed both models with a posterior probability of approximately 1 (*SI Appendix, "Choice models with unsigned prediction errors" in Results* and Fig. S4). Evidently estimation uncertainty is more relevant for guiding choice than unsigned prediction errors.

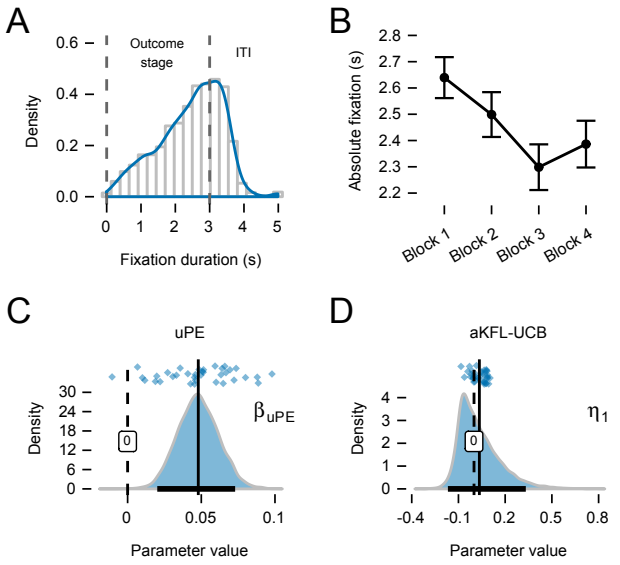

Fig. 4. Interactions between learning and fixation processes at outcome stage. (A) Density of absolute fixation in the outcome stage. Even though the option and feedback disappear after 3 s participants often kept fixating on the same location during the inter-trial interval (ITI). Fixations that extended 2 s into the ITI (i.e. 5 s in total) were also used in the analysis. (B) Like uncertainty and unsigned reward prediction errors, absolute fixation decreased over the course of learning. (C) Posterior of the group-level slope parameter in the model regressing unsigned reward prediction error (β_{uPE}) on absolute fixations in the outcome stage. Almost complete posterior is positive, including the 95% credible interval (CI, black bar on the x-axis), indicating a clearly positive relationship. (D) The group-level slope parameter (η_1) that biases the learning rate in the aKFL-UCB model has a positive mean, suggesting a reduced bias for longer fixation on the feedback; however, the CI includes zero.

Fixation in the outcome stage influences the learning rate. Given our finding that learning influences visual fixations in the outcome stage we next considered whether there was a relation in the other direction, i.e., whether fixations affected the course of learning. As for choice, we tested this by comparing the KFL-UCB model that fitted choices best to a similar model in which we allowed absolute fixation at the outcome stage to modulate the learning rate, now referred to as aKFL-UCB (*Materials and Methods*, Eq. 13, 2 and 4). We decomposed the laziness parameter η of the lazy Kalman filter into an intercept η_0 and a slope η_1 that multiplies the absolute fixation in the outcome stage.

The slope η_1 is the main parameter of interest in the aKFL-UCB model. While the larger portion of its posterior is positive, with a mean of 0.03, the 95% CI $[-0.17, 0.33]$ includes zero, suggesting the overall effect is weak (Eq. 13; Fig. 4E). To further assess its significance, we compared the aKFL-UCB model to the KFL-UCB model where learning is not modulated by absolute fixation. The KFL-UCB model outperformed the aKFL-UCB model, with a posterior probability of approximately 0.98, suggesting that absolute fixation does not modulate the learning rate.

Discussion

This study enriches our understanding of human reinforcement learning behaviour by looking at the four-way interaction between uncertainty, choice, learning and visual fixation. Our results offer evidence that people learn and choose in partial accordance with normative models, leveraging estimation uncertainty for both choice and learning. We show novel influences of fixation in reinforcement learning. Signatures of directed exploration can be seen in relative fixation at choice, which goes beyond previous findings on the effects of value and irreducible uncertainty on fixation at choice. Lastly, we provide novel evidence for the theoretical prediction that fixation at outcome is modulated by estimation uncertainty.

Examining choices alone supports a model where exploration is guided by both value and estimation uncertainty. The winning KFL-UCB model adds an "exploration bonus" to options' expected rewards (14, 38). This model can be viewed as an approximation to the optimal solution for multi-armed bandit problems (7, 42) and adds to a growing body of evidence that people use uncertainty-guided choice strategies (8–12). The KFL-UCB also includes a Bayesian learning component (Kalman filter) which adapts its learning rate according to uncertainty. This dovetails with previous studies demonstrating a dynamic modulation of learning rate by uncertainty (4, 6). Our results imply that people track uncertainty about estimated value and incorporate it in their choices. This aligns with evidence from perceptual decision making that people have well-calibrated confidence in their choices (43), and from bandit tasks that they have accurate sense of confidence in their value estimates (10, 44). Indeed, neuroimaging studies show that the brain tracks both mean and variance (45, 46), while studies of neuronal population activity support a coding scheme where both mean and variance are represented (47, 48).

Our analyses of visual fixation during choice provide novel evidence on the role of estimation uncertainty in choice. During the choice stage where participants considered which option to choose, we found that both value and estimation uncertainty, derived from estimation based on all previous trials, guided visual fixation in the current trial. Hence, directed exploration principles guide both choice and fixation. Examining choices alone do not always reveal the role of estimation uncertainty in exploration (5, 13), but including fixation may provide a more reliable method to decode its role. Previous studies (18, 21, 40) mostly focused on one-shot choices and hence could not examine whether and how visual fixation during choice is influenced by learning history, neither value nor estimation uncertainty. There are several exceptions. Perhaps the closest to the present study is recent work by Leong and colleagues (22), who show that fixation during choice is influenced by value learned from previous trials. However, the authors did not consider models that track uncertainty about value. Another recent study by Walker and colleagues (24) showed that irreducible uncertainty increases exploration in both choice and attention, i.e. less focus on best options. However, their study used a between-subjects design and cannot explain what components of learning drive fixation on a trial-by-trial basis. Consequently, their results are inconclusive about the role of estimation uncertainty. Several other studies that examined relation between choice and attention in reinforcement learning eliminated the exploration aspect of the task and hence did not examine the role of estimation uncertainty (25, 49).

We found that unsigned prediction errors guide visual fixation on the reward feedback during learning. Because estimation uncertainty modulates the learning rate, we expected it would guide fixation (3). Our additional prediction was that reward prediction errors might also influence fixation, as these

indirectly incorporate both estimation and irreducible uncertainty. As learning progresses, estimated value becomes more accurate and prediction errors correspondingly decrease, thus mimicking the decrease in estimation uncertainty over time. Because prediction errors are influenced by irreducible uncertainty, they track both fast-moving momentary uncertainty and slow-moving estimation uncertainty. Looking at relative fixations to aversive stimuli in a conditioning task, (16) also found evidence for the influence of momentary uncertainty during the outcome stage. Results of both studies jointly provide supportive evidence for a prediction based on (3) that fixation should be related to option uncertainty, following the precepts of Bayesian statistical learning. Interestingly, we did not find that performance of a model where we allowed absolute fixation at the outcome stage to modulate the learning process (aKFL-UCB) improved over a model without fixation modulation (KFL-UCB). This result suggests fixation reflects the update process rather than having an influence on it. By contrast, (16) and (22) found evidence for such modulation. In (16) learning process was directly observed and in (22) fixation measure was more detailed, tracking various features of options. These differences likely resulted in a greater sensitivity for detecting the fixation modulation in these studies.

Relative fixation in the choice stage exerted a stronger influence on choice than warranted by the information about value and estimation uncertainty contained in it. In fact, choices were better predicted from relative fixation alone than by the KFL-UCB model. This suggests that fixation carries additional choice relevant factors which are potentially unrelated to value and estimation uncertainty. For example, low level features of the symbols denoting individual options may have attracted gaze and biased choice toward those options (28). Such effects are anticipated by an attention modulated sequential sampling model (18). Here, we identify the magnitude of this modulation in a learning setting: our ability to predict choice nearly doubles, even for early trials that are usually difficult to predict by reinforcement learning models (Fig. 3F). This indicates that much can be gained by taking into account the visual search process in modelling learning and choices. The KFL-aUCB model, an example of how fixations can be incorporated into reinforcement learning models, explained choice better than relative fixation alone. This suggests that value and estimation uncertainty influenced choices both directly, through an internal valuation process, and indirectly, via fixation. This result invites an interesting conjecture about directed and random exploration (9). The source of directed exploration might be an internal choice process, while that of random exploration might lie in fixation specific factors unrelated to decision variables.

In tasks where people learn about options' values from reward feedback, looking at the options in the choice stage does not convey new information *per se*. In learning tasks, quantities such as estimated value and associated uncertainty must be represented in memory rather than externally. This raises the question of why participants' fixations in the choice stage were informative of their choices. To make an informed choice between the options, participants will likely retrieve experienced rewards or other indicators of options' value from memory. Looking at the stimuli, even though not informative per se, can facilitate memory retrieval and working memory operations (31, 32, 50, 51). This is akin to the rationale behind sequential sampling mechanisms in one-shot value based decision making. (18) hypothesized that the brain accumulates evidence by extracting the features of choice options, retrieving their learned values from the memory, and integrating these for each option. Similar assumptions underlie integrated reinforcement learning and sequential sampling models (20, 52–54). A negative side effect is that fixations can introduce bias, as suggested by (18). Our findings provide insight into the nature of this bias. Being shaped by the learning history the bias is partly adaptive, as a subset of fixations reflect cognitive processes behind directed exploration.

The attentional drift diffusion model by Krajbich and colleagues (18) is an appealing account of the within-trial choice process and how this may be influenced by fixation. Recent models combining reinforcement learning and sequential sampling have added across-trial learning dynamics (52–54). These models are not applicable in our task as the choice stage was fixed to 5 seconds and separated from the execution (Fig. 1B). Therefore, response times are not informative about the evidence accumulation process. When we allowed for self-selected choice times in pilot experiments, we discovered that participants plan their next choice immediately after the feedback and during the inter-trial interval, making the collection of useful eye-movement data difficult. While such separation seems artificial in a laboratory task, it arguably brings the task closer to real-world situations. For instance, purchasing a certain type of product in a supermarket might happen every few days, effectively separating the choice opportunities and forcing the consumer to make a final choice once they are in front of the shelf. Applying sequential sampling models would require experimental designs that solve the issue of deciding in non-choice time in a different way. One potential solution would be to use several bandit problems simultaneously and on each trial randomly assign one of these, thereby reducing the usefulness of planning a choice before choice options are presented. Another is to use a contextual bandit problem, where new options can be presented on every trial, while learning would allow making useful predictions about the value of these new options (10, 22, 55).

One pertinent question is how our results regarding visual fixations relate to the role of attention in reinforcement learning. In theoretical work on associative learning in nonhuman animals, the Mackintosh model (56) predicts that stimuli with high predictive value should attract attention, while the Pearce-Hall model (2) predicts that uncertainty has a primary role. These seemingly contradictory accounts of attention have both received empirical support (57). (3) reconciled the two accounts, proposing that both are correct, but at different stages: during choice, attention is guided by predictive value, whilst during learning it is guided by uncertainty. Our results are consistent with this latter account. Fixations during the outcome stage were mainly driven by unsigned prediction errors, the measure of surprise in the Pearce-Hall model (2). Our results for relative fixations in the choice stage support an extension of the (56) account based on approximately optimal solutions to the exploration-exploitation trade-off (14, 38). In this extension, both value and estimation uncertainty play a role in the choice stage.

Although imperfect, eye movements provide trial-by-trial empirical measures of attention. By recording fixations, attention need not be inferred solely from a computational model

(58–60). But there is scope for further integrating measured attention into our models. Rather than using fixations as exogenous modulators of learning and choice, as we have done here (see also 22), a more satisfying treatment would endogenise fixations in a model that learns to direct attention and choose both within and across trials. Research in vision science has suggested that in tasks such as scene viewing (61) and visual search (62), eye movements are guided by visual information gain. Sprague and Ballard (63) proposed a reinforcement learning model of eye movements where uncertainty guides eye movements. In their model, eye movements to visually uncertain stimuli are reinforced because learning about the identity or state of the stimuli result in decisions that maximize the amount of reward. Previous studies have provided qualitative support for the model, albeit not in a reinforcement learning context (64). Manohar and Husain (65) modelled fixations in one-shot choices between monetary gambles where the authors argued that visual attention aims to minimize uncertainty about the expected value of gambles. In the latter study, as well as those concerning visual scene detection, fixation directly provides novel information. This contrasts with our study, where fixating on an option can benefit memory retrieval, which in turn may serve a similar aim of information gain. This then paves the way to extending previous efforts to endogenise fixations to the current setting, a focus of future research that we plan.

In summary, we provide a detailed window on the interplay between learning, choice, and visual fixation, that allow us to trace the path through which uncertainty affects behaviour. Our study has theoretical and practical implications. First, it shows that attention and reinforcement learning processes might be more intertwined than previously thought, prompting a need for closer integration of the two in the future studies. It also raises new questions, such as whether the source of random exploration can be traced to the learning-independent properties of the fixation process. Second, it illustrates the utility of monitoring eye movements during learning and choice. The ability of reinforcement learning models to predict individual choice substantially improves when fixations are taken into account. Third, since fixations are shaped by learned values and associated uncertainties the potential for fixation to bias choice is smaller. Finally, the same result could explain everyday phenomena such as what shelf space in supermarkets people pay attention to and how companies can leverage this to induce exploration of new products.

Materials and Methods

Participants. We recruited 34 participants (18 female, $M_{\text{age}} = 26.8$ and $SD_{\text{age}} = 8.1$) from the Aarhus University subject pool. After applying *a priori* exclusion criteria separately to each game played by each participant, 23 participants remained (12 female, $M_{\text{age}} = 26.9$ and $SD_{\text{age}} = 8.4$), 36 games in total, 19 Decreasing variances and 17 V-shaped variances game (see *SI Appendix, Methods* for details). The experimental sessions were conducted individually in the COBE laboratory at Aarhus University and lasted for 75 minutes on average. Participants had normal or corrected to normal vision. The study was approved by the Aarhus University Research Ethics Committee and all participants provided written informed consent. Participants received a show-up fee of 100 Danish krone and an additional performance-contingent bonus (100 Danish krone on average).

Task. The experiment was comprised of two separate multi-armed bandit (MAB) tasks (games) with 60 trials each. In each task, participants made repeated choices between the same six options, represented by different symbols (Fig. 1A) and shown in the same location on each trial. Key stages of a trial were the *choice stage* and *outcome stage*. In the choice stage options were presented for a fixed duration of 5 s, during which participants considered which option to choose. They registered their choice in the execution stage that followed the choice stage. In the outcome stage participants were shown reward feedback overlaid over the chosen option for 3 s. Participants were instructed to maximize the cumulative sum of the rewards during each task.

The main difference between the games was in the variance of the rewards. In the *Decreasing variances* game, the variance of each option decreased from the best option to the worst (according to expected reward). In the *V-shaped variances* game the variance decreased from the best option to the third best, and then increased again from the fourth best to the worst option. To minimize carry-over effects between the games, we used a different set of letters from Gljagolica alphabet (Fig. 1A) and rescaled rewards differently for each game. The alphabet letters, options' locations, the order of the games, and the currencies and scaling factors associated with each game were randomized. At the end of each game participants received feedback about the experimental points they accumulated and corresponding earnings. After participants finished both games, we informed them which game was randomly selected for the payout, debriefed them, and paid their earnings. A detailed description of the time course of each trial, stimuli construction in each game and procedure is provided in *SI Appendix, Methods*.

Eye tracking. Participants sat in front of a screen with resolution of 1650×1050 pixels and physical size of 475×297 mm (widths and heights, respectively). They used a chinrest at approximately 60 cm distance from the screen. We recorded eye movements and pupillary responses using a desk-mounted EyeLink 1000 eye tracker (SR Research, https://www.sr-research.com/) with a monocular sampling rate of 500 Hz. We performed a 13-point calibration with the dominant eye, followed by a 13-point drift validation test. We accepted calibrations with offset less than 1°of visual angle. In gaze contingent stages of the trial – triggering the onset of the choice and execution stage – 90% of gaze locations within a 1 s window needed to be in a circular area with a 3 cm radius around the fixation cross. To make a response in the execution stage participants had to press a key and an eye data sample had to be recorded at the same time within a circle representing an option. We used the default algorithm provided by SR Research to detect fixations. In data analysis we drew an area of interest (AOI) with radius of 3 cm around the centre of every option and assigned all fixations falling into these AOI to the corresponding options. See *SI Appendix, Methods* for further details on the eye-tracking setup.

Data analysis. We present here an abbreviated overview of analyses and models. More detailed descriptions, together with model fitting and comparison procedures, are given in *SI Appendix, Methods*.

Mixed effect regressions. We examined learning effects in games and differences between game types using Bayesian mixed effect regressions. We computed averages across blocks and regressed an intercept, a block indicator (coded as $[-1.5, -0.5, 0.5, 1.5]$ for blocks 1 to 4) and a game type indicator (coded as -1 for Decreasing variances and 1 for V-shaped variances game), as well as their interaction on choice performance (chosen option rank) and fixation measures in the choice and outcome stage (total fixation duration, number fixations and number of options fixated). Intercept and blocks were entered as game-specific random effects while game type was entered as a fixed effect. Credible intervals were computed as highest posterior density intervals.

Modelling learning and choices. We fitted four main computational models to participants' choices. Each model consists of a learning and a choice component. The learning component is either a Kalman filter (KF) (8, 13, 36) or a "lazy" Kalman filter (KFL) model. For the choice component the models used either a softmax (SM; 37), or an upper confidence bound choice rule (UCB; 14).

The Kalman filter model assumes participants update their estimates $E_j(t+1)$ of the expected reward of choosing option j on

trial $t+1$ from the observed reward $R_j(t)$ on trial t as

$$E_j(t+1) = E_j(t) + I_j(t)K_j(t)[R_j(t) - E_j(t)]. \quad [1]$$

where the so-called "Kalman gain" term $K_j(t)$ acts as a learning rate. Term $I_j(t)$ is a simple indicator variable, with value of 1 if option j is chosen on trial t and 0 otherwise. The Kalman gain is updated on every trial and depends on current level of uncertainty

$$K_j(t) = \eta \frac{S_j(t) + \sigma_\zeta^2}{S_j(t) + \sigma_\zeta^2 + \sigma_{\epsilon,j}^2}, \quad [2]$$

where $S_j(t)$ is the variance of the posterior distribution of the mean reward, updated in every trial as $S_j(t+1) = [1 - I_j(t)K_j(t)][S_j(t) + \sigma_\zeta^2]$; σ_ζ^2 is the innovation variance and $\sigma_{\epsilon,j}^2$ the reward variance parameter which modulate the learning rate. Parameter $\eta \in (0, 1)$ determines a bias in the Kalman gain, allowing the filter to learn at slower pace (hence the term "lazy"). In the standard Kalman filter we fixed this parameter to $\eta = 1$, while in lazy versions it is an estimated parameter. In both variants we initialized estimate of the expected value to $E_j(0) = 0$. Initial variance was a free parameter σ_i^2 such that $S_j(0) = \sigma_i^2$. We take into account differences between variances of options by setting the $\sigma_{\epsilon,j}^2$ parameter to option's objective variance that we used to draw rewards from: $[2.75, 2.35, 1.95, 1.55, 1.15, 0.75]$ in Decreasing variances and $[2.75, 2.35, 1.95, 1.95, 2.35, 2.75]$ in V-shaped variances game.

In the softmax choice rule participants choose probabilistically according to relative estimated value

$$P(C(t) = j) = \frac{\exp[\theta E_j(t)]}{\sum_{k=1}^{6} \exp[\theta E_k(t)]}, \quad [3]$$

where $P(C(t) = j)$ is probability of choosing option j at trial t and the inverse temperature parameter $\theta > 0$ determines the sensitivity to differences in estimated values, and with it the amount of exploration.

The upper confidence bound choice rule combines estimated value and estimation uncertainty

$$P(C(t) = j) = \frac{\exp\{\theta(E_j(t) + \beta\sqrt{S_j(t)})\}}{\sum_{k=1}^{6} \exp\{\theta(E_k(t) + \beta\sqrt{S_k(t)})\}}, \quad [4]$$

where $\beta > 0$ is the weight a participant places on estimation uncertainty. While the original UCB rule chooses the option with the highest resulting value deterministically, we implemented a stochastic version by using a softmax transformation.

Modelling relative fixation in the choice stage. We used trial-by-trial subjective estimates of value and uncertainty from the KFL-UCB model fitting choices best, and regressed them on relative fixations in the choice stage. We controlled for potential differences between games by including a game-type indicator. Relative fixations were operationalised as the summed duration of fixations on each of the options divided by the sum of these quantities across all options.

We assume that relative fixations in the choice stage (RF) follow a Dirichlet distribution

$$\mathrm{RF}(t) \sim D(\boldsymbol{\alpha}(t), \kappa), \quad [5]$$

with the probability density function defined as

$$\frac{1}{\mathrm{B}(\boldsymbol{\alpha}(t)\kappa)} \prod_{j=1}^{6} \mathrm{RF}_j^{\alpha_j(t)\kappa - 1}, \quad [6]$$

where $\mathrm{B}(\boldsymbol{\alpha}(t)\kappa)$ is a multinomial beta function that acts as a normalising constant. The vector of concentration parameters $\boldsymbol{\alpha}(t)$ for each trial is obtained by passing values $(E_j(t))$ and estimation uncertainty $(S_j(t))$ of each option j obtained from the KFL-UCB model, as well as game type indicator as a control variable (G), through a softmax function

$$\boldsymbol{\alpha}(t) = \frac{\exp\{\beta_v E_j(t) + \beta_u \log S_j(t) + \beta_{gt} G\}}{\sum_{k=1}^{6} \exp\{\beta_v E_k(t) + \beta_u \log S_k(t) + \beta_{gt} G\}}, \quad [7]$$

where β_v and β_u are weights on value and uncertainty, while β_{gt} is the effect of game type. We log-transformed estimation uncertainty to linearise it. Games were coded as $G = -1$ for Decreasing variances and $G = 1$ for V-shaped variances game and this effect was included at a group level only. We assumed an additional precision parameter κ that multiplies the concentration parameters, governing how much probability mass is near the expected value.

Modelling absolute fixation in the outcome stage. We used trial-by-trial uncertainty, reward prediction errors, and value from the KFL-UCB model that fitted the choices the best and regressed them on absolute fixations in the outcome stage. We controlled for potential differences between games by including a game type variable. Absolute fixation measure was operationalised as a sum of durations of all fixations on the reward feedback.

We assumed fixation durations during outcome stage (F) follow a Skew Normal distribution

$$F(t) \sim N(\xi(t), \omega, \alpha), \quad [8]$$

truncated to interval $F(t) \in [0, 5]$. In the full model the location parameter $\xi(t)$ for each trial is a linear combination of intercept, uncertainty $(S_j(t))$, prediction error (PE), unsigned prediction error (uPE), and value $(E_j(t))$ of chosen option j obtained from the KFL-UCB model and game type indicator variable (G)

$$\xi(t) = \beta_i + \beta_u \log S_j(t) + \beta_{\mathrm{PE}}\mathrm{PE}_j(t) + \beta_{\mathrm{uPE}}|\mathrm{PE}|_j(t) + \beta_v E_j(t) + \beta_{gt} G, \quad [9]$$

where β_u, β_{PE}, β_{uPE} and β_v are weights on uncertainty, signed prediction errors, unsigned prediction errors and value, β_i is the intercept, and β_{gt} the effect of game type. We computed unsigned prediction errors as absolute value of the prediction error and we log-transformed estimation uncertainty to linearise it. Games were coded as $G = -1$ for Decreasing variances and $G = 1$ for V-shaped variances game and this effect was included at a group level only. We assumed an additional scale parameter ω and shape parameter α, modelled at an individual game level, without a group-wise parameter.

Modelling choices with visual fixations alone. We also regressed relative fixation in the choice stage alone on choices, without explicitly modelling the learning and choice process. We used a simple multinomial logistic regression model where relative fixation for option j in trial t, $\mathrm{RF}_j(t)$, is passed through a softmax function to obtain the probability $P(C(t) = j)$ of choosing option j at trial t

$$P(C(t) = j) = \frac{\exp[\tau \mathrm{RF}_j(t)]}{\sum_{k=1}^{6} \exp[\tau \mathrm{RF}_k(t)]}, \quad [10]$$

where the inverse temperature parameter $\tau > 0$ determines the sensitivity to differences in relative fixations.

To avoid the measure of relative fixation taking the value of zero for options that were not fixated on at all in certain trials, we assigned each option a minimum value of ϵ which was treated as a free parameter:

$$\mathrm{RF}_j(t) = \epsilon/6 + (1 - \epsilon)\frac{F_j(t)}{\sum_{k=1}^{6} F_k(t)}. \quad [11]$$

Modelling learning and choices modulated by visual fixations. We assumed visual fixations can modulate the choice or learning component of the KFL-UCB model. We mark the learning and choice component with an "a" prefix to indicate which aspect is modulated by fixations. For example, in the aKFL-UCB model, visual fixations modulate the learning process, while in the KFL-aUCB they modulate the choice process.

We assumed visual fixations in the choice stage enter the choice process by re-weighting the choice probabilities produced by the models based on options' estimated values and estimation uncertainty (Eq. 4). The relative fixation measure defined in Eq. 11 enters the UCB rule in an additive way:

$$P(C(t) = j) = \frac{\exp\{\tau \mathrm{RF}_j(t) + \theta(E_j(t) + \beta\sqrt{S_j(t)})\}}{\sum_{k=1}^{6} \exp\{\tau \mathrm{RF}_k(t) + \theta(E_k(t) + \beta\sqrt{S_k(t)})\}}. \quad [12]$$

We assumed visual fixations in the outcome stage influence the learning process by making the bias in the Kalman gain update dependent on how long the reward feedback was fixated on in total in the outcome stage of the trial. We implemented this by replacing

the η parameter in Eq. 2 with a baseline parameter η_0 and a slope parameter η_1 that depends on F, the absolute fixation duration in outcome stage:

$$\eta(t) = \Phi(\eta_0 + \eta_1 F(t)), \quad [13]$$

where Φ is the standard normal cumulative distribution function, used to constrain the resulting η parameter to the $(0, 1)$ range.

Data and code availability. The data, code used for our analyses, as well as other project-related files are publicly available at the Open Science Framework website: https://osf.io/539ps/ (66).

ACKNOWLEDGMENTS. We would like to thank Toby Wise, Eran Eldar, Nitzan Shahar and Rani Moran for their feedback on the project. We thank Anna Nason for help with collecting the data. H.S., J.L.O., and R.D. were funded by Lundbeckfonden, Grant number: R281-2018-27. H.S. and R.D. were funded by the Max Planck Society, Munich, Germany, https://www.mpg.de/en, Grant number: 647070403019. R.D. was also funded by the Wellcome Trust, https://wellcome.ac.uk/home, Grant number/reference: 098362/Z/12/Z. P.D. was funded by the Gatsby Charitable Foundation and the Max Planck Society.

1. BD Anderson, JB Moore, *Optimal filtering*. (Courier Corporation), (2012).
2. JM Pearce, G Hall, A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532 (1980).
3. P Dayan, S Kakade, PR Montague, Learning and selective attention. *Nat. Neurosci.* **3**, 1218 (2000).
4. TE Behrens, MW Woolrich, ME Walton, MF Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214 (2007).
5. E Payzan-LeNestour, P Bossaerts, Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput. Biol.* **7**, e1001048 (2011).
6. MR Nassar, RC Wilson, B Heasly, JI Gold, An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* **30**, 12366–12378 (2010).
7. J Gittins, K Glazebrook, R Weber, *Multi-armed bandit allocation indices*. (John Wiley & Sons), (2011).
8. M Speekenbrink, E Konstantinidis, Uncertainty and Exploration in a Restless Bandit Problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
9. RC Wilson, A Geana, JM White, EA Ludvig, JD Cohen, Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
10. H Stojic, E Schulz, PP Analytis, M Speekenbrink, It's new, but is it good? how generalization and uncertainty guide the exploration of novel options (2018).
11. SJ Gershman, Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42 (2018).
12. WB Knox, AR Otto, P Stone, B Love, The nature of belief-directed exploratory choice in human decision-making. *Front. Psychol.* **2**, 398 (2012).
13. ND Daw, JP O'Doherty, P Dayan, B Seymour, RJ Dolan, Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
14. P Auer, N Cesa-Bianchi, P Fischer, Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**, 235–256 (2002).
15. WR Thompson, On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **25**, 285–294 (1933).
16. T Wise, J Michely, P Dayan, RJ Dolan, A computational account of threat-related attentional bias. *PLOS Comput. Biol.* **15**, e1007341 (2019).
17. NJ Ashby, T Rakow, Eyes on the prize? evidence of diminishing attention to experienced and foregone outcomes in repeated experiential choice. *J. Behav. Decis. Mak.* **29**, 183–193 (2016).
18. I Krajbich, C Armel, A Rangel, Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).
19. A Konovalov, I Krajbich, Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nat. Commun.* **7**, 12438 (2016).
20. JF Cavanagh, TV Wiecki, A Kochar, MJ Frank, Eye tracking and pupillometry are indicators of dissociable latent decision processes. *J. Exp. Psychol. Gen.* **143**, 1476 (2014).
21. S Shimojo, C Simion, E Shimojo, C Scheier, Gaze bias both reflects and influences preference. *Nat. Neurosci.* **6**, 1317 (2003).
22. YC Leong, A Radulescu, R Daniel, V DeWoskin, Y Niv, Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron* **93**, 451–463 (2017).
23. M Schoemann, M Schulte-Mecklenbeck, F Renkewitz, S Scherbaum, Forward inference in risky choice: Mapping gaze and decision processes. *J. Behav. Decis. Mak.* **0** (In press).
24. AR Walker, D Luque, ME Le Pelley, T Beesley, The role of uncertainty in attentional and choice exploration. *Psychon. Bull. & Rev.* **26**, 1–6 (2019).
25. Y Hu, Y Kayaba, M Shum, Nonparametric learning rules from bandit experiments: The eyes have it! *Games Econ. Behav.* **81**, 215–231 (2013).
26. L Zhaoping, *Understanding Vision: Theory, Models, and Data*. (Oxford University Press, Oxford, UK), (2014).
27. JL Orquin, CJ Lagerkvist, Effects of salience are both short-and long-lived. *Acta Psychol.* **160**, 69–76 (2015).
28. JL Orquin, SM Loose, Attention and choice: A review on eye movements in decision making. *Acta Psychol.* **144**, 190–206 (2013).
29. M Usher, JL McClelland, The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.* **108**, 550 (2001).
30. R Ratcliff, G McKoon, The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
31. E Awh, J Jonides, Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* **5**, 119–126 (2001).
32. E Awh, EK Vogel, SH Oh, Interactions between attention and working memory. *Neuroscience* **139**, 201–208 (2006).
33. R Johansson, M Johansson, Look here, eye movements play a functional role in memory retrieval. *Psychol. Sci.* **25**, 236–242 (2014).
34. L Holm, T Mäntylä, Memory for scenes: Refixations reflect retrieval. *Mem. & Cogn.* **35**, 1664–1674 (2007).
35. M Usher, JD Cohen, D Servan-Schreiber, J Rajkowski, G Aston-Jones, The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554 (1999).
36. WK Zajkowski, M Kossut, RC Wilson, A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* **6**, e27430 (2017).
37. RS Sutton, AG Barto, *Reinforcement learning: An introduction*. (MIT Press, Cambridge, MA, US), (1998).
38. S Kakade, P Dayan, Dopamine: Generalization and bonuses. *Neural Networks* **15**, 549–559 (2002).
39. JK Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (Academic Press), (2014).
40. KC Armel, A Beaumel, A Rangel, Biasing simple choices by manipulating relative visual attention. *Judgm. Decis. making* **3**, 396–403 (2008).
41. RS Sutton, Gain adaptation beats least squares in *Proceedings of the 7th Yale workshop on adaptive and learning systems*. Vol. 161168, (1992).
42. P Whittle, Multi-Armed Bandits and the Gittins Index. *J. Royal Stat. Soc. Ser. B (Methodological)* **42**, 143–149 (1980).
43. R Moran, AR Teodorescu, M Usher, Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cogn. Psychol.* **78**, 99–147 (2015).
44. A Boldt, C Blundell, B De Martino, Confidence modulates exploration and exploitation in value-based learning. *Neurosci. Conscious.* **2019**, niz004 (2019).
45. M Symmonds, ND Wright, DR Bach, RJ Dolan, Deconstructing risk: Separable encoding of variance and skewness in the brain. *Neuroimage* **58**, 1139–1149 (2011).
46. HD Critchley, CJ Mathias, RJ Dolan, Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* **29**, 537–545 (2001).
47. WJ Ma, M Jazayeri, Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).
48. WJ Ma, JM Beck, PE Latham, A Pouget, Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432 (2006).
49. T Beesley, KP Nguyen, D Pearson, ME Le Pelley, Uncertainty and predictiveness determine attention to cues during human associative learning. *The Q. J. Exp. Psychol.* **68**, 2175–2199 (2015).
50. J Theeuwes, A Belopolsky, CN Olivers, Interactions between working memory, attention and eye movements. *Acta Psychol.* **132**, 106–114 (2009).
51. A Kiyonaga, T Egner, Working memory as internal attention: toward an integrative account of internal and external selection processes. *Psychon. Bull. & Rev.* **20**, 228–242 (2013).
52. R Ratcliff, MJ Frank, Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Comput.* **24**, 1186–1229 (2012).
53. ML Pedersen, MJ Frank, G Biele, The drift diffusion model as the choice rule in reinforcement learning. *Psychon. Bull. & review* **24**, 1234–1251 (2017).
54. MJ Frank, et al., fmri and eeg predictors of dynamic decision parameters during human reinforcement learning. *J. Neurosci.* **35**, 485–494 (2015).
55. E Schulz, E Konstantinidis, M Speekenbrink, Putting bandits into context: How function learning supports decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 927–943 (2018).
56. NJ Mackintosh, A theory of attention: variations in the associability of stimuli with reinforcement. *Psychol. Rev.* **82**, 276 (1975).
57. JM Pearce, NJ Mackintosh, Two theories of attention: A review and a possible integration in *Attention and associative learning: From brain to behaviour*, eds. C Mitchell, M LePelley. (Oxford University Press), pp. 11–39 (2010).
58. AJ Yu, P Dayan, Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
59. Y Niv, et al., Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).
60. D Marković, J Gläscher, P Bossaerts, J O'Doherty, SJ Kiebel, Modeling the evolution of beliefs using an attentional focus mechanism. *PLoS computational biology* **11**, e1004558 (2015).
61. NDB Bruce, JK Tsotsos, Saliency, attention, and visual search: An information theoretic approach. *J. Vis.* **9**, 1–24 (2009).
62. JM Wolfe, Visual search. *Curr. Biol.* **20**, R346–R349 (2010).
63. N Sprague, D Ballard, Eye movements for reward maximization in *Advances in neural information processing systems*. (Neural Information Processing Systems Foundation, Inc.), pp. 1467–1474 (2004).
64. BT Sullivan, L Johnson, CA Rothkopf, D Ballard, M Hayhoe, The role of uncertainty and reward on eye movements in a virtual driving task. *J. Vis.* **12**, 1–17 (2012).
65. S Manohar, M Husain, Attention as foraging for information and value. *Front. Hum. Neurosci.* **7**, 711 (2013).
66. H Stojic, JL Orquin, P Dayan, R Dolan, M Speekenbrink, Project files for "Uncertainty in learning, choice and visual fixation." (2019) Open Science Framework. Available at https://osf.io/539ps. November 2019.

PNAS
www.pnas.org

Supplementary Information Appendix for

Uncertainty in learning, choice and visual fixation

H. Stojić, J.L. Orquin, P. Dayan, R. Dolan and M. Speekenbrink

Hrvoje Stojić.
E-mail: h.stojic@ucl.ac.uk

This PDF file includes:

Supporting Information Text

SI Methods

Exclusions. We recruited 34 participants (18 female, $M_{\text{age}} = 26.8$ and $SD_{\text{age}} = 8.1$) in total. We removed data from two participants before processing, one due to problems with eye tracker calibration throughout the experiment, and one due to constant movement during the experiment. We applied two *a priori* exclusion criteria separately to each game played by each participant. The first criterion was failing to respond in time in more than 10 trials in a game. We excluded one participant who had more than 10 such trials in both games, most likely due to issues with calibration, as the choices were gaze contingent. The second criterion was failing to exceed chance performance (mean choice rank of 3.5) in the last 15 trials of a game. Specifically, we excluded games in which there was not at least weak evidence of above chance performance, as evidenced by the Bayes factor. In particular, we excluded games for which $BF_{01} > 1$ (more evidence for the null hypothesis of chance performance). Based on these two criteria we excluded 28 games, 13 Decreasing variances games and 15 V-shaped variances games. After these exclusions, we were left with 23 participants (12 female, $M_{\text{age}} = 26.9$ and $SD_{\text{age}} = 8.4$), 36 games in total, 19 Decreasing variances and 17 V-shaped variances games.

Trial time course. The trial structure was as follows. The inter-trial interval (ITI), indicated by a fixation cross, was randomly drawn from a uniform distribution, $U(5, 5.5)$. After this interval, the fixation cross would rotate, indicating to participants that they should fixate on the cross in order to trigger the presentation of the stimuli. They had to maintain their fixation on the cross for 1 s for this to happen. Six options faded in over the course of 0.5 s and were presented for a total of 5 s fixed time, during which participants would consider which option to choose. Stimuli would then smoothly fade away over the course of 0.5 s, and the fixation cross would reappear, this time with a normal (unrotated) orientation. Participants would again have to fixate on the cross for 1 s to trigger the onset of the execution stage, in which they would make their choice. All six options appeared at their own location, smoothly fading in over 0.5 s. To register their choice, participants had to fixate on the option they wanted to choose and simultaneously press the SPACE key on the keyboard. They had at most 1.5 s to make a choice. Unchosen options would smoothly fade away over 0.5 s while the chosen option would remain on screen for a random duration drawn from a uniform distribution, $U(2, 2.5)$. In the following outcome stage, the chosen option was made slightly transparent and reward feedback was overlaid on top of it for a fixed duration of 3 s. This was followed by the random inter-trial interval, and then the start of the next trial.

Stimuli construction in games. Rewards were drawn from a Gaussian distribution with an option specific mean and variance, μ_k and σ_k^2. Means and variances differed between the options in each game. The best option in both games had a mean reward of 6. The main difference between the games was in the variance of the rewards. In both games the variance of the best option was set to 2.75. In the *Decreasing variances* game, the variance of each option decreased from the best option to the worst (according to the mean value) by 0.4. In the *V-shaped variances* game the variance decreased from the best option to the third best by 0.4, and then increased again from the fourth best to the worst option by 0.4. To make the difficulty of both games approximately equal, we set $d^{'} = 0.4$ between option pairs adjacent in their rank, and then determined exact means of the second best to the worst option according to $\mu_j = \mu_{j-1} - 0.4 \times \sqrt{(\sigma_{j-1}^2 + \sigma_j)/2}$, where j is the rank of the option starting from 1.

To minimize carry-over effects between the games, we used a different set of letters from Gljagolica alphabet (1, letter area and color was adapted) and presented rewards using two different currencies (kuna and lek), with exchange rates of either 10 or 40, determined so that average earnings in both games are approximately equal. Rewards were scaled through multiplication with these exchange rates. Before each game began we informed participants about the exchange rate that would be used in the game and according to which we converted the points earned to money at the end of the experiment. The letters used in the games, options' locations, the order of the games, and the currencies and scaling factors associated with each game were fully randomized.

Procedure. Upon entering the laboratory, participants completed an informed consent form and provided basic sociodemographic data. Next we tested participants for eye dominance and seated them in front of the eye tracker. We then presented them with instructions about the task and earnings on-screen. We explained they would play two games and in each game they had to repeatedly (60 times) choose between six options, receiving a reward after each choice, with the goal to maximise the sum of rewards earned. We also explained that while the rewards were noisy, the average reward of the options would not change over time. We also indicated that the locations were chosen randomly and that nothing in the spatial arrangement was predictive of the options' values. Finally, we explained in detail how their earnings in the experiment were related to their choices in the games.

Participants completed seven practice trials before starting the games, in order to familiarise themselves with the interface, timings, and how to make gaze contingent transitions between the stages and choices. In the first two trials we showed brief instructions at the different stages, explaining what to do and how to perform actions. We increased the duration of each stage in these trials, to allow for sufficient time for reading the instructions. We used a different set Gljagolica letters in these practice trials.

Throughout the games, including the practice trials as well, if participants failed to respond in time (1.5 s), the trial was repeated. To provide an incentive to respond in time we deducted a significant number of experimental points when they failed to respond in time, in the amount of the expected value of the highest ranking option.

On finishing each game we provided feedback to participants how many experimental points they accumulated and to what earnings would that correspond if the game were to be selected to be paid out. After participants finished both games, we informed them about the game randomly selected for the payout and their final earnings, debriefed them, and paid out their earnings.

For the sake of full transparency, we recorded the following variables in our experiment: participants' choices, response times, eye gaze locations, and pupillary responses in the MAB tasks, and basic socio-demographic data (age and gender).

Eye tracking. Eye movements and pupillary responses were recorded using a desk-mounted EyeLink 1000 eye tracker (SR Research, https://www.sr-research.com/) with a monocular sampling rate of 500 Hz, a screen resolution of 1650×1050 pixels and physical size of 475×297 mm (widths and heights, respectively). We recorded pupil area using the centroid fit algorithm and CR tracking mode. The screen subtended a visual angle of 46.5°horizontally and 30.1°vertically. Participants used a chin-rest at approximately 60 cm viewing distance from the screen. We recorded exact physical layout of the equipment following (2).

Before beginning with the practice trials and each of the games, we performed calibration with the dominant eye using a 13-point calibration procedure, followed by a 13-point drift validation test. Background colour and calibration point colours were adjusted according to the rest of the experiment. We considered acceptable a calibration offset less than 1.0°of visual angle. Pupil tracking parameters were determined through EyeLink's automatic method. We repeated the calibration procedure within the game if participants reported difficulties in gaze contingent stages of the trial.

We presented the experimental stimuli using the PsychoPy library (3). In the MAB task each option was represented by a circle with a radius of 3 cm and a letter with 2×2 cm size centred in it. To prevent participants from foveating more than one option at a time, they were placed in a circle separated horizontally and vertically by at least 3°of visual angle, 9 cm from the centre of the screen. In gaze contingent stages of the trial – triggering the onset of the choice and execution stage – participants had to have 90% of gaze locations within a 1 s window in a circle area with a 3 cm radius around the fixation cross. To make a response in the execution stage participants had to press a key and an eye data sample had to be recorded at the same time within a circle representing an option. Reward feedback was presented overlaid over the chosen option as text, with letters of 1 cm in height.

Several aspects of the task implementation were concerned with collecting good quality pupillary data. We presented everything on the screen using two isoluminant colours – #457CA9 as a background colour and #A4694F for text and stimuli (HTML colour code). We chose letters from the alphabet that have approximately the same area. We used a longer inter-trial interval (5.25 s on average) to allow for pupil size to return to the baseline. Before each game we collected the game pupil baseline in a 40 s long stage with a fixation cross only where we instructed participants to look at the fixation cross during that time. They had another 40 s long stage where they looked directly at a camera, to be able to compute the pupil area in non-arbitrary physical units (2). Finally, the experiment was conducted in a darkly lit room, with constant lighting conditions across participants.

We used the default velocity, acceleration and motion-based algorithm provided by SR Research to detect fixations (4). In data analysis we drew an area of interest (AOI) with a radius of 3cm around the centre of every option and assigned all fixations falling into these AOIs to the corresponding option.

Data analysis.

Mixed effect regressions. We examined learning effects within games and differences between game types using Bayesian mixed effect regressions, implemented in the `brms` package in R (5, 6). We computed averages across blocks and regressed the intercept, a block indicator variable (coded as $[-1.5, -0.5, 0.5, 1.5]$ for blocks 1 to 4) and a game type indicator variable (coded as -1 for Decreasing variances game and 1 for V-shaped variances game), as well as their interaction, on choice performance (chosen option rank) and various fixation statistics in the choice and outcome stage (total fixation duration, number fixations and number of options fixated). We modelled choice ranks as a normal distribution truncated to the $[1, 6]$ interval. The total number of fixations and the number of options fixated on were also modelled as normal distributions, truncated to the $[0, \infty)$ interval. Total fixation duration was modelled as a skew normal distribution, truncated to the $[0, \infty)$ interval. Intercept and blocks were entered as game-specific random effects, while game type was entered as a fixed effect. `brms` package uses the No-U-Turn-Sampling MCMC algorithm implemented in Stan (7) to fit the models (see "Fitting using MCMC" in *SI Methods*). We used default `brms` (version 2.8.0) priors and a centred parametrization of group-level parameters.

Modelling learning and choices. We fitted four main computational models to participants' choices. All four consisted of a learning component and a choice component. The learning component was either a Kalman filter (KF) (8–10) or a "lazy" Kalman filter (KFL) model. Both use a form of the delta rule (11) to update estimated value based on a reward prediction error. What makes Kalman filter models different is that they track the (posterior) variance of the estimated value of each option (i.e. estimation uncertainty) and use this to dynamically adjust the learning rate. The lazy Kalman filter introduces a bias to the learning rate, allowing for slower learning. For the choice component the models used either a Softmax (SM; 12), or an upper confidence bound choice rule (UCB; 13). In the Softmax model exploration happens randomly – participants choose options with probability roughly proportional to the differences in estimated value between the options. By contrast, the UCB choice rule uses estimation uncertainty to approximate the information gained by choosing an option, and adds this as an "uncertainty bonus" to the estimated values (14), making exploration driven by information gain. We use a probabilistic form of the UCB rule where values are passed through a Softmax function, in contrast to the original deterministic form (13).

In the Kalman filter model we assumed participants update their estimates $E_j(t+1)$ of the expected reward of choosing option j on trial $t+1$ from the observed reward $R_j(t)$ on trial t as

$$E_j(t+1) = E_j(t) + I_j(t)K_j(t)[R_j(t) - E_j(t)]. \quad [1]$$

where the so-called "Kalman gain" term $K_j(t)$ acts as a learning rate. Term $I_j(t)$ is a simple indicator variable, with a value of 1 if option j is chosen on trial t and 0 otherwise. The Kalman gain is updated on every trial and depends on current level of uncertainty. This makes it a dynamic learning rate

$$K_j(t) = \eta \frac{S_j(t) + \sigma_\zeta^2}{S_j(t) + \sigma_\zeta^2 + \sigma_{\epsilon,j}^2}, \quad [2]$$

where $S_j(t)$ is the variance of the posterior distribution of the mean reward, updated in every trial as $S_j(t+1) = [1 - I_j(t)K_j(t)][S_j(t) + \sigma_\zeta^2]$. Parameters σ_ζ^2 and $\sigma_{\epsilon,j}^2$ are the innovation variance and option's reward variance respectively, which modulate the learning rate. Parameter $\eta \in (0, 1)$ determines the bias in updates of the Kalman gain, causing it to potentially learn at slower pace (hence the term "lazy"). In the standard Kalman filter we fixed this parameter to $\eta = 1$, while in lazy versions it is a free parameter, thus allowing for imperfect updates. In both variants we initialized estimate of the expected value to $E_j(0) = 0$. Initial variance was a free parameter σ_i^2 such that $S_j(0) = \sigma_i^2$. We take into account differences between variances of options by setting the $\sigma_{\epsilon,j}^2$ parameter to option's objective variance that we used to draw rewards from: $[2.75, 2.35, 1.95, 1.55, 1.15, 0.75]$ in Decreasing variances and $[2.75, 2.35, 1.95, 1.95, 2.35, 2.75]$ in V-shaped variances game. This imposes different learning rates for each option, influenced by its objective reward distribution variance. To estimate the Kalman filter learning model one of the three parameters (σ_ζ^2, σ_ϵ^2 and σ_i^2) needs to be fixed and we found that fixing σ_ϵ^2 results in more stable estimations with better convergence properties (see "Hierarchical Bayesian parameter estimation" in *SI Methods*).

We consider two choice rules that describe how the estimated values are used to make a choice $C(t)$ between the options. In the Softmax choice rule (12) exploration occurs by chance – participants choose probabilistically according to relative estimated value

$$P(C(t) = j) = \frac{\exp[\theta E_j(t)]}{\sum_{k=1}^{6} \exp[\theta E_k(t)]}, \quad [3]$$

where $P(C(t) = j)$ is the probability of choosing option j on trial t and the inverse temperature parameter $\theta > 0$ determines the sensitivity to differences in estimated values, and with it the amount of exploration.

The upper confidence bound choice rule (13) uses estimation uncertainty to approximate an option's informativeness, or how much value estimates can be improved by trying an option. A multiple of the estimation uncertainty, defined as the standard deviation of the posterior distribution of the mean reward, is added to the posterior mean reward as an "exploration bonus". While the original UCB rule chooses the option with the highest resulting value deterministically, we implemented a stochastic version of the UCB rule by using a softmax transformation. In the Kalman filter the estimation uncertainty is explicitly modelled (posterior variance S in Eq. 2), resulting in the following form of the UCB choice rule

$$P(C(t) = j) = \frac{\exp\{\theta(E_j(t) + \beta\sqrt{S_j(t)})\}}{\sum_{k=1}^{6} \exp\{\theta(E_k(t) + \beta\sqrt{S_k(t)})\}}, \quad [4]$$

where $\beta > 0$ is the weight a participant places on estimation uncertainty. We fitted all models using hierarchical Bayesian parameter estimation (see "Hierarchical Bayesian parameter estimation" in *SI Methods*).

Modelling visual fixation in the choice stage. We used trial-by-trial subjective estimates of value and uncertainty from the KFL-UCB model fitting choices best and regressed them on relative fixations in the choice stage. We controlled for potential differences between the games by including a game type indicator. Relative fixations were operationalised as a sum of durations of fixations on each of the options in the choice stage, normalized by dividing by the sum of these quantities across all options.

We assume that relative fixations in the choice stage (RF) follow a Dirichlet distribution

$$\mathrm{RF}(t) \sim D(\boldsymbol{\alpha}(t), \kappa), \quad [5]$$

with the probability density function defined as

$$\frac{1}{\mathrm{B}(\boldsymbol{\alpha}(t)\kappa)} \prod_{j=1}^{6} \mathrm{RF}_j^{\alpha_j(t)\kappa - 1}, \quad [6]$$

where $\mathrm{B}(\boldsymbol{\alpha}(t)\kappa)$ is a multinomial beta function that acts as a normalising constant. The vector of concentration parameters $\boldsymbol{\alpha}(t)$ for each trial is obtained by passing values ($E_j(t)$) and estimation uncertainty ($S_j(t)$) of each option j obtained from the KFL-UCB model, as well as the game-type indicator (G) as a control variable, through a Softmax function

$$\boldsymbol{\alpha}(t) = \frac{\exp\{\beta_v E_j(t) + \beta_u \log S_j(t) + \beta_g G\}}{\sum_{k=1}^{6} \exp\{\beta_v E_k(t) + \beta_u \log S_k(t) + \beta_{gt} G\}}, \quad [7]$$

where β_v and β_u are weights on value and uncertainty, while β_{gt} is the effect of game type. We log-transformed estimation uncertainty to linearise it. Games were coded as $G = -1$ for the Decreasing variances and $G = 1$ for the V-shaped variances game and this effect was included at a group level only. We assumed an additional precision parameter κ that multiplies the concentration parameters, governing how much probability mass is near the expected value. We tested also several reduced models where either uncertainty or value was left out. We fitted all models using hierarchical Bayesian parameter estimation (see "Hierarchical Bayesian parameter estimation" in *SI Methods*).

Modelling visual fixation in the outcome stage. We used trial-by-trial uncertainty, reward prediction errors, and value from the KFL-UCB model fitting choices best and regressed them on absolute fixations in the outcome stage. We controlled for potential differences between games by including a game-type indicator. Absolute fixation was operationalised as the sum of durations of all fixations on the reward feedback.

We assumed fixation durations during outcome stage (F) follow a Skew Normal distribution

$$F(t) \sim N(\xi(t), \omega, \alpha), \quad [8]$$

truncated to the interval $F(t) \in [0, 5]$. In the full model the location parameter $\xi(t)$ for each trial is a linear combination of intercept, uncertainty ($S_j(t)$), reward prediction error (PE), unsigned reward prediction error (uPE), and value ($E_j(t)$) of chosen option j obtained from the KFL-UCB model and the game-type indicator variable (G)

$$\xi(t) = \beta_i + \beta_u \log S_j(t) + \beta_{\text{PE}}\text{PE}_j(t) + \beta_{\text{uPE}}|\text{PE}|_j(t) + \beta_v E_j(t) + \beta_{gt} G, \quad [9]$$

where β_u, β_{PE}, β_{uPE} and β_v are weights on uncertainty, prediction errors, unsigned prediction errors and value, while β_i is the intercept parameter and β_{gt} the game-type effect. The intercept parameter is the baseline or mean absolute fixation across the whole experiment, while other parameters act as deviations from the baseline. We computed unsigned prediction errors as absolute value of prediction errors (this worked better than squaring the prediction error) and log-transformed estimation uncertainty to linearise it. Games were coded as $G = -1$ for the Decreasing variances and $G = 1$ for the V-shaped variances game and this effect was included at a group level only. We assumed an additional scale parameter ω and shape parameter α, modelled at an individual game level, without a group-wise parameter. We tested several reduced models where uncertainty, one of the reward prediction errors, or value are left out. We fitted all models using hierarchical Bayesian parameter estimation (see "Hierarchical Bayesian parameter estimation" in *SI Methods*).

Modelling choices with visual fixations alone. We can model choices by relative fixations in the choice stage alone by regressing the latter onto the former, without explicitly modelling the learning and choice process. We used a simple multinomial logistic regression model where relative fixation for option j in trial t, $\text{RF}_j(t)$, is passed through a Softmax function to obtain the probability $P(C(t) = j)$ of choosing option j at trial t

$$P(C(t) = j) = \frac{\exp[\tau \text{RF}_j(t)]}{\sum_{k=1}^{6} \exp[\tau \text{RF}_k(t)]}, \quad [10]$$

where the inverse temperature parameter $\tau > 0$ determines the sensitivity to differences in relative fixations.

To avoid our relative fixation measure taking the value 0 for options that were not fixated on at all in certain trials, we assigned each option a minimum value ϵ which was treated as a free parameter:

$$\text{RF}_j(t) = \epsilon/6 + (1 - \epsilon)\frac{F_j(t)}{\sum_{k=1}^{6} F_k(t)}. \quad [11]$$

Estimating the ϵ parameter can tell us how useful the fixation data is. Overall, this regression model has two parameters: θ and ϵ. We fitted the model using hierarchical Bayesian parameter estimation (see "Hierarchical bayesian parameter estimation" in *SI Methods*).

Modelling learning and choices modulated by visual fixations. We assumed visual fixations can modulate the choice or learning component of the KFL-UCB model. We mark the learning and choice component with an "a" prefix to indicate which aspect is modulated by fixations. For example, in the aKFL-UCB model, visual fixations modulate the learning process, while in the KFL-aUCB they modulate the choice process.

We assumed visual fixations in the choice stage enter the choice process by re-weighting the choice probabilities produced by the models based on options' estimated values and estimation uncertainty (Eq. 4). The relative fixation measure defined in Eq. 11 enters the UCB rule in an additive way:

$$P(C(t) = j) = \frac{\exp\{\tau \text{RF}_j(t) + \theta(E_j(t) + \beta\sqrt{S_j(t)})\}}{\sum_{k=1}^{6} \exp\{\tau \text{RF}_k(t) + \theta(E_k(t) + \beta\sqrt{S_k(t)})\}}. \quad [12]$$

This has the effect of increasing the choice probabilities for options that received relatively more fixation time and diminishing them for the options that received relatively little fixation time.

We assumed visual fixations in the outcome stage influence the learning process by making the bias in the Kalman gain update dependent on how long the reward feedback was fixated on during the outcome stage of a trial. We implemented this

by replacing the η parameter in Eq. 2 with a baseline parameter η_0 and a slope parameter η_1 that depends on F, the absolute fixation duration in outcome stage:

$$\eta(t) = \Phi(\eta_0 + \eta_1 F(t)), \quad [13]$$

where Φ is the standard normal cumulative distribution function, which we use to constrain the resulting η parameter to the $(0, 1)$ range.

Overall, there are three model variants with attention modulation: The aKFL-UCB model has all the parameters that the KFL-UCB model has and an additional η_1 parameter. The KFL-aUCB model instead has additional ϵ parameter, while the aKFL-aUCB has both additional parameters. We fitted all models using hierarchical Bayesian parameter estimation (see "Hierarchical bayesian parameter estimation" in *SI Methods*).

Hierarchical bayesian parameter estimation. We used a Bayesian hierarchical estimation procedure to estimate the parameters of each model (15). The unit of analysis was a game, rather than a participant, as games differed in their design and some participants had only one game left in the dataset after exclusions were performed. We used hierarchical models which treat each game as drawn from a common group-level distribution, where parameters at the game level are assumed to be generated by the same group-level prior distribution. We used this approach for all parameters in our models unless explicitly stated otherwise (e.g. game type variable). Parameters at the game level and so-called hyperparameters at the group level mutually constrain each other and we estimate them jointly using a Markov Chain Monte Carlo (MCMC) sampling procedure. We formulated all game-level parameters using a non-centred (probit) parametrization which facilitates MCMC sampling with hierarchical models (16).

We sample hyperparameters from hyperprior distributions, for which we used moderately informative distributions but broad enough to allow data to shift them. We show priors and hyperpriors for each parameter in Table S1 and Table S2. We use superscript g on each parameter to refer to the dependence of parameters on game. For example, we assume that the uncertainty guidance parameter from UCB choice rule, β^g, is sampled from a prior which is a linear combination of three hyperparameters: μ_β determines the mean of the prior, while ζ_β and ν_β together determine deviations from the mean (this is the non-centered parametrization). This particular parameter should be non-negative and we use an exponential transformation to ensure that. The hyperparameters are in turn sampled from their hyperpriors: μ_β from a Normal distribution with mean 0 and standard deviation 1, ζ_β from a Half-Normal distribution with mean 0 and standard deviation 1 (half refers to truncation to the $[0, \infty)$ interval), and ν_β from a Normal distribution with mean 0 and standard deviation 1. Credible intervals were computed as highest posterior density intervals.

Fitting using MCMC. We fitted the models to the data using the No-U-Turn-Sampling MCMC algorithm implemented in Stan (7). This algorithm approximates the posterior distribution of parameters by generating samples from this posterior distribution given the observed behavioural data. We initialized five independent chains with randomly generated starting values and collected 5000 samples of each chain at a thin rate of 1, after discarding the first 5000 of burn-in samples of each chain. We confirmed that all chains successfully converged by visually inspecting the traceplots of the chains and examining the number of efficient samples and $\hat{R}$ statistic.

Model comparison with bridge sampling. We performed model comparison by estimating the model evidence through bridge sampling (17, 18), using the *bridgesampling* package in R (19). Bridge sampling uses samples from MCMC chains to estimate the log marginal likelihood of a model, which can then be used to compute posterior probabilities of a model being the true one among a set of models (18). Following recommendations in (19) we used the "warp3" method and repeated the estimation multiple times ($N = 50$) to obtain an empirical estimate of error in estimated model evidence (the interquartile range of the estimates). In the figures that illustrate model evidence we report median log marginal likelihoods and posterior probabilities across the repetitions for each model.

SI Results

Additional properties of visual fixation.

In the choice stage, we examined several other measures of visual fixation: number of fixations made across all options and number of distinct options that were fixated. Mean number of fixations decreased over time and there was no difference between games (mixed effects regression estimates: intercept = 9.26, 95% CI $[8.57, 9.95]$; block = -0.52, 95% CI $[-0.78, -0.26]$; game = 0.10, 95% CI $[-0.60, 0.80]$; block×game = -0.24, 95% CI $[-0.51, 0.02]$). Mean number of unique option fixations decreased over time and there was a weak difference between games (mixed effects regression estimates: intercept = 2.99, 95% CI $[2.67, 3.31]$; block = -0.15, 95% CI $[-0.25, -0.04]$; game = 0.06, 95% CI $[-0.26, 0.38]$; block×game = -0.14, 95% CI $[-0.25, -0.04]$), as evidenced by block and block-game interaction estimates.

In the outcome stage, besides assessing trial-by-trial variability in absolute fixation we also examined variability in the number of fixations. The mean number of fixations decreased over time, same as for absolute fixation, but here there was no difference between games (mixed effects regression estimates: intercept = 4.76, 95% CI $[4.40, 5.11]$; block = -0.17, 95% CI $[-0.30, -0.05]$; game = -0.12, 95% CI $[-0.49, 0.23]$; block×game = 0.08, 95% CI $[-0.05, 0.20]$). The negative effect of block suggests that the decrease in number of fixations, but not fixation durations, is likely driving the decrease in absolute fixation. The standard deviation of number of fixations is sizeable (mixed effects regression estimates: intercept = 1.97, 95% CI

[1.87, 2.08]; block = 0.04, 95% CI [−0.02, 0.10]; game = 0, 95% CI [−0.10, 0.10]; block×game = −0.01, 95% CI [−0.07, 0.05]), as evidenced by intercept estimate.

As an additional check of differences between games in terms of visual fixations, we regressed out expected values of options in each game from relative fixation time in the choice stage using Dirichlet regressions. Residuals from such regressions should reveal potential differences between games due to differences in the pattern of variances. The results however show no difference between standardised residuals of the two games (Fig. S1). This provides additional evidence that the differences between the games, relating to the variance of the lower ranking options only, were too subtle to result in large differences.

Finally, we investigated some relations between aspects of the fixation process and choices. In particular, we were interested in checking the predictions of how visual fixations influence choice made by (20), based on their attentional drift diffusion modelling approach. Even though we could not directly apply this approach to our paradigm because we fixed the duration of the choice stage, we examined to what extent their predictions are born out by our data. First, one assumption (20) made was that first fixation is unbiased by options' values. In our learning task, participants learn the value of options over time and since the locations of options do not change across the trials we expected this assumption to be violated. Indeed, we found a larger probability of first fixating the best option than expected by chance (mean of 0.31, $SE = 0.04$; Fig. S2B). Another of (20) predictions was that final fixations would be shorter than middle fixations, as fixations are interrupted when the evidence accumulation process hits the bound. In our task there was no free response and we did not expect to see that pattern. Indeed, examining fixation duration by fixation type – first, one of the middle, or last fixation – shows the opposite pattern, where the last fixation duration is longer than the middle ones (Fig. S2A). A final prediction made by (20) which we assessed is that participants should choose the option they looked at last, unless the option is much worse. This prediction did hold in our data as well. Participants indeed increasingly chose the option fixated the last, unless the option was low ranking one (Fig. S3).

Modelling learning and choices – control models.

Rescorla-Wagner and Choice kernel based models. We fitted four additional control models to the choices. These models also consisted of a learning component and a choice component. We used three types of learning models: a Choice kernel (CK) model (21), a Rescorla-Wagner (RW) model (11), and a combination of both (21). These models do not explicitly track estimation uncertainty, but have been often successfully used to model learning in tasks like ours. With these learning models we used two choice models: a Softmax choice model (SM; 12) that we used with all three types of the learning models, and a nonparametric version of an upper confidence bound choice model (UCB; 13) that we used only with the RW model.

The CK model assumes that participants estimate a so-called choice kernel, $\mathrm{CK}_j(t)$, which keeps track of how frequently they have chosen option j in the recent past. This choice kernel is updated as

$$\mathrm{CK}_j(t+1) = \mathrm{CK}_j(t) + \gamma[I_j(t) - \mathrm{CK}_j(t)], \quad [14]$$

where $\gamma \in (0, 1)$ is a fixed learning rate parameter and $I_j(t) = 1$ if option j was chosen on trial t, and 0 otherwise. We initialized estimates of the choice kernel to $\mathrm{CK}_j(0) = 0$.

The RW model assumes participants update their estimates $E_j(t+1)$ of the expected value of choosing option j on trial $t+1$ from the reward $R_j(t)$ on trial t

$$E_j(t+1) = E_j(t) + I_j(t)\alpha[R_j(t) - E_j(t)], \quad [15]$$

where $\alpha \in (0, 1)$ is a fixed learning rate parameter. We initialized estimates of mean values to $E_j(0) = 0$.

The Softmax choice rule uses the estimated kernels or values to make a choice $C(t)$ between the options. In this choice rule exploration occurs by chance – participants choose probabilistically according to relative estimated kernel or value. The CK-SM model used

$$P(C(t) = j) = \frac{\exp[\tau \mathrm{CK}_j(t)]}{\sum_{k=1}^{6} \exp[\tau \mathrm{CK}_k(t)]}, \quad [16]$$

where the parameter $\tau > 0$ determines the sensitivity to differences in estimated kernels, and with it amount of exploration. The model thus had two parameters: τ and γ. The RW-SM model used

$$P(C(t) = j) = \frac{\exp[\theta E_j(t)]}{\sum_{k=1}^{6} \exp[\theta E_k(t)]}, \quad [17]$$

with an inverse temperature parameter $\theta > 0$ in addition to α. Finally, the RWCK-SM model used

$$P(C(t) = j) = \frac{\exp[\theta E_j(t) + \tau \mathrm{CK}_j(t)]}{\sum_{k=1}^{6} \exp[\theta E_k(t) + \tau \mathrm{CK}_j(t)]}, \quad [18]$$

with all four parameters from CK-SM and RW-SM model.

The last model was the RW-UCB model. The UCB choice rule uses estimation uncertainty to approximate an option's informativeness, or how much value estimates can be improved by trying an option. While Kalman filter models explicitly track estimation uncertainty, the RW model does not. Therefore we use a nonparametric form of the UCB, using current trial t and number of times the options were chosen N_j as proxies

$$P(C(t) = j) = \frac{\exp\{\theta(E_j(t) + \beta\sqrt{\log(t)/N_j})\}}{\sum_{k=1}^{6} exp\{\theta(E_k(t) + \beta\sqrt{\log(t)/N_j})\}}, \quad [19]$$

where $\beta > 0$ is the weight a participant places on estimation uncertainty.

We fitted these models to the choice data using hierarchical Bayesian model estimation (see "Modelling choices with visual fixations alone" in *SI Methods*). Model evidence shows that the models consisting of Kalman filter learning and UCB choice rule fit the data better than those based on Rescorla-Wagner, choice kernel learning, or both (Fig. S4). The lazy Kalman filter model with a UCB choice rule described participants' choices best (KFL-UCB), with a posterior probability of approximately 0.99. All other models have a posterior probability of approximately zero. The Rescorla-Wagner UCB model (RW-UCB) performed particularly poorly, indicating that a nonparametric form of the UCB choice rule based on the number of times an option has been chosen does not describe behaviour well.

KFL-UCB model with unconstrained β parameter. The only difference between the KFL-UCB model that fitted the choices the best and its Softmax counterpart, KFL-SM, is the β parameter, that acts as a weight on uncertainty in the UCB choice rule. The strong evidence favouring the KFL-UCB model over the KFL-SM model indicates that the β parameter is reliably different from zero. In estimating the β parameter in the KFL-UCB model, we assumed that it can not be negative. As an additional check, we here also estimate an unconstrained model where β can be positive or negative. While the notion of directed exploration rests on a positive value of β, it is possible that participants are averse to irreducible uncertainty and negative values of the β parameter can capture such uncertainty or risk aversion (22). This makes interpretation of a negative value of β complicated, as the parameter may be negative when there is both directed exploration *and* risk aversion.

We fitted the unconstrained KFL-UCB model using hierarchical Bayesian parameter estimation (see "Hierarchical bayesian parameter estimation" in *SI Methods*). Comparing only the two models of interest, the KFL-UCB model with the non-negative β parameter outperformed the unconstrained KFL-UCB model with a posterior probability of approximately 0.99. Moreover, the β parameter of the unconstrained model was overwhelmingly positive (posterior mean of 0.48, 95% CI $[0.21, 0.76]$). This result further affirms that the β parameter is positive. In addition, we also compared the model to all other choice models (KFL-UCB unc; Fig. S4). The unconstrained KFL-UCB model is a second best model, after the constrained KFL-UCB model. These two models, together with KF-UCB, outperformed all others by a large margin.

Choice models with unsigned prediction errors. The analyses of interactions between the learning and fixation process showed that unsigned prediction error was the strongest predictor of absolute fixations in the outcome stage of the trial. This provided evidence for a theory-driven expectation that time spent looking at the reward feedback is guided by uncertainty (23). This result also suggests that unsigned prediction errors could reflect a more important form of uncertainty for guiding choices than estimation uncertainty. Hence, a model that uses unsigned prediction errors instead of estimation uncertainty in the UCB choice rule could potentially explain choices better than the currently best-fitting KFL-UCB model. Here we investigate this further.

In the analysis of fixation data we regressed unsigned prediction error from *the current trial* on the total fixation duration in the outcome stage of the trial. For predicting choices, the prediction error from the previous trial would not be sensible as uncertainty would not be defined for those options which were not chosen on the previous trial. Instead, it is reasonable to maintain estimates of uncertainty based on unsigned prediction errors that are updated from trial to trial.

Based on this idea we implemented two models. The first model is a KFL-UPE model that uses a simple delta-rule to learn slow-moving estimates of unsigned prediction errors coming from the lazy Kalman filter learning model. Hence, besides the usual lazy Kalman filter learning (Eq. 1 and 2), the KFL-UPE model assumes participants update their estimates of unsigned prediction errors $U_j(t+1)$ based on chosen option j on trial $t+1$ from the reward $R_j(t)$ on trial t and estimated value $E_j(t)$ (from Eq. 1)

$$U_j(t+1) = U_j(t) + I_j(t)\zeta|R_j(t) - E_j(t)|, \quad [20]$$

where $\zeta \in (0, 1)$ is a fixed learning rate parameter. We initialized estimates to $U_j(0) = 100$. Term $I_j(t)$ is an indicator variable, with value of 1 if option j is chosen on trial t and 0 otherwise.

These estimates were then used in a UCB-like choice rule where instead of estimation uncertainty we used estimates of unsigned prediction errors, $U_j(t)$. The probability of choosing option j at trial t is given by

$$P(C(t) = j) = \frac{\exp\{\theta(E_j(t) + \beta\sqrt{U_j(t)})\}}{\sum_{k=1}^{6}\exp\{\theta(E_k(t) + \beta\sqrt{U_k(t)})\}}, \quad [21]$$

where $\beta > 0$ is the weight a participant places on uncertainty and the inverse temperature parameter $\theta > 0$ determines the sensitivity to differences in the values.

The second model is a K2-UPE model that uses the K2 learning model which computes estimates of unsigned prediction errors in a more principled manner, following (24). In the K2 model these uncertainty estimates are used to dynamically modulate the learning rate, making the K2 model an alternative to the Kalman learning. The K2 model assumes participants update their estimates $E_j(t+1)$ of the expected reward of choosing option j on trial $t+1$ from the observed reward $R_j(t)$ on trial t as

$$E_j(t+1) = E_j(t) + I_j(t)K_j(t)[R_j(t) - E_j(t)]. \quad [22]$$

where term $K_j(t)$ is a dynamic learning rate, similar to the Kalman filter learning model. Term $I_j(t)$ is an indicator variable, with value of 1 if option j is chosen on trial t and 0 otherwise. The learning rate is updated on each trial as

$$K_j(t) = \frac{S_j(t)}{S_j(t) + \hat{R}_j}, \quad [23]$$

where $S_j(t)$ is the uncertainty estimate. The K2 algorithm adapts $S(t)$ by performing gradient descent in a corresponding set of parameters $b(t)$, where the two are related by

$$S_j(t) = \exp\{b_j(t+1)\}. \quad [24]$$

The parameters $b(t)$ are updated as

$$b_j(t+1) = b_j(t) + \nu I_j(t)[(R_j(t) - E_j(t))^2 - \hat{R}_j - \sum_{k=1}^{6} S_k(t)], \quad [25]$$

where $\nu > 0$ is a step-size parameter and $b(t)$ parameters are initialized as

$$b_j(0) = \log \hat{R}_j. \quad [26]$$

Finally, $\hat{R}$ is an estimate of irreducible uncertainty. Here we used the objective reward variances, $[2.75, 2.35, 1.95, 1.55, 1.15, 0.75]$ in the Decreasing variances and $[2.75, 2.35, 1.95, 1.95, 2.35, 2.75]$ in the V-shaped variances game, rescaled by a free parameter σ_ϵ^2.

We combined the K2 learning model with a UCB-like choice rule where instead of estimation uncertainty we used uncertainty estimates ($S_j(t)$) from the K2 learning model. The probability of choosing option j at trial t is given by

$$P(C(t) = j) = \frac{\exp\{\theta(E_j(t) + \beta\sqrt{S_j(t)})\}}{\sum_{k=1}^{6} \exp\{\theta(E_k(t) + \beta\sqrt{S_k(t)})\}}, \quad [27]$$

where $\beta > 0$ is the weight a participant places on uncertainty and the inverse temperature parameter $\theta > 0$ determines the sensitivity to differences in the values.

We fitted the new model using hierarchical Bayesian parameter estimation (see "Hierarchical Bayesian parameter estimation" in *SI Methods*). The KFL-UCB model outperformed both the K2-UPE and KFL-UPE model with a posterior probability of approximately 1. This result affirms that for explaining choices it is estimation uncertainty that matters the most, not unsigned prediction error. In addition, we also compared the new models to all other choice models (Fig. S4). Even though in simulations the K2 algorithm seems like a good competitor to the Kalman filter algorithm (24), the K2-UPE model does not fit the behaviour of our participants well. It outperforms only the RW-UCB and CK-SM model by a convincing margin, and fits choices similar to the KF-SM model. The KFL-UCb model fits the behaviour somewhat better than the K2-UPE, but is still substantially worse than Kalman filter models combined with the UCB choice rule, as well as RWCK-SM model.

Comparison of learning and choice models modulated by visual fixation. In our original analysis we assumed visual fixations can modulate the choice or learning component of the KFL-UCB model that best fitted the behaviour (see "Modelling learning and choices modulated by visual fixations" in *SI Methods*). However, it is possible that certain components of the winning KFL-UCB model may become unnecessary once fixation information is taken into account. For example, the Softmax choice rule might outperform the UCB rule or the "laziness" parameter in the Kalman filter may become unnecessary once we include the fixation information. Hence, as a robustness check, we fitted additional attention-modulated models.

To assess the robustness of our finding that the UCB choice rule is relevant, we considered the Softmax choice rule combined with the same KFL learning component. As in the attention-modulated UCB rule, we here use the relative fixation measure defined in Eq. 11 to re-weight the values in the Softmax (Eq. 3)

$$P(C(t) = j) = \frac{\exp\{\tau \mathrm{RF}_j(t) + \theta E_j(t)\}}{\sum_{k=1}^{6} \exp\{\tau \mathrm{RF}_k(t) + \theta E_k(t)\}}. \quad [28]$$

This re-weighting again has the effect of increasing the choice probabilities for options that received relatively more fixation time and diminishing them for the options that received relatively little fixation time. The learning process in the KFL component is modulated in the same way as in Eq. 13. Overall, we have three additional KFL-SM model variants where either learning (aKFL-SM), the choice process (KFL-aSM), or both (aKFL-aSM) are modulated by attention.

To confirm the usefulness of the laziness parameter we also considered the non-lazy Kalman filter model (KF), combined with either an attention-modulated Softmax choice rule (KF-aSM; as in Eq. 28) or an attention-modulated UCB choice rule (KF-aUCB; as in Eq. 12). As we could not modulate the learning process of the KF component without effectively making it a "lazy" version, we only modulate the choice process.

We fitted all five models using hierarchical Bayesian parameter estimation (see "Hierarchical Bayesian parameter estimation" in *SI Methods*). The results of all five models, together with the three original models (aKFL-UCB, KFL-aUCB and aKFL-aUCB), are illustrated in Fig. S6. The results show that the KFL-aUCB is still the best fitting model, with a posterior probability of approximately 0.77. The aKFL-aUCB model, in which learning is also modulated by fixation at the outcome stage, obtained the remaining posterior probability of approximately 0.23. All other attention modulated models received negligible evidence. What is clearly visible from the ordering of the model performances is that the UCB component is needed (Fig. S6). All models with the UCB component clearly outperformed models with the SM component. Next, attention modulation of the choice process has a large impact on explaining choices. The aKFL-SM and aKFL-UCB model without it performed poorly, coming last in the ordering. The "laziness" parameter is important as well, all models with the "laziness" parameter outperform models without it.

Table S1. Parameters, priors and hyperpriors for choice models. Each panel denotes parameters for a group of models as described in *Materials and Methods* and *SI Methods*.

Parameter	Prior	Hyperpriors
	Modelling learning and choices	
Initial variance, $\sigma_i^{2,g}$	$\exp(\mu_{\sigma_i^2} + \zeta_{\sigma_i^2}\nu_{\sigma_i^2})$	$\mu_{\sigma_i^2} \sim \text{Normal}(2,1)$, $\zeta_{\sigma_i^2} \sim \text{Half-Normal}(0,1)$, $\nu_{\sigma_i^2} \sim \text{Normal}(0,1)$
Innovation variance, $\sigma_\zeta^{2,g}$	$\exp(\mu_{\sigma_\zeta^2} + \zeta_{\sigma_\zeta^2}\nu_{\sigma_\zeta^2})$	$\mu_{\sigma_\zeta^2} \sim \text{Normal}(0,1)$, $\zeta_{\sigma_\zeta^2} \sim \text{Half-Normal}(0,1)$, $\nu_{\sigma_\zeta^2} \sim \text{Normal}(0,1)$
Inverse temperature, θ^g	$\exp(\mu_\theta + \zeta_\theta\nu_\theta)$	$\mu_\theta \sim \text{Normal}(0,1)$, $\zeta_\theta \sim \text{Half-Normal}(0,1)$, $\nu_\theta \sim \text{Normal}(0,1)$
Uncertainty guidance, β^g	$\exp(\mu_\beta + \zeta_\beta\nu_\beta)$	$\mu_\beta \sim \text{Normal}(0,1)$, $\zeta_\beta \sim \text{Half-Normal}(0,1)$, $\nu_\beta \sim \text{Normal}(0,1)$
Laziness, η^g	$\Phi(\mu_\eta + \zeta_\eta\nu_\eta)$	$\mu_\eta \sim \text{Normal}(1,1)$, $\zeta_\eta \sim \text{Half-Normal}(0,1)$, $\nu_\eta \sim \text{Normal}(0,1)$
	Modelling learning and choices modulated by visual fixation	
RF sensitivity, τ^g	$\exp(\mu_\tau + \zeta_\tau\nu_\tau)$	$\mu_\tau \sim \text{Normal}(0,2)$, $\zeta_\tau \sim \text{Half-Normal}(0,1)$, $\nu_\tau \sim \text{Normal}(0,1)$
RF min attention, ϵ^g	$\Phi(\mu_\epsilon + \zeta_\epsilon\nu_\epsilon)$	$\mu_\epsilon \sim \text{Normal}(-1,1)$, $\zeta_\epsilon \sim \text{Half-Normal}(0,1)$, $\nu_\epsilon \sim \text{Normal}(0,1)$
Laziness intercept, η_0^g	$\mu_{\eta_0} + \zeta_{\eta_0}\nu_{\eta_0}$	$\mu_{\eta_0} \sim \text{Normal}(1,1)$, $\zeta_{\eta_0} \sim \text{Half-Normal}(0,1)$, $\nu_{\eta_0} \sim \text{Normal}(0,1)$
Laziness slope, η_0^g	$\mu_{\eta_0} + \zeta_{\eta_0}\nu_{\eta_0}$	$\mu_{\eta_0} \sim \text{Normal}(0,1)$, $\zeta_{\eta_0} \sim \text{Half-Normal}(0,1)$, $\nu_{\eta_0} \sim \text{Normal}(0,1)$
	Modelling choices with visual fixation alone	
RF min attention, ϵ^g	$\Phi(\mu_\epsilon + \zeta_\epsilon\nu_\epsilon)$	$\mu_\epsilon \sim \text{Normal}(-1,1)$, $\zeta_\epsilon \sim \text{Half-Normal}(0,1)$, $\nu_\epsilon \sim \text{Normal}(0,1)$
RF sensitivity, τ^g	$\exp(\mu_\tau + \zeta_\tau\nu_\tau)$	$\mu_\tau \sim \text{Normal}(0,2)$, $\zeta_\tau \sim \text{Half-Normal}(0,1)$, $\nu_\tau \sim \text{Normal}(0,1)$
	Modelling learning and choices – Control models	
CK learning rate, γ^g	$\Phi(\mu_\gamma + \zeta_\gamma\nu_\gamma)$	$\mu_\gamma \sim \text{Normal}(-1,1)$, $\zeta_\gamma \sim \text{Half-Normal}(0,1)$, $\nu_\gamma \sim \text{Normal}(0,1)$
RW learning rate, α^g	$\Phi(\mu_\alpha + \zeta_\alpha\nu_\alpha)$	$\mu_\alpha \sim \text{Normal}(-1,1)$, $\zeta_\alpha \sim \text{Half-Normal}(0,1)$, $\nu_\alpha \sim \text{Normal}(0,1)$
K2 learning rate, ν^g	$\exp(\mu_\nu + \zeta_\nu\nu_\nu)$	$\mu_\nu \sim \text{Normal}(-1,1)$, $\zeta_\nu \sim \text{Half-Normal}(0,1)$, $\nu_\nu \sim \text{Normal}(0,1)$
K2 reward variance, $\sigma_\epsilon^{2,g}$	$\exp(\mu_{\sigma_\epsilon^2} + \zeta_{\sigma_\epsilon^2}\nu_{\sigma_\epsilon^2})$	$\mu_{\sigma_\epsilon^2} \sim \text{Normal}(0,1)$, $\zeta_{\sigma_\epsilon^2} \sim \text{Half-Normal}(0,1)$, $\nu_{\sigma_\epsilon^2} \sim \text{Normal}(0,1)$
UPE learning rate, ζ^g	$\Phi(\mu_\zeta + \zeta_\zeta\nu_\zeta)$	$\mu_\zeta \sim \text{Normal}(0,1)$, $\zeta_\zeta \sim \text{Half-Normal}(0,1)$, $\nu_\zeta \sim \text{Normal}(0,1)$
CK sensitivity, τ^g	$\exp(\mu_\tau + \zeta_\tau\nu_\tau)$	$\mu_\tau \sim \text{Normal}(0,1)$, $\zeta_\tau \sim \text{Half-Normal}(0,1)$, $\nu_\tau \sim \text{Normal}(0,1)$
Inverse temperature, θ^g	$\exp(\mu_\theta + \zeta_\theta\nu_\theta)$	$\mu_\theta \sim \text{Normal}(0,1)$, $\zeta_\theta \sim \text{Half-Normal}(0,1)$, $\nu_\theta \sim \text{Normal}(0,1)$
Uncertainty guidance, β^g	$\exp(\mu_\beta + \zeta_\beta\nu_\beta)$	$\mu_\beta \sim \text{Normal}(0,3)$, $\zeta_\beta \sim \text{Half-Normal}(0,1)$, $\nu_\beta \sim \text{Normal}(0,1)$

Note. All models use non-centered reparametrization as indicated in *Prior* column, often transformed to constrain the parameters to be non-negative (exp) or to a certain range (Probit function, Φ). In addition to the parameter specification listed in *Modelling learning and choices modulated by visual fixations* panel, these models have the same parameters and associated priors as models in *Modelling learning and choices* panel. RF = relative fixation, CK = Choice kernel model, RW = Rescorla-Wagner model, UPE = unsigned prediction error, K2 = learning model from (24).

Table S2. Parameters, priors and hyperpriors for modelling relative fixation in the choice stage and for modelling absolute fixation in the outcome stage. Each panel denotes parameters for a group of models as described in *Materials and Methods* and *SI Methods*.

Parameter	Prior	Hyperpriors
	Modelling relative fixation in the choice stage	
Value, β_v^g	$\mu_{\beta_v} + \zeta_{\beta_v}\nu_{\beta_v}$	$\mu_{\beta_v} \sim \text{Normal}(0, 2)$, $\zeta_{\beta_v} \sim \text{Half-Normal}(0, 1)$, $\nu_{\beta_v} \sim \text{Normal}(0, 1)$
Uncertainty, β_u^g	$\mu_{\beta_u} + \zeta_{\beta_u}\nu_{\beta_u}$	$\mu_{\beta_u} \sim \text{Normal}(0, 2)$, $\zeta_{\beta_u} \sim \text{Half-Normal}(0, 1)$, $\nu_{\beta_u} \sim \text{Normal}(0, 1)$
Game type, β_{gt}	$\text{Normal}(0, 5)$	–
Precision, κ^g	$\exp(\mu_\kappa + \zeta_\kappa\nu_\kappa)$	$\mu_\kappa \sim \text{Normal}(0, 1)$, $\zeta_\kappa \sim \text{Half-Normal}(0, 1)$, $\nu_\kappa \sim \text{Normal}(0, 1)$
	Modelling absolute fixation in the outcome stage	
Intercept, β_i^g	$\mu_{\beta_i} + \zeta_{\beta_i}\nu_{\beta_i}$	$\mu_{\beta_i} \sim \text{Normal}(2, 0.5)$, $\zeta_{\beta_i} \sim \text{Half-Normal}(0, 0.5)$, $\nu_{\beta_i} \sim \text{Normal}(0, 0.5)$
Value, β_v^g	$\mu_{\beta_v} + \zeta_{\beta_v}\nu_{\beta_v}$	$\mu_{\beta_v} \sim \text{Normal}(0, 0.2)$, $\zeta_{\beta_v} \sim \text{Half-Normal}(0, 0.2)$, $\nu_{\beta_v} \sim \text{Normal}(0, 0.2)$
Uncertainty, β_u^g	$\mu_{\beta_u} + \zeta_{\beta_u}\nu_{\beta_u}$	$\mu_{\beta_u} \sim \text{Normal}(0, 0.2)$, $\zeta_{\beta_u} \sim \text{Half-Normal}(0, 0.2)$, $\nu_{\beta_u} \sim \text{Normal}(0, 0.2)$
PE, β_{PE}^g	$\mu_{\beta_{\text{PE}}} + \zeta_{\beta_{\text{PE}}}\nu_{\beta_{\text{PE}}}$	$\mu_{\beta_{\text{PE}}} \sim \text{Normal}(0, 0.2)$, $\zeta_{\beta_{\text{PE}}} \sim \text{Half-Normal}(0, 0.2)$, $\nu_{\beta_{\text{PE}}} \sim \text{Normal}(0, 0.2)$
uPE, β_{uPE}^g	$\mu_{\beta_{\text{uPE}}} + \zeta_{\beta_{\text{uPE}}}\nu_{\beta_{\text{uPE}}}$	$\mu_{\beta_{\text{uPE}}} \sim \text{Normal}(0, 0.2)$, $\zeta_{\beta_{\text{uPE}}} \sim \text{Half-Normal}(0, 0.2)$, $\nu_{\beta_{\text{uPE}}} \sim \text{Normal}(0, 0.2)$
Game type, β_{gt}	$\text{Normal}(0, 0.2)$	–
Scale, ω^g	$\text{Half-Normal}(0, 1)$	–
Shape, α^g	$\text{Normal}(-2, 1)$	–

Note. All models use non-centered reparametrization as indicated in *Prior* column, often transformed to constrain the parameters to be non-negative (exp). Game type parameter in both groups of models was assumed to be fixed effect and hence only prior was necessary. We used same assumption for scale and shape parameter in models of absolute fixation in the outcome stage. PE = Prediction error parameter, uPE = unsigned prediction error parameter.

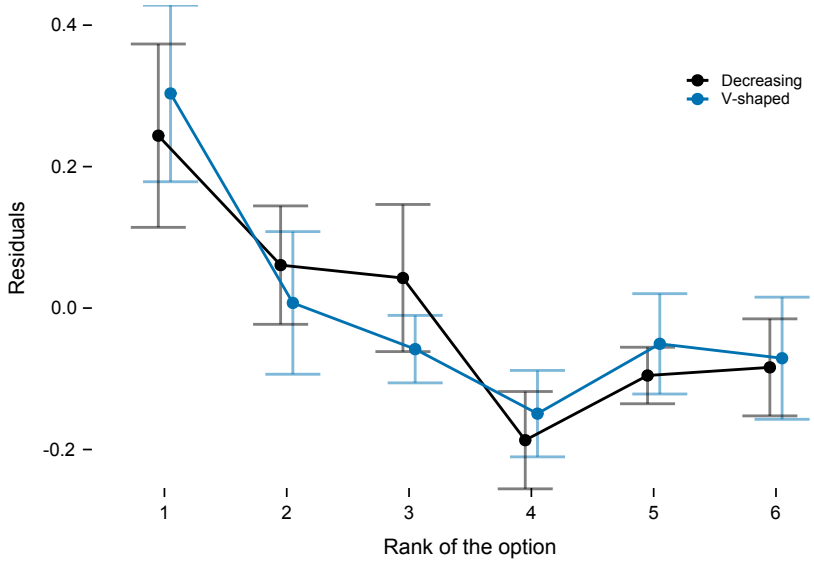

Fig. S1. Standardised residuals after regressing out expected values of options from relative fixations in choice stage do not differ between the games. We fitted a Dirichlet regression model to each game, similar to models of relative fixation in choice stage (see "Modelling choices with visual fixations alone" in *SI Methods*). Expected values were passed through a softmax function to produce concentration parameters in Dirichlet distributed relative fixation data. We fitted the models to the first block of 15 trials where we expected the differences between the games to be the strongest.

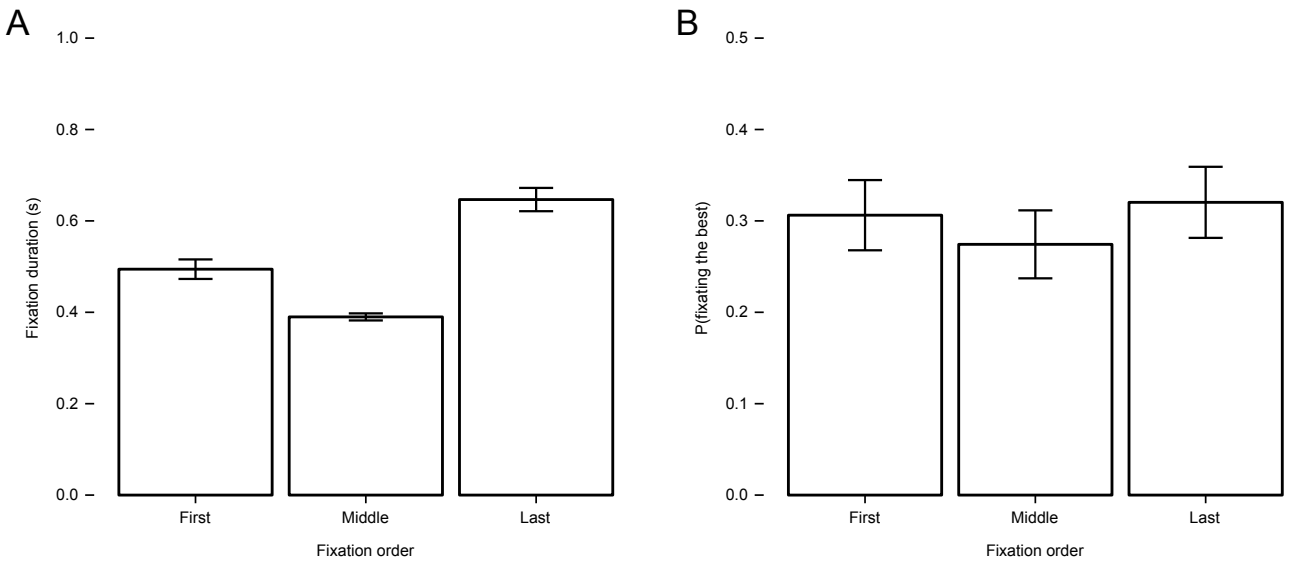

Fig. S2. (A) Last fixation duration in the choice stage of the trial is longer than the duration of middle fixations, contrary to the predictions of evidence accumulation process biased by what is looked at. (B) Probability of first fixating the best option in the choice stage is larger than chance level (0.17). Same holds for middle and last fixation as well.

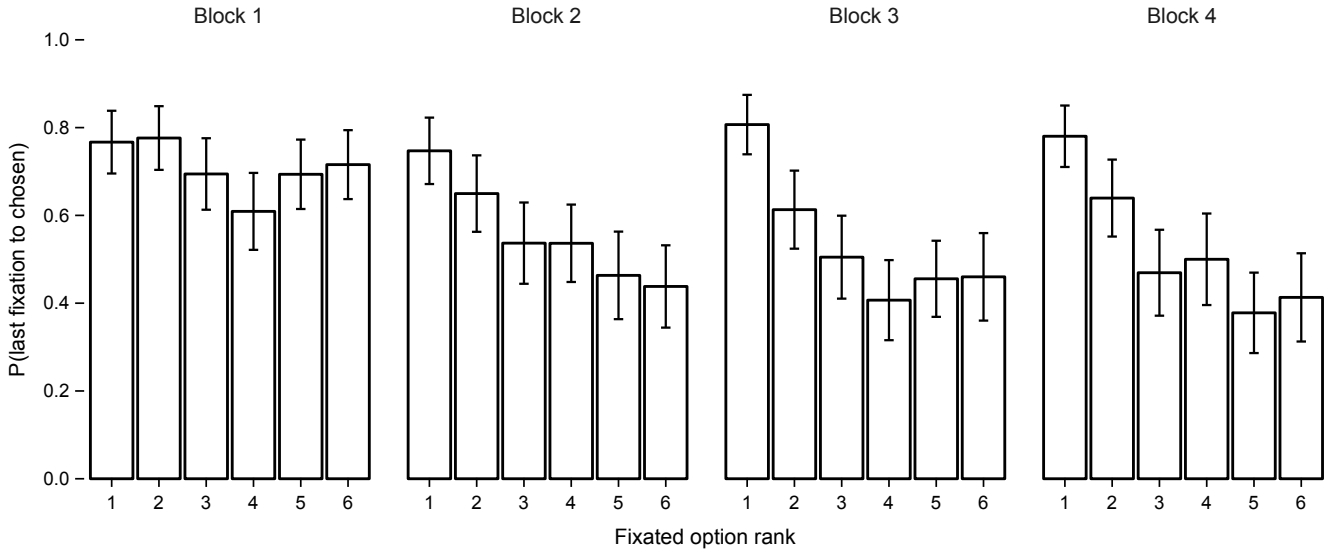

Fig. S3. Participants increasingly chose the option fixated the last, unless the option was low ranking one, as learning went by. This is evidenced by increase in probability that last fixation is to the chosen option for higher ranking options and decrease for lower ranking options, from first block of 15 trials to fourth block of 15 trials in the game.

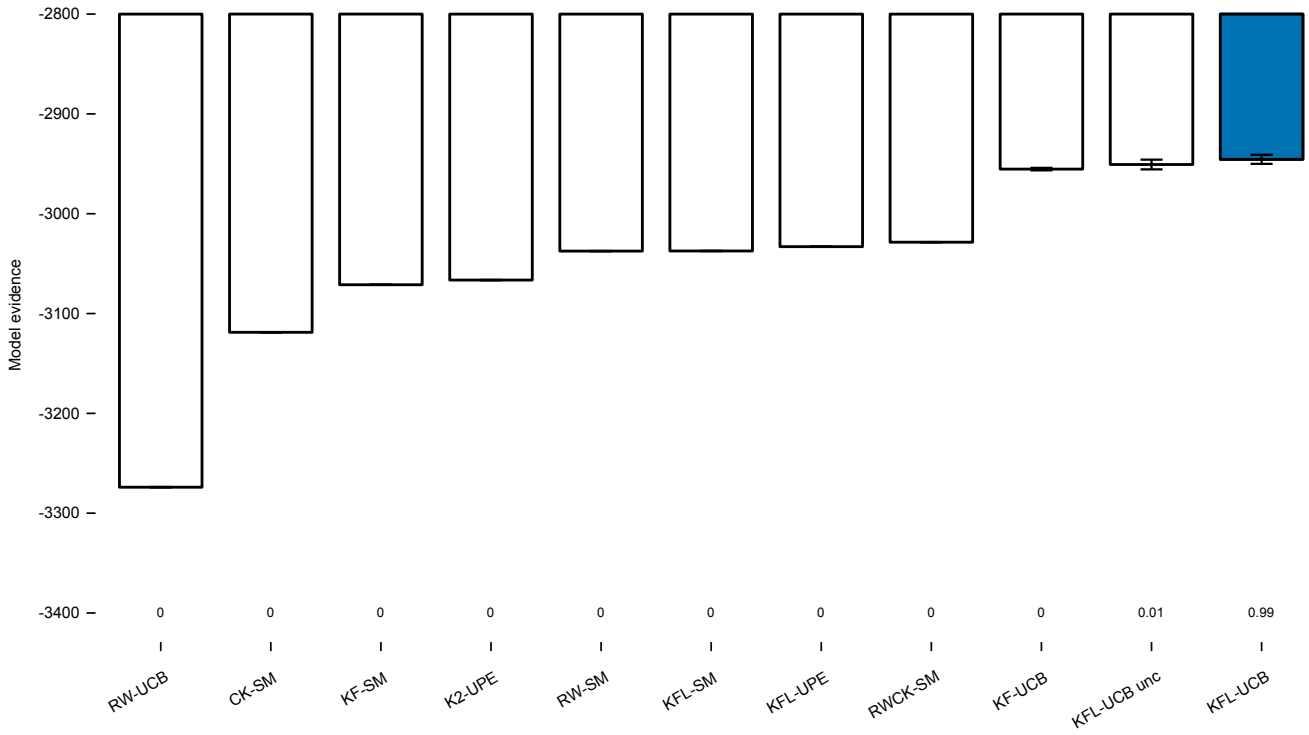

Fig. S4. Model evidence for all choice models. Bars show the median log marginal likelihoods and numbers below the bars show the median posterior probabilities from model comparisons. The lazy Kalman filter learning component combined with the upper confidence bound choice rule (KFL-UCB) describes the participants' choices best. Error bars reflect interquartile ranges of values across repetitions; for most models, these are too small to be visible.

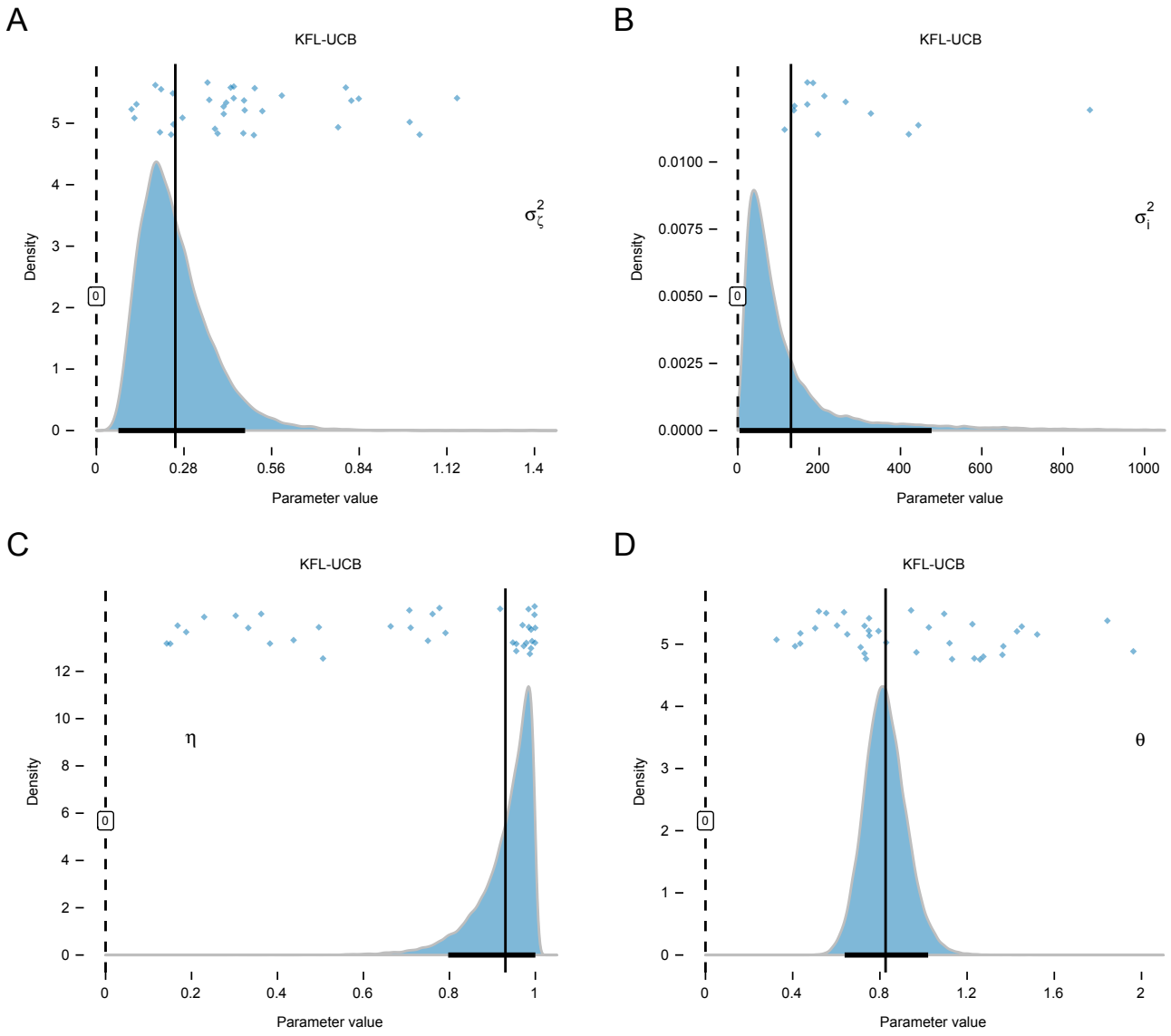

Fig. S5. Posterior distributions of the estimated group-level parameters for the KFL-UCB model not illustrated in the main text. (A) Innovation variance parameter in the Kalman filter learning model (σ^2_ζ). The posterior mean (vertical line) and 95% credible interval (black bar on the x-axis) illustrate the magnitude of the effect. Dots are means of posteriors of individual game level parameters; the vertical jitter is arbitrary. (B) Initial variance parameter in the Kalman filter learning model (σ^2_i). (C) Laziness parameter in the Kalman filter learning model (η). (D) Temperature parameter in the Upper confidence bound rule (θ).

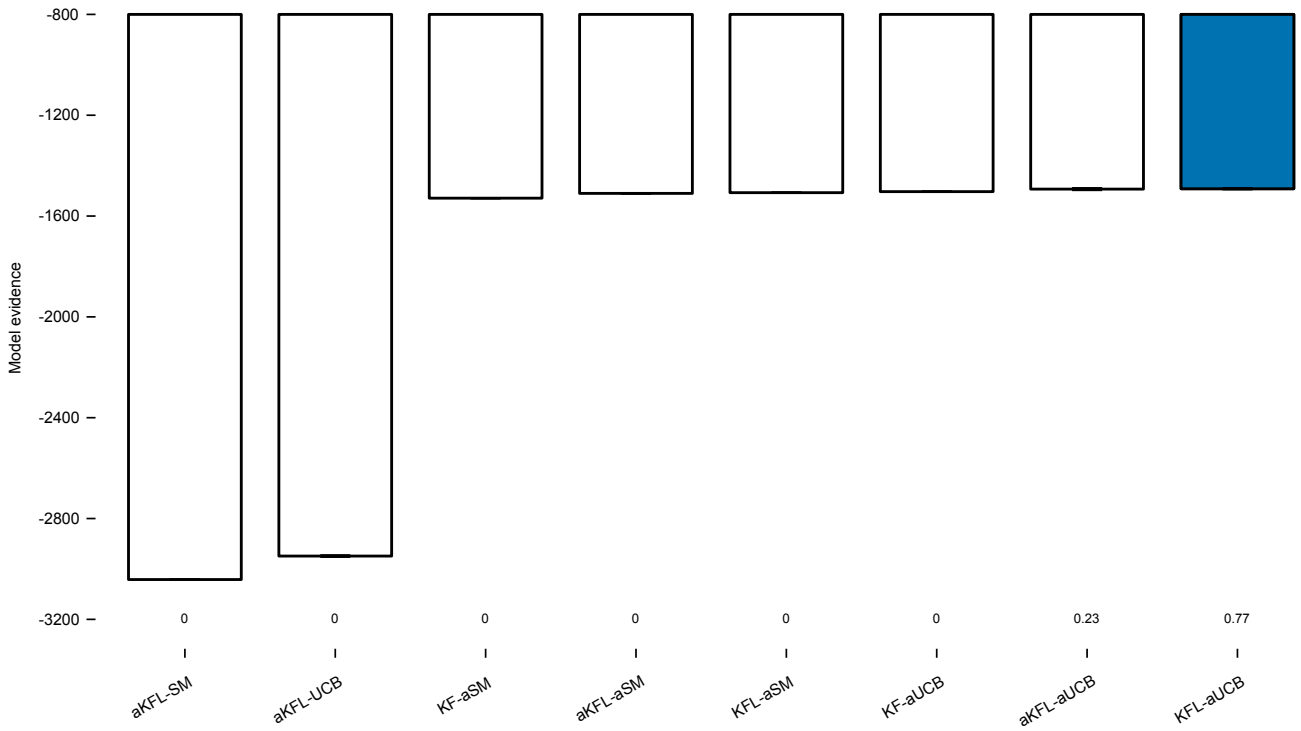

Fig. S6. Model evidence for all attention modulated models. Bars show the median log marginal likelihoods and numbers below the bars show the median posterior probabilities from model comparisons. The lazy Kalman filter learning component combined with the attention-modulated upper confidence bound choice rule (KFL-aUCB) describes the participants' choices best. Error bars reflect interquartile ranges of values across repetitions; for most models, these are too small to be visible.

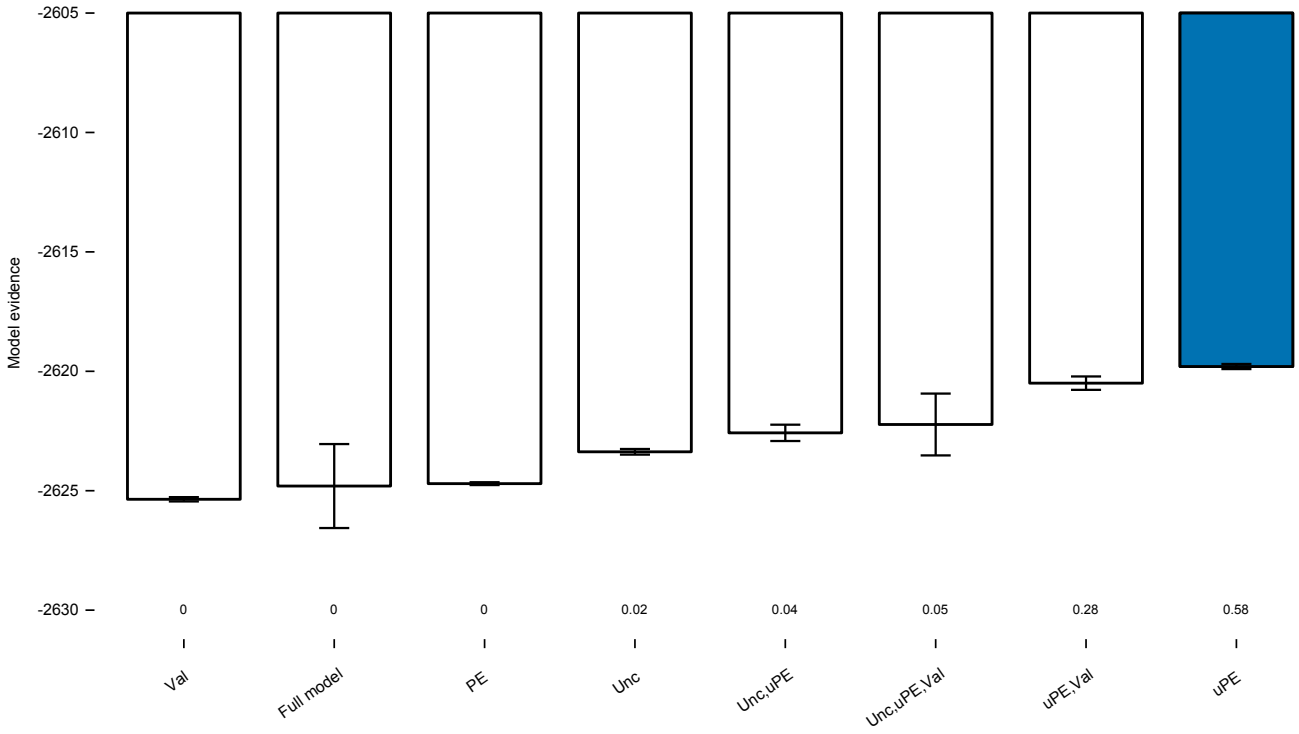

Fig. S7. Interactions between learning and fixation processes at outcome stage. Model evidence (bars) and model comparison (numbers below bars) for the full model – regressing uncertainty (Unc), reward prediction error (PE), unsigned reward prediction error (uPE) and value (Val) on absolute fixations in the outcome stage, and simpler model variants where we excluded some of the predictors. The model with unsigned prediction errors alone explains absolute fixations the best among the compared models. Error bars are interquartile ranges of bridge sampling repetitions (modelling details in SI Methods).

References

1. MacedonianBoy, File:glagolitic zhivete.svg, file:glagolitic zemlja.svg, file:glagolitic slovo.svg, file:glagolitic ljudi.svg, file:glagolitic fita.svg, file:glagolitic izhe.svg, file:glagolitic iota.svg, file:glagolitic kako.svg, file:glagolitic mislete.svg, file:glagolitic on.svg, file:glagolitic ot.svg, file:glagolitic pokoi.svg, file:glagolitic rtsi.svg, file:glagolitic shta.svg, file:glagolitic tverdo.svg, file:glagolitic yerj.svg, file:glagolitic vedi.svg, file:glagolitic az.svg — wikimedia commons, the free media repository (2018) https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_zhivete.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_zemlja.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_slovo.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_ljudi.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_fita.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_izhe.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_iota.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_kako.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_mislete.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_on.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_ot.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_pokoi.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_rtsi.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_shta.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_tverdo.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_yerj.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_vedi.svg, https://commons.wikimedia.org/w/index.php?title=File:Glagolitic_az.svg, [CC BY-SA 3.0 Licence; Online, accessed 14-March-2018].
2. TR Hayes, AA Petrov, Mapping and correcting the influence of gaze position on pupil size measurements. *Behav. Res. Methods* **48**, 510–527 (2016).
3. JW Peirce, PsychoPy - Psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).
4. K Holmqvist, et al., *Eye tracking: A comprehensive guide to methods and measures*. (OUP Oxford), (2011).
5. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2017).
6. PC Bürkner, brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
7. Stan Development Team, RStan: the R interface to Stan (2018) R package version 2.18.2.
8. M Speekenbrink, E Konstantinidis, Uncertainty and Exploration in a Restless Bandit Problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
9. ND Daw, JP O'Doherty, P Dayan, B Seymour, RJ Dolan, Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
10. WK Zajkowski, M Kossut, RC Wilson, A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* **6**, e27430 (2017).
11. RA Rescorla, AR Wagner, A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement in *Classical conditioning II: Current research and theory*, eds. AH Black, WF Prokasy. (Appleton-Century-Crofts, New York, NY, US), pp. 64–99 (1972).
12. RS Sutton, AG Barto, *Reinforcement learning: An introduction*. (MIT Press, Cambridge, MA, US), (1998).
13. P Auer, N Cesa-Bianchi, P Fischer, Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **47**, 235–256 (2002).
14. S Kakade, P Dayan, Dopamine: Generalization and bonuses. *Neural Networks* **15**, 549–559 (2002).
15. JK Kruschke, *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (Academic Press), (2014).
16. M Betancourt, M Girolami, Hamiltonian Monte Carlo for hierarchical models. *Curr. trends Bayesian methodology with applications* **30**, 79–101 (2015).
17. XL Meng, WH Wong, Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sinica* **6**, 831–860 (1996).
18. QF Gronau, et al., A tutorial on bridge sampling. *J. Math. Psychol.* **81**, 80–97 (2017).
19. QF Gronau, H Singmann, *bridgesampling: Bridge Sampling for Marginal Likelihoods and Bayes Factors*, (2018) R package version 0.6-0.
20. I Krajbich, C Armel, A Rangel, Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).
21. RC Wilson, AG Collins, Ten simple rules for the computational modeling of behavioral data (2019).
22. CM Wu, E Schulz, K Gerbaulet, TJ Pleskac, M Speekenbrink, Under pressure: The influence of time limits on human exploration (2019).
23. P Dayan, S Kakade, PR Montague, Learning and selective attention. *Nat. Neurosci.* **3**, 1218 (2000).
24. RS Sutton, Gain adaptation beats least squares in *Proceedings of the 7th Yale workshop on adaptive and learning systems*. Vol. 161168, (1992).