Reward prediction errors induce risk-seeking

Manuscript

Moritz Moeller*1, Jan Grohn*2, Sanjay Manohar**12+, Rafal Bogacz**1.

*: equal contributions

**: equal contributions

1: Nuffield Department of Clinical Neurosciences, University of Oxford

2: Department of Experimental Psychology, University of Oxford

+: corresponding author (sanjay.manohar@psy.ox.ac.uk)

Abstract

Reinforcement learning theories propose that humans choose based on the estimated values of available options, and that they learn from rewards by reducing the difference between the experienced and expected value. In the brain, such prediction errors are broadcasted by dopamine. However, choices are not only influenced by expected value, but also by risk. Like reinforcement learning, risk preferences are modulated by dopamine: enhanced dopamine levels induce risk-seeking. Learning and risk preferences have so far been studied independently, even though it is commonly assumed that they are (partly) regulated by the same neurotransmitter. Here, we use a novel learning task to look for prediction-error induced risk-seeking in human behavior and pupil responses. We find that prediction errors are positively correlated with risk-preferences in imminent choices. Physiologically, this effect is indexed by pupil dilation: only participants whose pupil response indicates that they experienced the prediction error also show the behavioral effect.

Introduction

Reward-guided learning in humans and animals can often be modelled simply as reducing the difference between the obtained and expected reward—a reward prediction error. This well-established behavioral phenomenon [Rescorla, 1972] has been linked to the neurotransmitter dopamine [Schultz, 1997]. It has been shown that bursts of dopaminergic activity broadcast prediction errors to brain areas that are relevant for reward learning, such as the striatum, the amygdala, and the prefrontal cortex.

Another behavioral phenomenon that has been well studied is the effect of uncertainty and risk on decision making [Kahneman, 2013]. Here again, a different line of research has established an association between dopamine and risk-taking: dopamine-enhancing medication has been shown to increase risk-seeking in rats [St Onge, 2009], and drive excessive gambling when used to treat Parkinson's disease [Voon, 2006] [Gallagher, 2007] [Weintraub, 2010]. More recently, it has been demonstrated that phasic responses in dopaminergic brain areas modulate moment-by-moment risk-preference in humans: the tendency to take risks correlated positively with the magnitude of task-related dopamine responses [Chew, 2019]. A family of mechanistic theories of the basal ganglia network provides an explanation for these risk effects [Mikhael, 2016] [Moeller, 2019]. According to these models, positive and negative outcomes of actions are encoded separately in direct and indirect pathways of the basal ganglia. The balance between those pathways is controlled by the dopamine level. An increased dopamine level promotes the direct pathway and hence puts emphasis on potential gains, thus rendering risky options more attractive.

In summary, dopamine bursts are related to distinct behavioral phenomena—learning and risk-taking—by way of 1) acting as reward prediction errors, affecting synaptic weights during reinforcement learning, and 2) inducing risk-seeking behavior directly. There is no obvious a priori reason for those functions to be bundled together; in fact, one would perhaps expect them to work independently, and

their conflation might lead to interactions, unless some separation mechanism exists. There have been different suggestions for such separation mechanisms: it has been proposed that the tonic level of dopamine might modulate behavior directly, while phasic dopamine bursts provide the prediction errors necessary for reward learning [Niv, 2007]. Alternatively, cholinergic interneurons might flag dopamine activity that is to be interpreted as prediction errors by striatal neurons [Berke, 2018]. However, it has also been suggested that the architecture of the basal ganglia is well set-up to both learn from reward-prediction errors and use them to regulate risk-preferences [Mikhael, 2016] [Moeller, 2019].

Curiously, even though the multi-functionality of dopamine has been noted and separation mechanisms have been proposed, interference between the different functions has never been investigated experimentally. Here, we investigate this: if dopamine indeed provides both prediction errors for learning and modulates risk preferences, do these two processes interfere with each other, or are they cleanly separated by some mechanism? Can prediction errors induce risk-seeking?

A proven method to provoke prediction-error related dopamine bursts in humans is to present cues and outcomes in sequential decision-making tasks, hence causing prediction errors both when options are presented, and at the time of outcome [Seymour, 2004] [Pessiglione, 2006] [Niv, 2012]. To test whether such prediction errors induce risk seeking, we used a learning task in which prediction errors are followed by choices between options with different levels of risk. If there was a clear separation of roles, then risk preferences should be independent of prediction errors. Incomplete separation, in contrast, should result in a correlation between risk preferences and preceding prediction errors. In particular, we hypothesized that positive prediction errors (expectations exceeded) should induce risk seeking, while negative prediction errors should lead to risk aversion.

To consider the possibility that this effect might not appear equally strongly in all participants—which could be due to differences in behavioral strategy, neural information processing or risk- and learning-

related traits—we also tracked pupil dilation, which is comparatively robust, and known to reflect surprising events such as prediction errors events [Preuschoff, 2011] [Browning, 2015] [Lawson, 2017] [Cavanagh, 2014]. In particular, we hypothesized that participants who experience stronger prediction errors, as indexed by pupil dilation, also show a stronger behavioral effect of risk preferences.

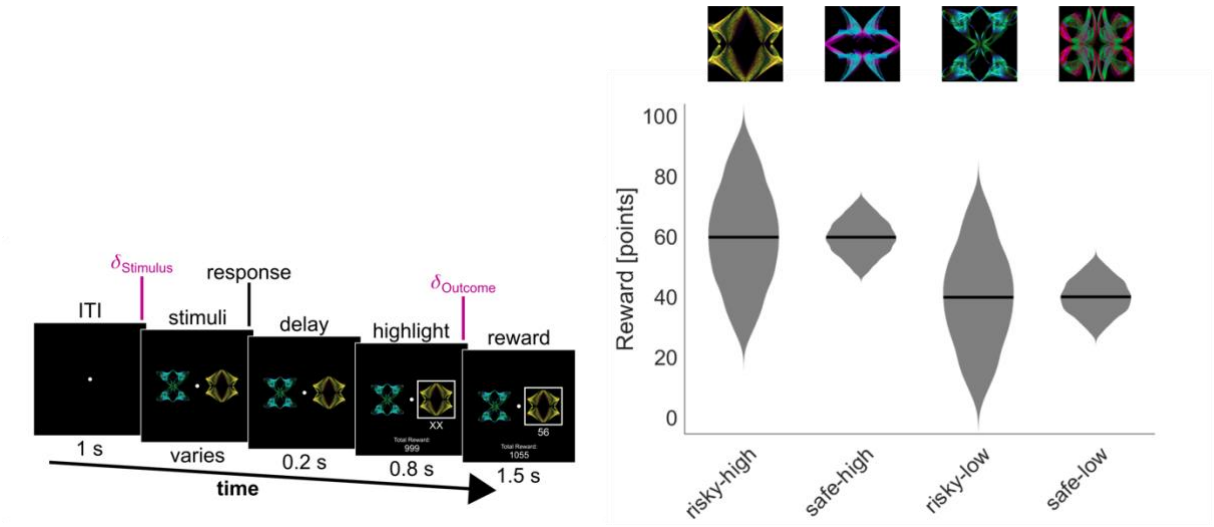
Our analysis proceeds in three steps: first, we conduct a model-free analysis of behavioral data to look for effects on the group level—do prediction errors make participants more risk seeking on average? Second, we move on to uncover individual differences. The effect we are interested in is likely not expressed homogenously; therefore, we use a trial-by-trial learning model to determine the effect size for individuals. Thirdly, we harness these individual differences by linking the strength of the behavioral effect to pupil dilation. This way, we validate our model on independent data, as well as explore a potential reason for the identified individual differences.

Results

The task

Our task consisted of sequences of two-alternative forced choice trials. On each trial, after an inter-trial interval (ITI) of 1 s, two stimuli (fractal images, Fig 1A) were drawn from a set of four stimuli and shown to the participant, who had to choose one. Following the choice, after a short delay of 0.8 s a numerical reward between 1 and 99 was displayed under the chosen stimulus for 1.5 s. Then, the next trial began. Participants were instructed to try to maximize the total number of reward points throughout the experiment. The reward on each trial depended on the participant's choice: each stimulus was associated with a specific reward distribution from which rewards were sampled. The four reward distributions associated with the four stimuli were approximately Gaussian and followed a two-by-two design: the mean of the Gaussian could be either high or low (60 or 40), and the standard deviation could be either large or small (20 or 5), resulting in four reward distributions in total (risky-high, risky-low, safe-high and safe-low, Fig 2B). Each participant (N=27, 3 excluded, see Methods and Fig S1) performed four blocks of 120 trials. During each block, all six possible stimuli pairings occurred equally often. Each block used a new set of four stimuli, mapped to the same four distributions.

93 **A** **B**



94

95 *Fig 1: A) Task structure. On each trial, participants were shown two out of four possible stimuli. They had*
96 *to choose one of the two, which resulted in a reward. The reward was sampled from a distribution linked*
97 *to the chosen stimulus. During each trial, prediction errors occurred at two times (indicated by purple*
98 *lines). B) Reward structure. Each reward distribution is linked to one stimulus and is sampled from if that*
99 *stimulus was chosen. The reward distributions are approximately normal; their parameters follow a two-*
100 *by-two design: the mean could either be at 40 or at 60, the standard deviation could either be 5 or 20.*

101

102 During each trial two distinct prediction errors occur. At stimulus onset it is revealed to participants
103 whether the potential reward on this trial will likely be above or below average. This can be determined
104 by considering the difference between the learned means of the two available options and the average
105 reward associated with all four possible options. A positive prediction error occurs when the displayed
106 options promise a higher than average reward, while a negative prediction error occurs when the
107 expected reward is lower than average. This update of the reward prediction at stimulus onset will be
108 called stimulus prediction error $\delta_{Stimulus}$. Stimulus prediction errors have previously been investigated:

they are associated with phasic responses of dopamine neurons [Schultz, 1997], and have, for example, been used to assess the impact of dopamine on the formation of episodic memories [Jang, 2019]. After considering the options, the participant will make a choice, and be presented with a reward. The difference between the expectation and the actual outcome corresponds to a second reward prediction error, which we call the outcome prediction error $\delta_{outcome}$.

In our task, risk corresponds to the variability of the rewards associated with a stimulus. The reward distributions associated with some options are broad, while those related to other options are narrow (Fig 2B). It is "risky" to pick a stimulus associated with a broad reward distribution, since outcomes might deviate a lot from the expected outcome. Correspondingly, it is "safe" to pick a stimulus with a narrow distribution, since the outcomes will mostly be as expected. Note that some stimuli were matched to produce the same reward on average, while differing in variability. If participants have accurately learned the average reward of these options, then choices between those stimuli cannot be based on a value difference (since on average there is none); residual preferences must therefore be interpreted as risk preferences. Those choices between matched-mean stimuli were our way of reading out risk preferences, and we refer to such trials as matched-mean trials. In the other trials, one of the options provides substantially more reward than the other option (20 points difference on average). We refer to those trials as different-mean trials.

Behavior

To confirm that participants had understood the task and had learned the values associated with the four options, we first analyzed their choices in different-means trials. We observed a gradual shift from initial indifference to a strong preference for the high mean stimuli (proportion of correct choices after trial 40 > 0.5, t-test, t = 38.6, p = 1.73 x 10^{-24}; see Fig 2A, first column). This suggests that participants understood the instructions and learned values accurate enough to maximize reward points.

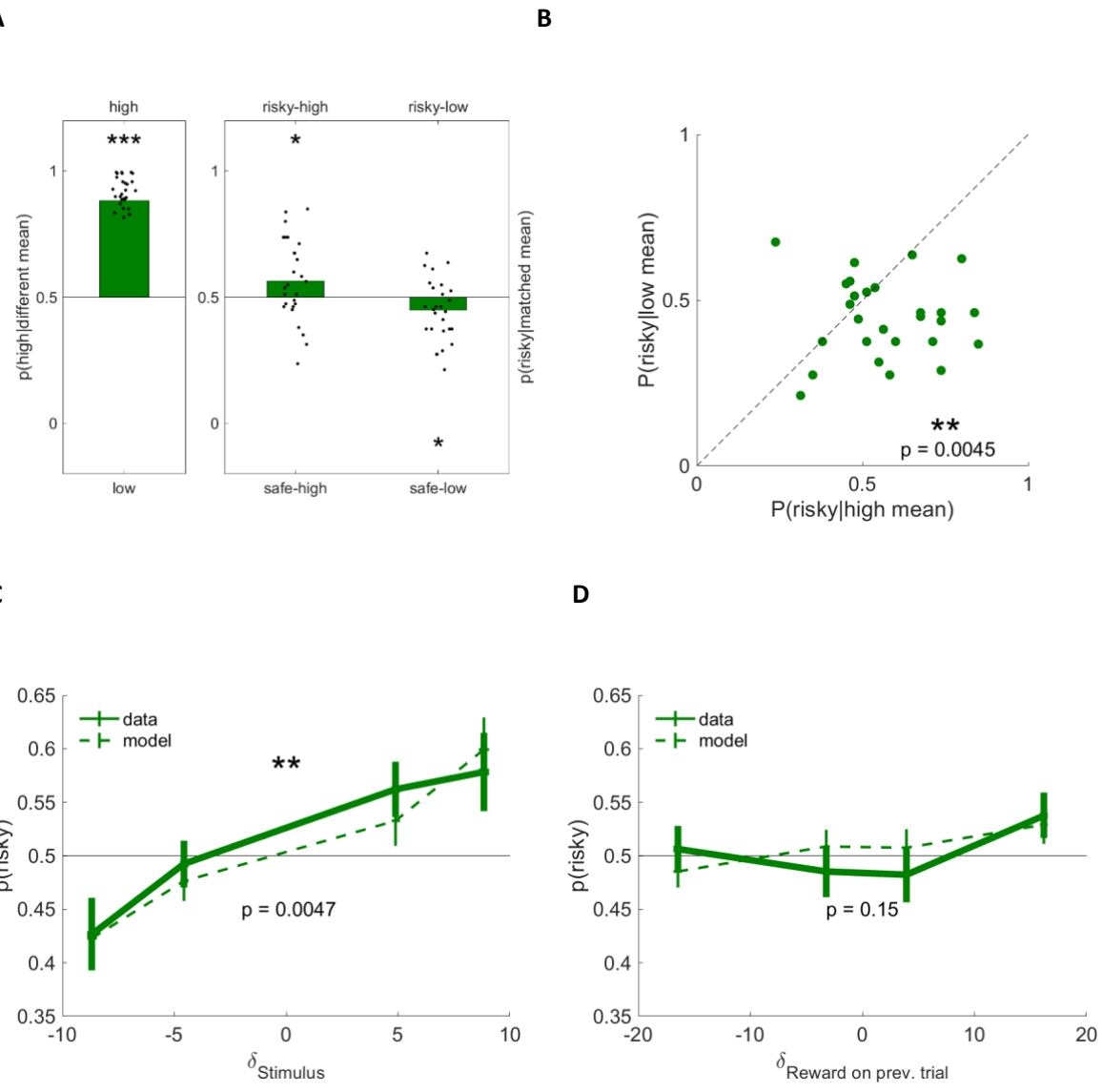


Fig 2: A) Choice proportions. The bars mark the mean proportion of choices across the entire cohort, the black dots mark mean choice proportions for each participant. Panel 1 shows the proportions of choices between the high-mean stimulus and the low-mean stimulus after trial 40. Panels 2 and 3 show the proportions between the two high-mean stimuli (2, risky-high versus safe-high) and the two low-mean stimuli (3, risky-low versus safe-low), respectively. B) Correlation between choice proportions. Each point represents one participant. If a point falls below the diagonal, the participant was more risk seeking for high-mean stimuli than for low-mean stimuli. C) Impact of stimulus prediction errors on risk preference.

Prediction errors and value estimates were obtained by fitting a Rescorla-Wagner model to the choice data. Choices in matched-mean trials were binned by participant, value difference and stimulus prediction errors. The proportion of risky choices was then averaged across all but the prediction error bin. The solid green line shows the residual dependency of proportion of risky choices on stimulus prediction errors (error bars indicate the standard error of the mean). This binning method controls for confounding effects related to incidental differences in learned values and differences between individuals. The dashed line was obtained using the same binning method on predicted choice probabilities, obtained through a logistic regression fitted to predict choices from value differences, outcome prediction errors and participant ID (see Methods for details). D) Impact of outcome prediction errors on risk: identical to C), except this time using outcome prediction errors on the previous trial instead of stimulus prediction errors as predictor.

In addition to this clear preference for high-mean options, we found a weak but significant preference towards the risky stimulus in risky-high versus safe-high choices (Fig 2A, second column; t-test: $p = 0.0343$, $t = 2.23$), and a weak but significant preference against the risky stimulus in risky-low versus safe-low choices (Fig 2A, third column; t-test: $p < 0.0317$, $t = -2.27$). This suggests that on average, participants acted risk-seeking in high reward contexts, and risk-averse in low reward contexts. In addition to this group-level analysis, we investigated how preferences differed between the matched-mean conditions within each participant. We found that most of the participants were more risk seeking in the high-mean condition than in the low mean condition (Fig 2B; paired t-test: $t = 3.11$, $p = 0.0045$). These results are in line with previous findings [Wulff, 2018] [Madan, 2014], see Discussion for details.

Next, we investigated whether these risk preferences could be due to prediction errors. Our hypothesis was that the dopamine release triggered by prediction errors might bias the participants' preferences

towards the risky option. To test this hypothesis, we fitted a basic Rescorla-Wagner (RW) model to each participant's behavior to obtain trial-by-trial estimates of subjective values and prediction errors (see Modelling and Methods for model specifications and fitting procedure). We then extracted both the stimulus prediction error $\delta_{Stimulus}$ and the outcome prediction error $\delta_{Outcome}$, and checked whether these prediction errors were correlated with the risk preference displayed in the following choice. This was done by fitting logistic regressions to the choices recorded in matched-mean trials (see Methods for details of the procedure). We found that the probability of choosing the risky option was predicted by the stimulus prediction error, but not by the previous trial's outcome prediction error (Stimulus prediction error: Fig 2C; chi-squared test, chi-squared = 8.00, df = 1, p: 0.00468. Outcome prediction error: Fig 2D; chi-squared test, chi-squared = 2.11, df = 1, p = 0.146). This suggests that the stimulus prediction error immediately before the choice (0.97 s delay on average, with standard deviation 0.51) but not the outcome prediction error on the previous trial (3.47 s delay on average, with standard deviation 0.51) modulates risk preferences on a trial-by-trial basis.

Modelling

Having established that there was a correlation between prediction errors and risk-seeking, we tried to capture the effect in a reinforcement learning model. We designed a model that tracks the stimulus-specific mean rewards Q, as well as the stimulus-specific spreads S. More explicitly: Q_i represents an estimate of the average reward obtained after choosing stimulus i, while S_i represents an estimate of the mean absolute deviation or "spread" of that reward. Spread is one way to quantify risk, since it measures how unpredictable the stimulus is. Our model updates Q using a conventional Rescorla-Wagner rule,

$$\Delta Q_i = \alpha_Q \delta_{outcome},$$

where $\delta_{outcome} = r - Q_i$ is the outcome prediction error, and α_Q is the learning rate for value. The model updates the estimated spread S using a similar rule,

$$\Delta S_i = \alpha_S(|\delta_{outcome}| - S_i),$$

where α_S is the learning rate for risk. After sufficient burn-in, this rule produces S that fluctuate around the mean absolute deviation of the reward distributions, and hence provides an estimate of the risk associated with each stimulus. Our learning rules are analogous to plasticity rules that feature in a computational model of the basal ganglia (where the mean is encoded in the difference between synaptic weights of the direct and indirect pathway, while the spread is encoded in the sum of these weights) [Mikhael, 2016] [Moeller, 2019]. In those models, choices are based on subjective values V which are assembled from the mean rewards Q and the dopamine-weighted spreads S. Following these models, we define the subjective value of reward in the following equation, where the spread is weighted by the stimulus prediction error—which is indicative of dopamine activity—on that trial:

$$V_i = Q_i + \gamma \ \delta_{stimulus} S_i \qquad \text{(Eq. 1)}$$

where $\delta_{Stimulius} = \frac{1}{2}\sum_{i \in options} Q_i - \frac{1}{4}\sum_j Q_j$ (the stimulus prediction error represents the change in reward expectation before and after the presentation of the options). Note that we include the stimulus prediction error, but not the outcome prediction error on the previous trial, because only the former showed an effect on choices in our previous analysis (see Fig 2C and 2D). The parameter γ thus captures the extent to which recent dopaminergic prediction errors might modulate risk preference.

On every trial, subjective values V are computed from the learned Q and S for both available options. Those values are then softmax-transformed into a probability distribution, from which choices are sampled. This model, which we call Prediction Error Induced Risk Seeking (PEIRS), can be understood as

209 a generalization of the RW model, which is contained in it as a special case ($\gamma = 0$ decouples risk from
210 choices and recovers the conventional value based RW model).

211 We fitted our PEIRS model (as well as a conventional RW model) to the choice data of each participant
212 individually, to obtain estimates on the strength and direction of prediction error induced risk seeking
213 for each participant (see Methods for details). A model comparison for each participant individually
214 showed that 11 out of 27 participants were better described by PEIRS than by RW (Fig 3B). This means
215 that for 11 out of 27 participants in our cohort (about 40 %), prediction error induced risk seeking is
216 strong enough to merit extra parameters.

217 We next investigated the posterior parameter distributions for γ that we obtained from the fit. We
218 found that they are grouped around a positive mean significantly different from zero (Fig 3C; one-tailed
219 t-test: t = 2.67, p = 0.0064). That γ tends to be positive across the cohort is in line with the dopaminergic
220 interaction we propose. Note that the model did not feature any bias for γ to be positive; the positive
221 tendency that we observe in the fitted values is due entirely to biases in our participant's behavior.

222 Overall, our basic analysis of behavior as well as the model comparison both suggest that there is a
223 significant positive interaction between prediction errors and ensuing risk-seeking.

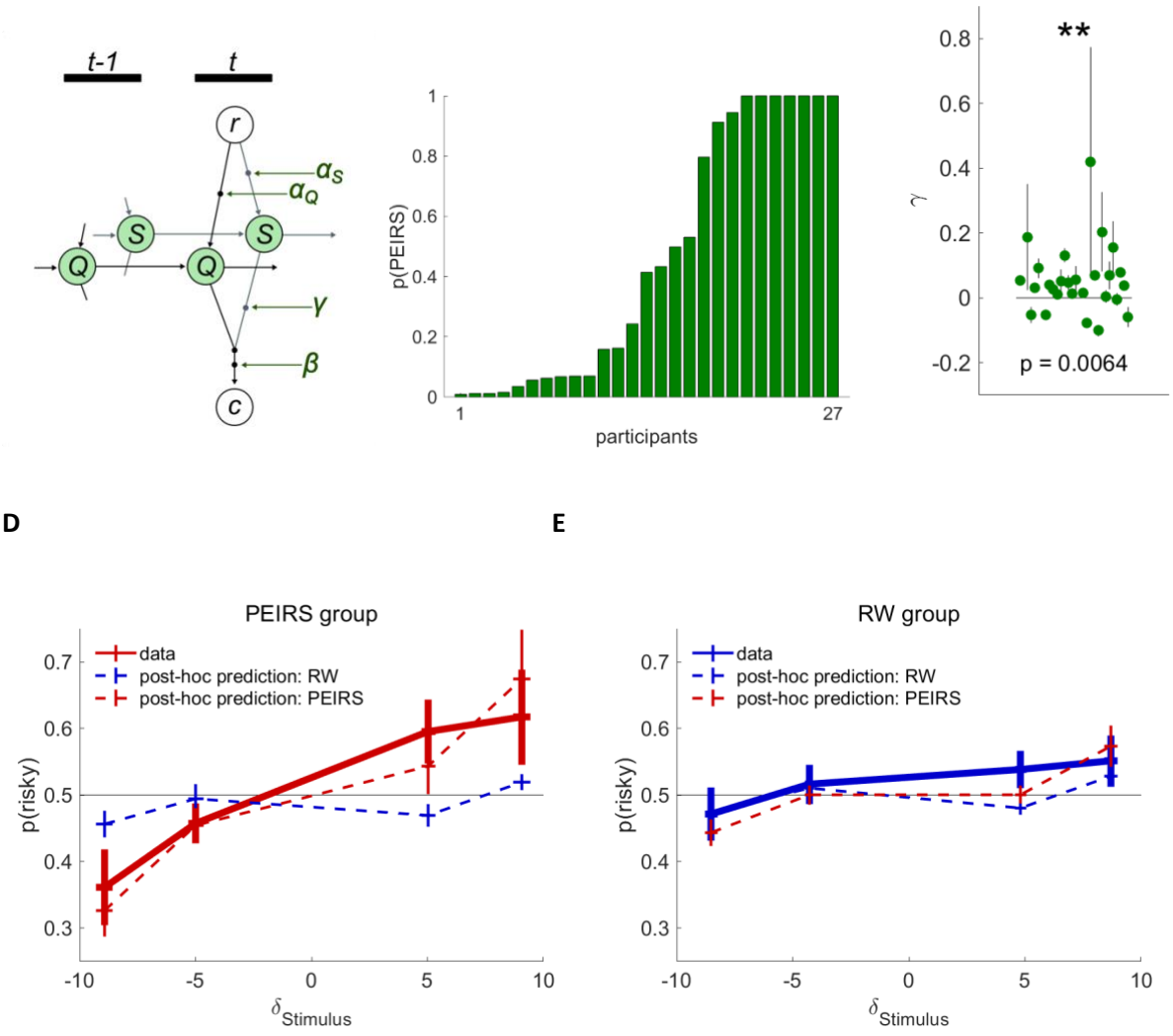


Fig 3: A) Schematic representation of our generative model. The latent variables (estimated mean rewards Q and mean spreads S) are depicted as pale-green circles; the observable variables (rewards r and choices c) correspond to white circles. Black arrows represent the dependencies between those variables. The model parameters are annotated in dark green. B) Probabilities that participant shows prediction-error induced risk seeking. Each bar indicates the probability that single a participant generated data according to the PEIRS model rather than generating data according to an RW model. Participants were sorted according to this probability. C) Parameter estimates extracted from the fit.

Estimates of the parameter γ from all participants are shown (green dots). The error bars represent the standard deviation of the corresponding posterior distribution. D) Prediction-error induced risk seeking in participants for which PEIRS wins the model comparison. The red solid line represents choice data of these participants, the dashed lines represent choice predictions extracted from model fits. Choice data and choice predictions are plotted in the same way as in Fig 2C and 2D, merely displaying posterior predictions from our generative models instead of predictions of simple logistic models. D) Prediction-error induced risk seeking in participants for which RW wins the model comparison. Similar to C), but based on a complementary subgroup of participants.

To check whether the model indeed captured the effect that it was intended to capture, we performed post-hoc simulations [Palminteri, 2017]. Using the posterior predictive density over choices that the we obtain as an output of the fit, we generated post-hoc predictions for all choices (i.e. we used the fitted models to predict probabilities for all choices). Such predictions were generated for all participants, both from the PEIRS model and the RW model, leaving us with three data sets: a data set simulated from the fitted RW model, a data set simulated from the fitted PEIRS model and the data set obtained from humans in our experiment. We then split all these data sets according to the model comparison results (i.e. whether a participant's choices are best described by the PEIRS or by the RW model), and used the same binning scheme as for the recorded choices to check whether the two models predicted any dependency of risk preferences on reward prediction errors (Fig 3D and 3E, dashed lines).

The behavior simulated from the RW model did not show any substantial dependency between prediction errors and risk-taking, even when fitted to participants whose choices were better explained by the PEIRS model (blue dashed lines in Fig 3D and 3E). The PEIRS model, on the other hand, produced an approximately linear dependency between risk-taking and prediction errors for the participants best

described by the PEIRS model, but did not produce any dependency for the participants best described by the RW model (red dashed lines in Fig 3D and 3E). The tendencies simulated from the PEIRS model coincide with the tendencies that were observed (experimentally observed tendencies correspond to the solid lines in Fig 3D and 3E; compare the thick lines to the blue dashed lines). We concluded that the PEIRS model successfully captured our participant's risk preferences both qualitatively (linear upwards trend) and quantitatively (both intercept and slope coincide). The RW model, on the other hand, was not able to capture the risk preferences, even with fitted parameters.

We ran three additional tests to check the robustness of our results and the validity of our conclusions. First, we assessed the reliability of our parameter estimates by performing a standard parameter recovery analysis for both models. We found that all parameters could be recovered with little distortion, for both models and realistic parameter settings (see Fig S2). Second, we tested whether reward predictions (rather than prediction errors) might be the cause of risk-seeking. A model where reward predictions induced risk-seeking did not fit the data as well as the PEIRS model. Additional analyses based on linear models confirmed that prediction errors are more likely than reward predictions to cause the observed risk preferences (see Fig S3). Third, we tested whether our results depended on of the linearity of the utility function. We found that the interaction between risk-seeking and reward prediction errors was present even when we accounted for a nonlinear utility function (see Fig S4). These tests suggest that our results are robust if assumptions are modified, and provide additional support for our conclusions.

Pupillometry

A range of studies have demonstrated a dilation of the pupil in response to surprising events [Preuschoff, 2011] [Browning, 2015] [Lawson, 2017] [Cavanagh, 2014]. Those phenomena have recently been synthesized into a coherent theory: pupil dilation is triggered by belief updates and scales with the

mismatch between prior and posterior beliefs [Zénon, 2019]. We sought to capitalize on this effect, to 1) establish the occurrence of the two prediction errors during our task through a physiological marker, and 2) to understand the individual differences suggested by our behavioral modelling better.

As a first step, we investigated whether pupil dilation reflected updates in reward expectation (i.e. prediction errors). We used the absolute value of the two task-related prediction errors, $|\delta_{Stimulus}|$ and $|\delta_{Outcome}|$, as a measure of mismatch between prior and posterior reward expectation. Trial-by-trial estimates of those prediction errors were extracted from the PEIRS model fits. Regression analyses were used to determine whether pupil dilation after stimulus onset encoded $|\delta_{Stimulus}|$, and whether dilation after reward presentation encoded $|\delta_{Outcome}|$. To avoid confounding factors such as reward or outcome prediction errors, we censored all data points collected after reward presentation in the analysis of the stimulus prediction error. For both prediction errors, we found delayed phasic responses which peaked 1.6 s after stimulus onset and 0.9 s after reward presentation, respectively (Stimulus prediction error: Fig 4A; t-test: t = 2.89, p = 0.0079. Outcome prediction error: Fig 4B; t-test: t = 4.61, p = 0.00010. Statistical significance was established through leave-one-out unbiased peak detection, see Methods). The responses were similar for both prediction errors, except for a longer delay between stimulus prediction error onset and the peak of the pupil response. There might be many reasons for this difference in delay. Among those, differences in information processing might play a role: generating a stimulus prediction error involves two stimuli, hence attention mechanisms, in addition to retrieval of value estimates from memory. Generating the outcome prediction error, on the other hand, just requires the processing of a number.

We concluded that both prediction errors occur as assumed in our modelling, not only as cognitive variables, but as measurable physiological events with appropriate timing. This means that our model must at least partially represent the neural processes that occur during decision making, since it provided us with latent variables that are correlated with physiological variables in a meaningful way.

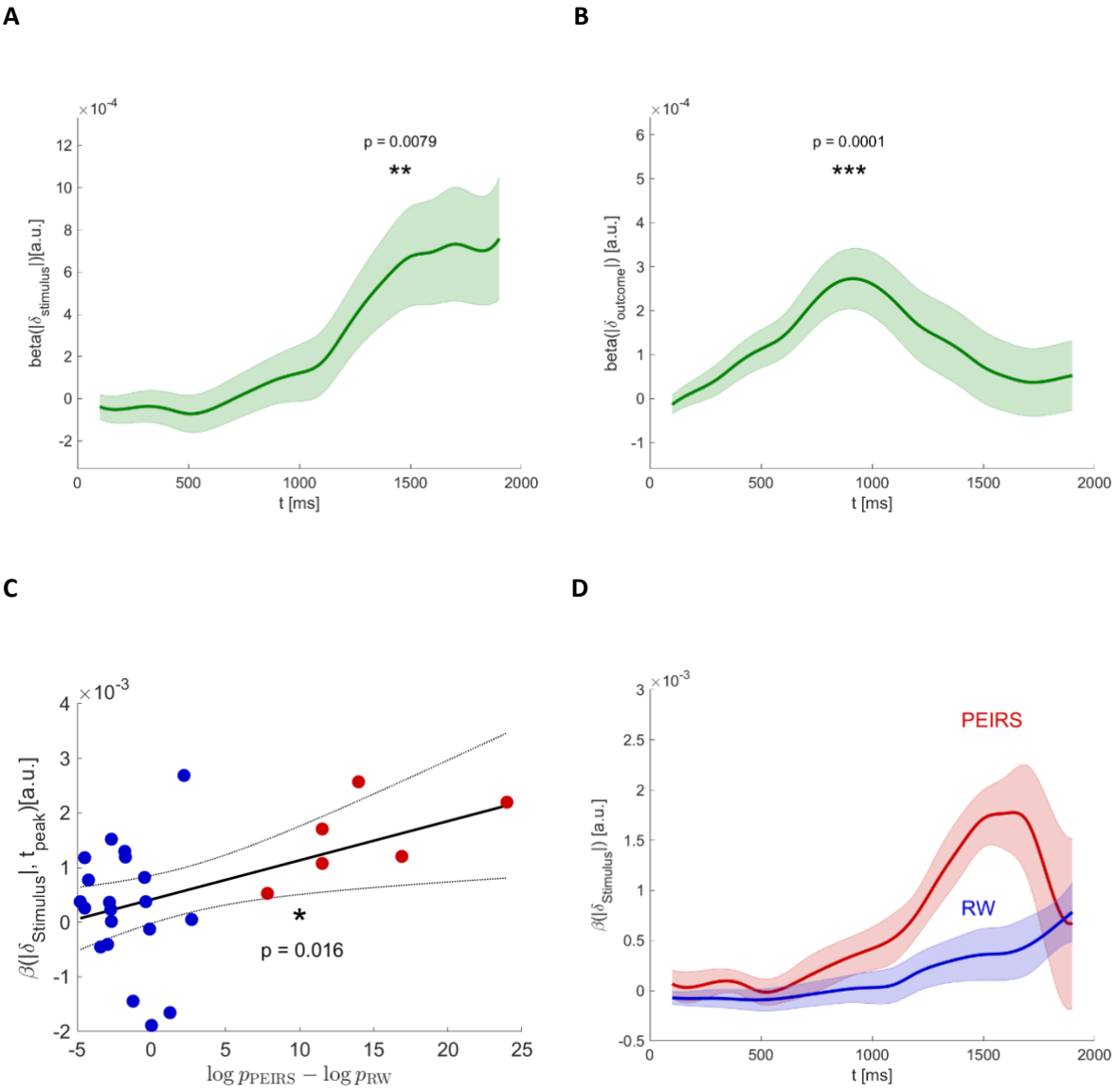


Fig 4: A) Pupil dilation encodes the magnitude of the stimulus prediction errors. The green line represents the regression coefficient across participants as a function of time, extracted from a mixed-effects model. Responses were aligned at stimulus onset. The shaded area indicates the standard error of the estimate, as provided by the regression model. For display, the trace was smoothed using spline interpolation. B) Pupil dilation encodes the magnitude of the outcome prediction errors. Similar to A), with responses aligned at reward presentation. C) Effect strength in behavior predicts effect strength in pupil dilation. The plot shows the correlation between the log-odds that a participant generated choices according to

316 *the PEIRS model (x-axis) and the regression coefficient for the stimulus prediction error as a predictor of*
317 *pupil dilation, at time of maximum effect strength (y-axis). The coloring of the dots corresponds to the*
318 *split of the cohort used in D). D) Group split to illustrate how behavior predicts pupil response. The red*
319 *curve represents the mean pupil response across those participants that show strong prediction error*
320 *induced risk-seeking (p(PEIRS) > 0.95); the blue curve represents the mean pupil response of all other*
321 *participants.*

322

323 Next, we aimed to correlate individual differences in behavior with individual differences in pupil
324 responses. Our behavioral modelling allowed us to stratify our cohort with respect to the strength of
325 individual prediction-error induced risk seeking (see section Modelling). We quantified the effect
326 strength through the logarithmic odds ratios of the probability that a participant is better described by
327 the PEIRS model than by the RW model (see Fig 3A for a plot of those probabilities). The strength of the
328 pupil response was quantified through the respective correlation coefficient at the time of strongest
329 effect (determined though leave-one-out peak detection to avoid bias). Using a linear model to relate
330 pupil effect strength and behavioral effect strength, we found that participants responded stronger to
331 stimulus prediction errors if they showed more pronounced prediction-error induced risk seeking in
332 their behavior (Fig 4C, adjusted R^2: 0.1864, $p < 0.05$). To better illustrate this, we split our cohort into
333 two groups: those that showed significant prediction error induced risk seeking (i.e. $p(RW) < 0.05$, PEIRS
334 group) and the rest ($p(RW) > 0.05$, RW group). While there is no noticeable response in the RW group,
335 the PEIRS group shows a very pronounced effect (Fig 4D). Overall, this suggests that the degree to which
336 some individual responds to the stimulus prediction error (as measured through the strength of the
337 pupil response) is correlated to the strength of prediction error induced risk-seeking in the behavior of
338 that individual.

339 To rule out a possible circularity that might confound these results (both the log-odds-ratio and the
340 predictor variable for the response come from the same model fit), we conducted additional analyses
341 based on estimates of the stimulus prediction error that were independent from the model
342 (Supplemental Materials, Fig S5). We found that the effect appears unchanged for model-free estimates
343 of the stimulus prediction error, which rules out the possibility of a model artefact.

Discussion

Different behavioral phenomena—learning from prediction errors and biased risk-preferences --are attributed to the same neuromodulator, dopamine. Using a task where reward prediction errors are immediately followed by decisions that involve risk, we showed that reward prediction errors and the probability of risk-taking are positively correlated: positive reward prediction errors induce risk seeking, negative ones inhibit it. In particular, our results show that the strength of the reward prediction error (as indexed by pupil dilation) determines the effect on risk-preferences. This result suggests that the two roles of dopamine (teaching signal and risk modulator) interfere with each other. It provides evidence against the conjecture that the roles are well separated.

Decision making under uncertainty has been extensively studied in behavioral economics. One main finding in this field, codified in prospect theory, is that humans tend to be risk-averse if decisions concern gains, and risk-seeking if decisions concern losses [Kahneman, 2013]. However, those classic findings rely on explicit knowledge about the probabilities involved in the decisions. Several more recent studies indicate that those tendencies reverse when risks and probabilities are learned from experience (i.e. by trial and error): if learning is incremental and based on feedback, humans tend to make risky decisions about gains and risk-averse decisions about losses [Wulff, 2018]. This phenomenon has been termed the description-experience gap. In cognitive neuroscience and psychology, some studies have reproduced this phenomenon [Madan, 2014], while others report risk aversion in the gain domain [Niv, 2012]. This diversity might be associated with the degree of implicitness of the knowledge that is gained during the task: [Niv, 2012] used classical bimodal reward distributions (e.g. 40 points with probability 50 %, 0 points otherwise) which participants might be able to resolve after a few trials. Here, we used a strongly random reward schedule (normal distributions, see Fig. 1B), which made anything but implicit

366 learning intractable. Our main behavioral findings (Fig. 2A and 2B) are in line with the description-
367 experience-gap, and differ to those of [Niv, 2012].

368 The dopamine-related interpretation of our results is based on previously reported causes and
369 consequences of phasic dopamine signaling. On the causes side, it is well established that changes in
370 reward anticipation, brought about by reward-predicting cues, provoke dopamine bursts that originate
371 in brain areas such as VTA, and are broadcast to structures such as the striatum and the medial
372 prefrontal cortex [Schultz, 1997][Seymour, 2004] [Pessiglione, 2006] [Niv, 2012]. In our task, such
373 prediction errors occurred at the presentation of the available options. On the consequences side, it has
374 been shown that phasic dopamine activity can affect risk-preferences in decision making [Chew, 2019].
375 Our task featured decisions between options that provided the same average reward, but different
376 spreads of the individual rewards. Choices on those trials were not biased by value differences, and
377 hence well suited to read out risk preferences. The simultaneous occurrence of these two dopamine-
378 related phenomena explains our result: risk-seeking followed positive prediction errors and risk aversion
379 followed negative prediction errors.

380 For our behavioral results, interpretations other than our dopaminergic explanation may be evoked: the
381 behavior in a similar task [Madan, 2014] was interpreted as the result of memory replay: experiences
382 ("Obtained reward X after choosing option Y") might not only be used for immediate value updates but
383 might also be stored in a memory buffer. This buffer can then be used for offline learning from past
384 experiences in times of inactivity, such as during the inter-trial interval. It was proposed by [Madan,
385 2014] that experiences are more likely to enter the buffer if they are extreme. If entering the buffer is
386 biased in this way, then so are the values learned from replaying those experiences. In our task,
387 extreme might mean that the reward was extremely high or low. The corresponding bias would drive
388 choice towards the stimuli that produce the highest rewards, and away from those that produce the
389 lowest, and thereby lead to a pattern similar to the one we observed.

Which theory is closer to the truth? It is difficult to compare the memory theory directly to prediction-error induced risk-seeking; it is unclear how to obtain trial-by-trial choice predictions from the memory model, which rules out a formal model comparison. Indeed, the memory model has so far only been fitted to and assessed based on summary statistics of a large collection of trials. Further, the memory model has so far not been equipped with a mechanistic underpinning and was therefore not validated on physiological variables such as pupil dilation. In contrast, prediction-error induced risk-seeking can be fitted trial-by-trail, allowing it to make predictions not only about summary statistics but about the evolution of preferences during the task as well as about the immediate impact of extreme events such as large prediction errors. The corresponding latent variables can be correlated with physiological variables, proving that they can explain aspects of pupil dilation in addition to behavior (Fig. 3C and 3D).

If one interprets our results as resulting from dopaminergic interaction, one is forced to give up on the idea that direct and indirect dopaminergic effects are strictly separated. This conclusion is consistent with other recent findings: it has been shown that phasic dopamine correlates with motivational variables [Hamid, 2016] and movement vigor [da Silva, 2018] just as well as with reward prediction errors. These findings cast doubt on the separation into tonic and phasic and on separations in general.

In summary, our findings show that there is an interaction between prediction errors and risk seeking that matches what one would expect from dopaminergic interactions. We further show that this effect is detectable even on the individual level in a sizable part of our group, and that between-participant variability in behavior can be linked to differences in pupil responses—the stronger the pupil response to stimulus prediction errors, the stronger the prediction error induced risk seeking.

Methods

Participants

We tested 30 participants (15 female, median age: 26, range: 18-42). Our participants did not suffer from visual, motor or cognitive impairments. They were recruited and tested voluntarily, all experimental procedures were approved by the local ethics committee. Our results are based on 27 of the 30 participants. Three participants were excluded from the analysis due to their failure to understand the task. We evaluated the participants' understanding of the task by scoring their preferences in mixed-mean choices during the second half of the blocks. Participants were included in the analysis if they chose the high-valued option in more than 70 % of those trials (Fig S1).

Logistic regressions

Logistic regressions were conducted using mixed-effects modelling. The target variable y was defined as $y = 1$ if the risky option was chosen and $y = 0$ else. The predictors of interest were the prediction errors $\delta_{Stimulus}$ and $\delta_{Outcome}$ that preceded the choice. We further included $Q_{risky\ option} - Q_{save\ option}$ as a predictor to control for residual value differences. Individual differences were accounted for by a random intercept and random slopes for each predictor. The p-values we report for single predictors were obtained from chi-squared tests on likelihood ratio statistics. Those were computed through comparisons between the fit with all predictors included and the fit without the predictor of interest (but with the respective random slope).

Models

The RW model as well as the PEIRS model feature a softmax choice rule:

$$P(choice = i) = \frac{e^{\beta V_i}}{\sum_j e^{\beta V_j}}$$

The models differ in how those subjective values V are constructed. In the RW model, the subjective value of an option is simply the learned value of this option: $V_{RW,i} = Q_i$. In the PEIRS model, the subjective value is determined according to Eq. 1. For both the PEIRS and the RW model we set the initial value Q_0 to the empirical mean of 50. For the PEIRS model the initial value of the spread S_0 is left as a free parameter.All in all, the RW model features two free parameters (α_Q, β), while the PEIRS model features five ($\alpha_Q,\ \alpha_S, \beta, \gamma, S_0$).

Model fit, comparison and regularization

Fits and model comparisons were performed using the VBA toolbox [Daunizeau, 2014]. This toolbox implements a Variational Bayes scheme. It takes a set of measurements, a generative probabilistic model that describes how the measurements arise (which usually contains some latent, i.e. unobserved, variables) and prior distributions over the model parameters as input, and outputs among other things an approximate posterior distribution over model parameters, an approximate posterior distribution over the latent variables, and an upper bound for the model evidence. We fitted both models to each participant.

Our model comparison is based on the approximate model evidences $L(model|data)$ that the toolbox provides. Assuming that a participant generated data according one of the models, the probability of that participant using model m is given by $p(m|data) = \frac{L(m|data)}{\sum_{m'} L(m'|data)}$ (see the documentation of the VBA toolbox or [Stephan, 2009] for reference). We use these probabilities as an index of effect strength.

We estimate parameters using the posterior distributions over parameters that the toolbox outputs. Point estimates of parameters are obtained by computing the expected value of the posterior distributions. The same procedure is applied for latent variables, such as values and prediction errors.

The questions we pursue in this study involve physiological factors, such as dopamine and pupil dilation. To be useful to answer our questions and make valid predictions, it is important that our models operate in a physiologically plausible regime. One important requirement was that strong, systematic prediction errors should only occur during the learning phase at the beginning of each block, and not persist after choice behavior has stabilized. We found that our models did not fulfil this requirement by default, and hence introduced a regularization: from trial 61 onwards, we penalized prediction errors by introducing a prior centered around zero. This was implemented by adding an additional observed variable which was normally distributed around the outcome prediction error: $\delta_{outcome}^{observed} \sim N\left(\delta_{outcome}, \sigma^2\right)$. We then provided the model with "measurements" $\delta_{outcome}^{observed} = 0$ for trials 61 to 100. During model inversion, those "observations" penalized $\delta_{outcome}$ that differed strongly from zero. This regularization applies to both the RW and the PEIRS model.

Pupillometry

We recorded time series of pupil diameters for every trial, using an EyeLink 1000 system. The raw measurements where preprocessed (smoothing, blink correction) using standard methods [Manohar, 2019]. Then, the traces were aligned to the relevant temporal markers (stimulus onset, or reward onset). We used the mean over 500 ms prior to the alignment point to define a trial-wise baseline. All traces were divided and shifted by that baseline, resulting in traces reflecting the relative change of pupil diameter after the alignment point. Finally, traces were downsampled to 10 Hz.

To uncover the pupil response to the stimulus prediction error, we aligned the pupil time courses at stimulus onset. After stimulus onset, participants would eventually make a choice (with variable delay, the median reaction time was 0.86 s) and receive a reward (with a 1 s delay) after their choice. Since the reward or the resulting outcome prediction error might confound our regression analysis, we censored out all data after reward presentation. This means that the number of observations on which

475 regressions can be based rapidly declines after the median reward presentation time, which is at 1.86 s
476 after stimulus onset. Estimates obtained later are increasingly unreliable, since they are based on
477 insufficient data. We hence conducted our analyses for the interval 0 s to 1.9 s after stimulus onset. This
478 allows us to obtain reliable estimates of the statistics, while still avoiding confounding effects related to
479 reward presentation.

480 To test whether the pupil responses to the prediction errors are statistically significant, we needed to
481 perform a test in a single time-point corresponding to the largest effect. To avoid circularity, the time of
482 peak effect was identified using a leave-one-out method: We first calculated time-series of regression
483 weights for each participant individually. Then, for each participant, we determined the peak effect
484 strength of the response. To achieve this without introducing bias, we temporarily excluded the
485 participant in question and determined the time bin in which the responses of the other participants
486 where most significant. This was achieved by executing t-tests on the response strengths in each time
487 bin, and selecting the bin with the smallest p-value. We then took the left-out participant's response
488 strength from that time bin, considering it their response strength at the peak of the group response. In
489 a final step, we pooled all those individual response strengths at peak effect and used a t-test to check
490 whether they deviated significantly from zero.

Acknowledgements

This work has been supported by MRC grants MC_UU_12024/5, MC_UU_00003/1 and BBSRC grant BB/S006338/1 held by RB, an MRC clinician scientist fellowship MR/P00878X to SGM and studentships from the MRC (MR/K501256/1 and MR/N013468/1) and St John's College to JG.

References

[Schultz, 1997] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

[Rescorla, 1972] Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory, 2, 64-99.

[Berridge, 2012] Berridge, K. C. (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. *European Journal of Neuroscience*, *35*(7), 1124-1143.

[Niv, 2007] Niv, Y. (2007). Cost, benefit, tonic, phasic: what do response rates tell us about dopamine and motivation?. *ANNALS-NEW YORK ACADEMY OF SCIENCES*, *1104*, 357.

[Berke, 2018] Berke, J. D. (2018). What does dopamine mean? *Nature neuroscience*, *21*(6), 787.

[Pessiglione, 2006] Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042.

[Chew, 2019] Chew, B., Hauser, T. U., Papoutsi, M., Magerkurth, J., Dolan, R. J., & Rutledge, R. B. (2019). Endogenous fluctuations in the dopaminergic midbrain drive behavioral choice variability. *Proceedings of the National Academy of Sciences*, *116*(37), 18732-18737.

[St Onge, 2009] St Onge, J. R., & Floresco, S. B. (2009). Dopaminergic modulation of risk-based decision making. *Neuropsychopharmacology*, *34*(3), 681.

[Preuschoff, 2011] Preuschoff, K., t Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in neuroscience*, *5*, 115.

[Palminteri, 2017] Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, *21*(6), 425-433.

[Madan, 2014] Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering thebest and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic bulletin & review*, *21*(3), 629-636.

[Hamid, 2016] Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., ... & Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature neuroscience*, *19*(1), 117.

[Engelhard, 2019] Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., ... & Witten, I. B. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, 1.

[Daunizeau, 2014] Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology, 10*(1), e1003441.

[Zénon, 2019] Zénon, A. (2019). Eye pupil signals information gain. *Proceedings of the Royal Society B*, *286*(1911), 20191593.

[Browning, 2015] Browning, M., Behrens, T. E., Jocham, G., O'reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature neuroscience*, *18*(4), 590.

[Lawson, 2017] Lawson, R. P., Mathys, C., & Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature neuroscience*, *20*(9), 1293.

[Cavanagh, 2014] Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, *143*(4), 1476.

[Mikhael, 2016] Mikhael, J. G., & Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS computational biology*, *12*(9), e1005062.

[Moeller, 2019] Möller, M., & Bogacz, R. (2019). Learning the payoffs and costs of actions. *PLoS computational biology*, *15* (2), e1006285.

[Voon, 2006] Voon, V., Hassan, K., Zurowski, M., Duff-Canning, S., De Souza, M., Fox, S., ... & Miyasaki, J. (2006). Prospective prevalence of pathologic gambling and medication association in Parkinson disease. *Neurology*, *66*(11), 1750-1752.

[Weintraub, 2010] Weintraub, D., Koester, J., Potenza, M. N., Siderowf, A. D., Stacy, M., Voon, V., ... & Lang, A. E. (2010). Impulse control disorders in Parkinson disease: a cross-sectional study of 3090 patients. *Archives of neurology*, *67*(5), 589-595.

[Gallagher, 2007] Gallagher, D. A., O'Sullivan, S. S., Evans, A. H., Lees, A. J., & Schrag, A. (2007). Pathological gambling in Parkinson's disease: risk factors and differences from dopamine dysregulation. An analysis of published case series. *Movement disorders: official journal of the Movement Disorder Society*, *22*(12), 1757-1763.

553 [Niv, 2012] Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a
554 risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551-
555 562.

556 [Kahneman, 2013] Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under
557 risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127).

558 [Wulff, 2018] Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of
559 two modes of learning and the description-experience gap. *Psychological bulletin*, *144*(2), 140.

560 [Jang, 2019] Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction
561 errors during decision-making strengthen memory encoding. Nature human behaviour, 1.

562 [Stephan, 2009] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian
563 model selection for group studies. Neuroimage, 46(4), 1004-1017.

564 [da Silva, 2018] da Silva, J. A., Tecuapetla, F., Paixão, V., & Costa, R. M. (2018). Dopamine neuron activity
565 before action initiation gates and invigorates future movements. Nature, 554(7691), 244.

566 [Seymour, 2004] Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., ... &
567 Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. Nature,
568 429(6992), 664.

569 [Manohar, 2019] Manohar, S. G. (2019, October 19). Matlib: MATLAB tools for plotting, data analysis,
570 eye tracking and experiment design (Public). https://doi.org/10.17605/OSF.IO/VMABG

Reward prediction errors induce risk-seeking

Supplementary Materials

Moritz Moeller*1, Jan Grohn*2, Sanjay Manohar**12+, Rafal Bogacz**1.

*: equal contributions

**: equal contributions

1: Nuffield Department of Clinical Neurosciences, University of Oxford

2: Department of Experimental Psychology, University of Oxford

+: corresponding author (sanjay.manohar@psy.ox.ac.uk)

Performance evaluation

We assessed individual performances post-hoc, using the proportion of correct choices after trial 60 as the criterion. The data are provided in Fig S1.

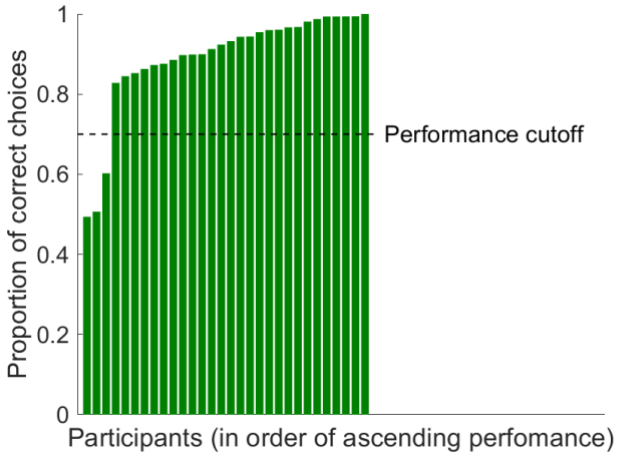


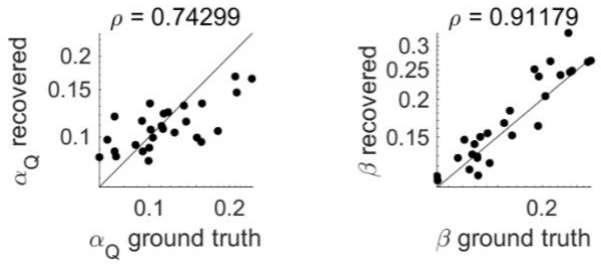
Fig S1: Individual performance evaluation. Every bar represents one participant. The height of the bar indicates the proportion of correct choices (i.e. choosing the high-valued option over the low-valued option) in the last 60 trials of each block. Three participants (corresponding to the first three bars in this graph) fell below our cutoff of 70 % and were hence excluded from the study.

Parameter recovery

To assess the reliability of the parameter estimates produced by our fitting procedure, and hence build confidence in the conclusions based on the fits, we conducted a parameter recovery analysis for both models. For each model and participant, we used the posterior distributions over parameter space obtained from the fit to get point estimates of the parameters that best describe the recorded behavior. The resulting parameter set was then used to run a simulation, aiming to produce simulated data with characteristics like those of the empirical data. As a next step, we fitted the same model that was used

for simulation to the simulated data, and again obtained estimates of the parameters, which could now be compared with the ground truth (the parameters that were used to simulate, and that were supposed to be recovered). For both models, we could robustly recover all parameters with minimal distortion (Fig S2).

A



B

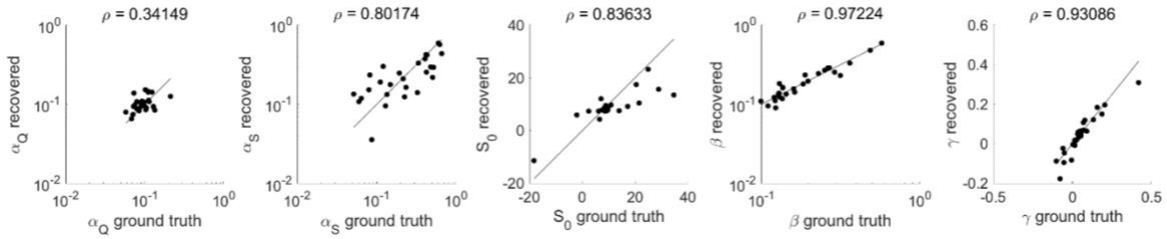


Fig S2: Parameter recovery for reinforcement learning models. A) RW model. Each panel corresponds to one parameter of the RW model. The x-axes correspond to the ground truth values (those used for the simulations), the y-axes correspond to the recovered parameters extracted from fits to the simulated data. Every dot corresponds to one participant. Above the plot, we report the correlation coefficient between the actual and the recovered parameters across the population. The black line indicates equality, i.e. $x = y$. B) Parameter recovery for the PEIRS model. Same conventions as in A).

We conclude that our models do not suffer from ambiguity or under-determination/over-parametrization. This means that both estimates of parameters and latent variables are to be taken as meaningful, unambiguous quantities.

Might reward predictions cause risk preferences?

We showed that seemingly irrational risk preferences can be explained by reward prediction errors (RPEs, changes in reward expectation) that occur immediately before the choice. A confounding variable in this analysis is the reward prediction (RP) itself: it could be that the anticipation of high rewards causes risk-seeking, while anticipating low rewards causes risk-aversion. If so, it would still seem as if RPEs cause risk preferences, since RPEs and RPs are highly correlated in our experiment (Fig S3 A).

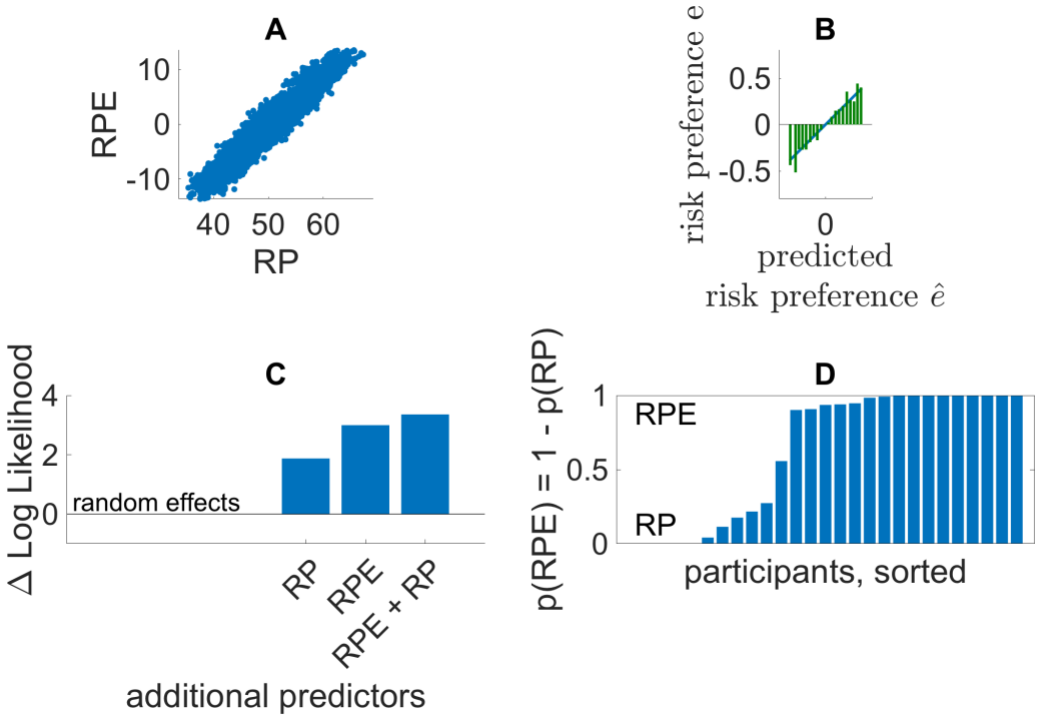


Fig S3: Reward predictions as a confounding variable. A) Correlation between trial-wise reward prediction errors (RPE) and the reward predictions (RP) that followed stimulus presentation. These prediction errors were estimated by fitting a standard RW model to each participant individually. B)

Model fitted to residual preferences. Risk preferences were predicted with a linear mixed effects model. Preference was modelled as a linear function of RP and RPE. Individual differences in regression coefficients were modelled as random effects of subject ID ($e \sim 1 + RP + RPE + (1 + RP + RPE|ID)$, the last term corresponds to individual differences in slopes and intercept). Residual preferences were binned according to predicted preferences, averaged per bin and are displayed as green bars. The predicted preferences are represented by the blue line. C) Relative log likelihoods of models with different sets of predictor variables. The bars indicate the increase in likelihood relative to a baseline model that used only random effects. D) Model comparison results on participant level. The bars represent the likelihood that the data recorded from a given participant was generated from the PEIRS model rather than the PIRS model.

To test whether risk preferences are due to RPEs rather than RPs, we conducted two additional analyses. First, we used linear models to test which signal—RPEs or RPs—is a better predictor for risk preferences. We started by extracting preferences e that could not be explained by standard learning effects. This was done by predicting choices $\hat{c}$ on matched-mean trials with a standard RW model, and subtracting them from the measured choices c ($c = 1$ when the risky option was chosen, and $c = 0$ otherwise). The residual preferences $e = c - \hat{c}$ contain the risk preferences we seek to explain. Next, we used linear models to predict the residual preferences e. As predictors, we considered RPE, RP and the corresponding random effects. Taken together, those signals partially explain the residual preferences (adjusted R^2: 0.0603, Fig S3 B). Finally, we checked how much explanatory power each signal contributes by comparing log likelihoods (LL) corresponding to different predictors. If risk preferences were due to RPEs, we should expect that 1) adding only RPEs should increase LL more than adding only RPs, and that 2) adding RPs on top of RPEs should not increase LL substantially. Point 2) specifically holds if RP does not contain additional relevant information about risk-preferences over and

above those that it shares with RPE. We found that 1) and 2) hold (Fig S3 C). This suggests that risk preferences are best explained by RPEs. In our experiment, they can also be predicted by RPs, but only because RPs are correlated with RPEs and thereby gain some of the RPEs' predictive power.

We run another analysis to corroborate this result: to test whether RPs could explain our data better than RPEs, we defined another trial-by-trial model (PIRS, "Predictions Induce Risk Seeking") similar to the PEIRS model. PIRS is identical to PEIRS with one exception: it is the prediction and not the prediction error that interacts with risk in the decision rule (Eq. 1 in the main text). We then performed a model comparison between PIRS and PEIRS. We found that our data is more likely to be generated by PEIRS than by PIRS (odds ratio about 3:1 for PEIRS), and that most participants are better fitted by PEIRS (Fig S3 D). This result aligns with the result we obtained using linear models to predict residual preferences, and suggests that it is the RPE, and not the RP, that might cause risk preferences.

Nonlinear utility

To set up models such as ours, one must choose a way to relate the point score that participants are shown to the abstract reward signal that features in RL models (i.e. one must chose a utility function that maps points to reward). For our analysis, we chose a simple linear mapping. Thus, we implicitly assume that points would directly translate into reward. However, it has been shown that often, utility functions are not as simple—for example, in behavioral economics the utility of money is frequently modeled using concave functions. Crucially, nonlinear utility functions can lead to apparent risk preferences. Do our results and conclusions still hold if we drop the assumption of a linear utility function, and allow for non-linear utility curves?

To test this, we started by choosing a parametric family of utility functions. We then determined the most likely utility function for each participant by fitting a RW-model with parametric utility to their

choices. Finally, we checked whether there was still a correlation between risk preferences and preceding prediction errors after the nonlinear utility curves were considered.

To model non-linear utility curves, we chose an exponential family centered at 50 points, defined by

$$u_\kappa(x) = 50 + \frac{50}{\kappa}\left(1 - e^{-\kappa\left(\frac{x}{50}-1\right)}\right).$$

The functions are shifted such that $u(50) = 50$ for all values of κ, to keep initial values independent of κ (see Fig S4 A for some exemplars). Next, we fitted standard RW models to the choices of each participants, using $r_\kappa = u_\kappa(x)$. From this, we obtained estimates of κ for each participant. We found that almost all κ were positive (Fig S4 B), suggesting concave mappings from points to subjective value for almost all participants (Fig S4 C). Finally, we performed the same analysis as in Fig 2C in the main manuscript, checking whether there was a correlation between the likelihood of risk-seeking and the magnitude of the prediction error immediately before the choice. We found a significant correlation similar to the corresponding curve based on linear utility (Fig S4 D). This suggests that our findings are robust, and hold even when the assumption of linear utility is relaxed.

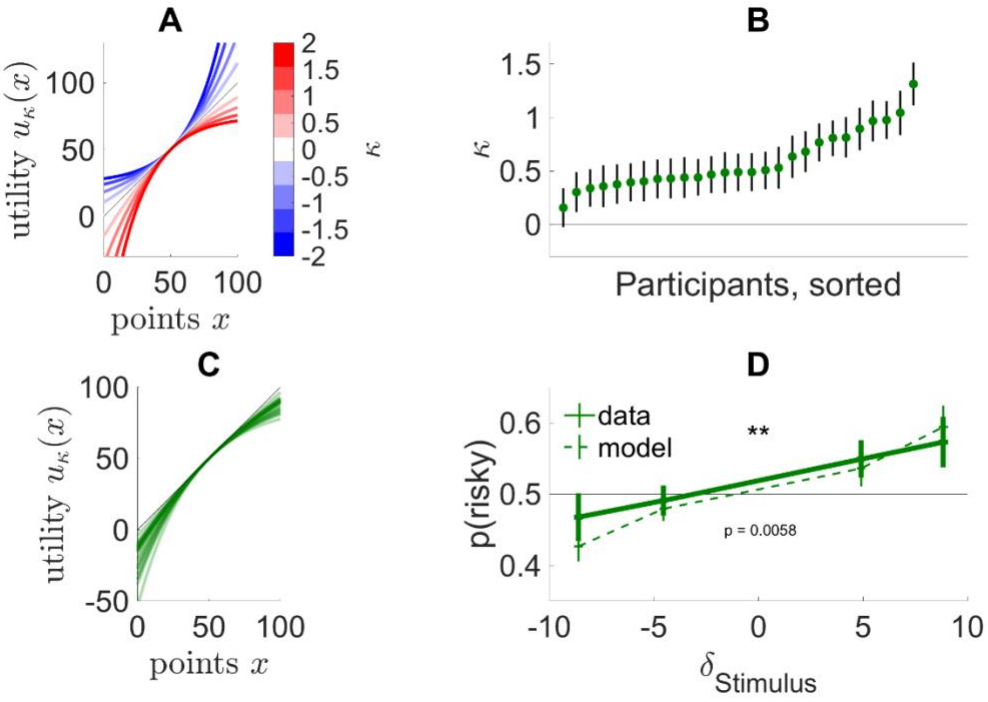


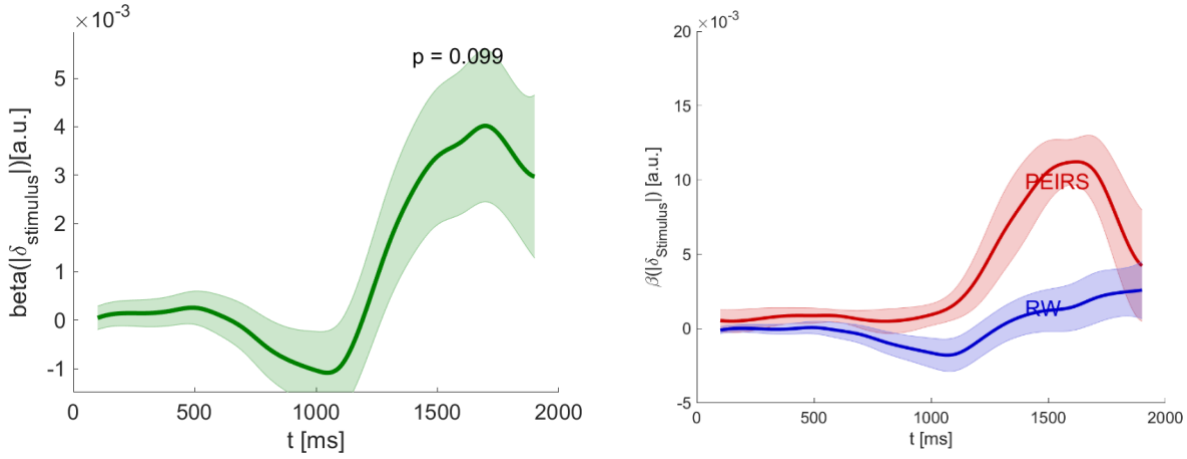
Fig S4: Robustness under nonlinear utility mappings. A) Family of utility functions. The parameter κ which controls the curvature is represented by color. B) Estimates for κ. Mean and standard deviation of the posterior distribution of κ are indicated by a green dot and black lines. The statistics for the posterior distribution are provided for each participant individually, ordered by the mean of the posterior. C) Estimated utility curves. Posterior estimates in C) were converted in utility functions and superimposed. Each green line corresponds to the most likely utility function of one participant. The lines are transparent to aid visibility. D) Similar to Fig 2C in the main text, but with stimulus prediction errors and values taken form a RW model with the non-linear utility functions depicted in C).

Ground truth pupillometry

The predictor variables of our pupil-related regression analyses (the absolute stimulus prediction error and the absolute outcome prediction error) are model-dependent variables. One might thus suspect that the correlations displayed in Fig 4C and Fig 4D might be spurious: pupil responses are defined with

125 respect to a model variable (the stimulus prediction error) and are predicted by another model-
126 depended quantity (the logarithmic odds ratio). To rule out potential confounding effects, and to make
127 sure that the pupil responses do in fact provide an external validation of our behavioral modelling, we
128 conducted the same analyses based on the ground truth (model-free) prediction error instead of the
129 model-based stimulus prediction error (Fig S5).

A **B**



C

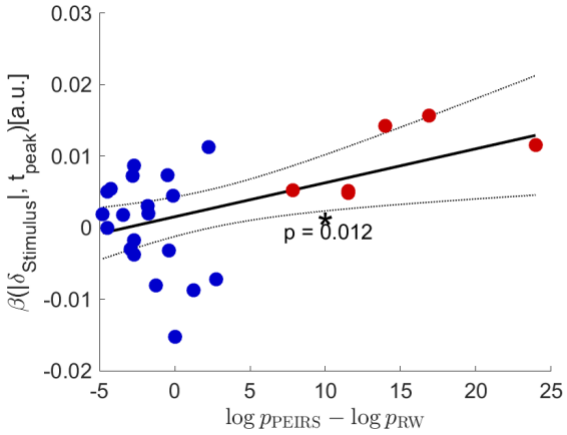


Fig S5: Ground truth pupillometry results. A, B and C) same as Fig 4 A, C and D) but using model-independent ground truth prediction error instead of the prediction error extracted from the model fit.

We found that all results described in the main text hold similarly if the analysis is conducted based on the ground truth instead of the model-based variable. Our reasoning is thus not circular, and the result are not due to modelling artifacts.

The ground truth stimulus prediction error is defined as

$$\delta_{Stimulus,GT} = E_{\ option\ shown}(R) - E_{\ all\ option}(R) = E_{\ option\ shown}(R) - 50$$

Since $E_{\ option\ shown}(R)$ could only take the values 40, 50 or 60 by experimental design, $\delta_{Stimulus,GT}$ could only take the values -10, 0 or 10, and the predictor variable $|\delta_{Stimulus,GT}|$ could only take the values 0 or 10. Therefore, the ground-truth prediction error used for the control analyses is equivalent to a contrast between matched-mean and different-mean conditions.