Diffusion modeling reveals reinforcement learning impairments in gambling disorder that are linked to attenuated ventromedial prefrontal cortex value representations.

Antonius Wiehler[1,2], Jan Peters[1,3]

[1]Department of Systems Neuroscience, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany

[2]Institut du Cerveau et de la Moelle épinière (ICM), INSERM U 1127, CNRS UMR 7225, Sorbonne Universités Paris, France.

[3]Department of Psychology, Biological Psychology, University of Cologne, Cologne, Germany

Contact: jan.peters@uni-koeln.de

Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (PE 1627/5-1 to J.P.). J.P. and A.W. designed the study. A.W. acquired the data. J.P. and A.W. analyzed the imaging data. J.P. performed the modeling and wrote the paper. A.W. provided revisions. J.P. supervised the project.

Abstract

Gambling disorder is associated with deficits in classical feedback-based learning tasks, but the computational mechanisms underlying such learning impairments are still poorly understood. Here, we examined this question using a combination of computational modeling and functional resonance imaging (fMRI) in gambling disorder participants (n=23) and matched controls (n=19). Participants performed a simple reinforcement learning task with two pairs of stimuli (80% vs. 20% reinforcement rates per pair). As predicted, gamblers made significantly fewer selections of the optimal stimulus, while overall response times (RTs) were not significantly different between groups. We then used comprehensive modeling using reinforcement learning drift diffusion models (RLDDMs) in combination with hierarchical Bayesian parameter estimation to shed light on the computational underpinnings of this performance impairment. In both groups, an RLDDM in which both non-decision time and response threshold (boundary separation) changed over the course of the experiment accounted for the data best. The model showed good parameter recovery, and posterior predictive checks revealed that in both groups, the model reproduced the evolution of both accuracy and RTs over time. Examination of the group-wise posterior distributions revealed that the learning impairment in gamblers was attributable to both reduced learning rates and a more rapid reduction in boundary separation over time, compared to controls. Furthermore, gamblers also showed substantially shorter non-decision times. Model-based imaging analyses then revealed that value representations in gamblers in the ventromedial prefrontal cortex were attenuated compared to controls, and these effects were partly associated with model-based learning rates. Exploratory analyses revealed that a more anterior ventromedial prefrontal cortex cluster showed attenuations in value representations in proportion to gambling disorder severity in gamblers. Taken together, our findings reveal computational mechanisms underlying reinforcement learning impairments in gambling disorder, and confirm the ventromedial prefrontal cortex and as a critical neural hub in this disorder.

Introduction

Gambling disorder is a behavioral addiction that shares core neurobiological features with substance-use-disorders[1]. Neuroimaging studies have often reported functional and to a lesser degree structural alteration in regions of the reward system in gamblers[2], in particular in the ventral striatum and the ventromedial prefrontal cortex (vmPFC), regions implicated in reinforcement learning and reward valuation [3,4]. However, as summarized extensively in a recent review[2], the directionality of dysregulation in these regions in gambling disorder is mixed, with studies reporting both increases and decreases in responses. These inconsistencies likely depend on task-specific and contextual effects [5–7], as extensively discussed previously[8–10].

Behaviorally, gambling disorder is characterized by maladaptive decision-making in a range of laboratory tasks. This includes increased temporal discounting (i.e. an increased preference for smaller-sooner over larger-later rewards)[6,11–14], and increased risk-taking[6,11,15]. Gamblers also show impairments in feedback-based learning. For example, on the Wisconsin Card Sorting Test (WCST), gamblers require more trials to adapt to changes in the relevant stimulus dimension, that is, they show increased perseveration behavior [16–21]. Likewise, gambling disorder is associated with impairments in probabilisitic reversal learning [16,22]. Although there is some heterogeneity across studies with respect to reversal learning impairments in gambling disorder [23], generally the directionality of these effects is quite consistent in the literature as highlighted in a recent meta-analysis [23].

Impairments in feedback-based learning might be of particular relevance in the context of gambling disorder. Modern Electronic Gaming Machines (EMGs) are designed to encourage "rapid and continuous play"[24], e.g. by combining high event frequencies with salient audio-visual reinforcers [25]. Such machine design features may then, in interaction with cognitive factors such as those outlined above (e.g. impaired reward learning, reduced self-control, increased risk-taking and perseveration)[25] give rise to a particularly rapid escalation of EMG gambling compared to more traditional forms of gambling [26].

From a computational reinforcement learning (RL) perspective[27], reversal learning impairments could arise due to changes in different processes. On the one hand, impairments could be due to response perseveration, where previous actions are repeated irrespective of learning or option values. Additionally, however, reinforcement learning in volatile environments requires balancing exploration (choosing options with unknown value for information gain) and exploitation (choosing options with known value for reward maximization)[27,28]. We have recently shown using computational modeling in a restless four

armed bandit task[29], that gamblers rely less on "directed exploration"[30], that is, the "strategic" exploration of choice options with high estimated uncertainty. Reversal learning impairments could therefore also be due to reduced exploration. Finally, slower learning in volatile environments (i.e. reduced learning rates) could also contribute to the observed impairments in gambling disorder.

Traditionally, RL models account for trial-wise categorical decisions by assuming that choices stochastically depend on the values of the available options. These values are learned via e.g. the delta learning rule, where values are updated based on reward prediction errors [27,31]. This learning rule is then combined with a choice rule such as softmax action selection[27]. Under softmax action selection, the degree to which choices depend on the learned values via a sigmoid function, where the slope parameter indexes the "value-dependency" or "stochasticity" of decisions with respect to a given learning model. However, such choice rules are agnostic with respect to the computational processes underlying changes in "stochasticity"[32].

For this reason, researchers have begun to move beyond examining binary decisions, instead taking the distribution of choice response time (RT) into account. Sequential sampling models such as the drift diffusion model [33,34] (DDM) are widely used in perceptual decision-making. These models assume that choices arise from a noisy evidence accumulation process that terminates as soon the accumulated evidence exceeds a threshold. In its simplest form, the DDM has three free parameters. The drift rate v reflects the average rate of evidence accumulation. The boundary separation parameter α governs the response threshold, and thus controls the speed-accuracy trade-off – a smaller boundary separation emphasizes speed over accuracy, whereas the reverse is true for a larger boundary separation. The non-decision time τ models perceptual and/or motor components or the RT that are unrelated to the evidence accumulation process. If one assumes that the quality of evidence in favor of a decision option is reflected in the trial-wise drift rate (via some *linking function*[35]), the DDM can be used to model trial-wise decisions in reinforcement learning[32,36,37] and value-based decision-making more generally[38–40]. The benefits of such an approach over e.g. traditional softmax choice rules are both of technical and theoretical nature. In technical terms, inclusion of RTs during model estimation has been shown to improve parameter recovery and reliability [37,41], which is of particular relevance when the number of observations is small, e.g. when working with clinical samples. In theoretical terms, a combined reinforcement learning DDM (RLDDM) is not only a more complete model, as it accounts for both binary decisions and RTs, but also allows for a more fine-grained analysis of the dynamics underlying the decision process. For

example, choices could be more random or stochastic due to a reduced impact of values on drift rates or due to a more liberal (smaller) boundary separation. RLDDM models can dissociated these different possibilities, whereas softmax choice rules only contain a single "stochasticity" parameter, and thus cannot disentangle effects of value-dependency from threshold changes. We have recently shown that DDMs can reveal alterations in decision-making following prefrontal cortex damage[38] and following pharmacological dopamine challenges[39,40]. Others have shown that RLDDMs can reveal pharmacological effects on learning in attention-deficit hyperactivity disorder [32] and effects of reward magnitude on reinforcement learning [36].

Here we extended these recently developed modeling schemes to comprehensively analyze the computational basis of reinforcement learning deficits in gambling disorder. Gamblers and matched controls performed a simple stationary reinforcement learning task[42] while brain activity was measured using functional magnetic resonance imaging (fMRI). Modeling using hierarchical Bayesian parameter estimation revealed that an extension of previously proposed RLDDMs[32,35,36] that allowed both non-decision times and boundary separation to vary over the course of learning accounted for the data best, showed good parameter recovery, and in both groups accounted for the evolution of accuracy and response times over the course of learning. Finally, fMRI revealed that vmPFC value representations in gamblers were attenuated, and these effects were in part associated with performance as reflected in model parameters as well as clinical symptom severity.

Methods

Participants

In total, n=23 gamblers and n=19 matched controls participated in the study. The demographic and clinical characteristics of the groups have been reported in detail elsewhere[30]. In short, groups were matched on age (M[SD] gamblers: 25.91 [6.47], controls: 26.58 [6.52], t=-.33, p=.74), gender (all participants were male), smoking according to the Fagerström Test for Nicotine Dependence (FTND)[43] (M[SD] gamblers: 2.14 [2.58], controls: 2.21 [2.18], t=-.10, p=.92), alcohol use according to the Alcohol Use Disorders Test (AUDIT)[44] (M[SD] gamblers: 6.09 [7.14], controls: 6.84 [4.91], t=-.40, p=.69) and education (school years M[SD] gamblers: 11.64 [1.77], controls: 11.79 [1.40], t=-.31, p=.76). Gamblers scored higher on depression according to the Beck Depression Inventory (BDI-II)[45] (M[SD] gamblers: 15.41 [11.41], controls: 8.47 [8.46], t=2.26, p=.03).

Sixteen gamblers fulfilled four or more DSM-5 criteria for gambling disorder (previously classified as pathological gamblers). Seven gamblers fulfilled one to three DSM-5 criteria (previously classified as problem gamblers). Addiction severity was further characterized using the German "Kurzfragebogen zum Glücksspielverhalten" (KFG)[46] (M[SD] gamblers: 25.90 [14.15], controls: 0.58 [0.32], t=8.55, p<.001) and the South Oaks Gambling Screen (SOGS)[47] (M[SD] gamblers: 8.64 [4.46], controls: 0.21 [0.54], t=8.99, p<.001). These scores were highly correlated in the gamblers (r=.95) and were thus combined for correlational analyses into a single "addiction severity" score by averaging their respective z-scores.

Due to an incidental finding during fMRI which may have impacted spatial normalization, one control participant was excluded from the imaging data analysis. This participant was however retained for all behavioral and modeling analyses. Participants were recruited via postings in local internet bulletin boards, and reported no history of neurological or psychiatric disorder except for depression. No participants were currently undergoing any psychiatric treatment. Current drug abstinence on the day of scanning was verified via a Urine drug test. All participants provided informed written consent prior to their participation, and the study procedure was approved by the local institutional review board (Hamburg Board of Physicians).

Reinforcement learning task

Following completion of our previously reported restless four-armed bandit task [30], participants had a short break inside the scanner. Then they performed 60 trials in total of a

simple reinforcement learning task[42] using two pairs of stimuli (n=30 trials per pair). Per pair, one stimulus was associated with a reinforcement rate of 80% (optimal stimulus) whereas the other was associated with a reinforcement rate of 20% (suboptimal stimulus). Options were randomly assigned to the left/right side of the screen, and trials from the two option pairs were presented in randomized order. Following presentation of the two options, participants had 3 seconds to choose one of the two stimuli via button press (see Figure 1). Participants then received binary feedback, either in the form the display of a 1€ coin (*reward* feedback, see Figure 1) or as a crossed 1€ coin (*no reward* feedback). A jitter of variable duration (2-6sec, uniformly distributed) was included following presentation of the selection feedback and following presentation of the reward feedback (see Figure 1). Prior to scanning, participants performed a short practice version of the task in order to familiarize themselves with the task and the response deadline.

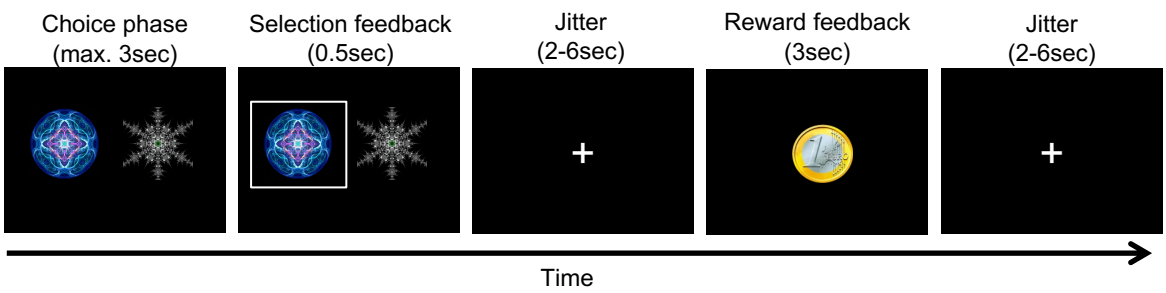


Figure 1. Illustration of a single trial from the reinforcement learning task. Stimuli were presented for a maximum of 3sec, during which participants were free to make their selection. The selection was then highlighted for 500ms, followed by a jitter of variable duration (2-6sec). Reward feedback was then presented for 3sec, followed by another jitter of variable duration (2-6sec). Stimuli consisted of two pairs of abstract fractal images (80% vs. 20% reinforcement rate), which were presented in randomized order, and participants completed 30 trials per pair.

Q-learning model

We applied a simple Q-learning model[27] to formally model the learning process. Here, participants are assumed to update the value (Q-value, Eq. 1) of the chosen action based on the reward prediction error that is computed on each trial as the difference between the obtained reward and the expected reward (Eq. 2):

$$Q_{i,t+1} = Q_{i,t} + \eta * \delta_t \quad (1)$$

$$\delta_t = r_t - Q_{i,t} \quad (2)$$

Q-values of unchosen actions remain unchanged, and all Q-values were initialized with values of .5. As learning from positive and negative feedback is thought to depend on distinct striatal

circuits [48,49], also examined models with separate learning rates (η_+, η_-) for positive vs. negative prediction errors. Learning rates were estimated in standard normal space [-3, 3] and back-transformed to the interval [0, 1] via the inverse cumulative normal distribution function.

Softmax action selection

Softmax action selection models the choice probability of the chosen action *i* on trial *t* as a sigmoid function of the Q-value difference[27] between the chosen option *a* and the unchosen option *b*:

$$P(i)_t = \frac{1}{1 + \exp\left(-[Q_a - Q_b] * \beta\right)} \qquad (3)$$

The inverse temperature parameter β models the degree to which action probabilities depend on Q-values, such that choices are random for β = 0.

Reinforcement learning drift diffusion models (RLDDMs)

We next set out to more comprehensively analyze choice dynamics underlying learning performance. To this end, we examined a set a reinforcement learning drift diffusion models[32,36] (RLDDMs) in which the DDM replaces softmax action selection as the choice rule[35]. These models can account for the full response time (RT) distributions associated with decisions, and thus provide additional information regarding the dynamics of the choice process.

We applied an accuracy coding scheme, such that the upper response boundary was defined as selection of the optimal (80% reinforced) stimulus, whereas the lower response boundary was defined as selection of the suboptimal (20% reinforced) stimulus. RTs for choices of the suboptimal option where multiplied by -1 prior to model estimation, and we discarded for each participant the fastest 5% of trials. In a null model without a learning component (DDM_0), the RT on each trial *t* is then distributed according to the Wiener First Passage Time (*wfpt*):

$$RT_t \sim wfpt(\alpha, \tau, z, v) \qquad (4)$$

Here the boundary separation parameter α regulates the speed-accuracy trade-off, such that smaller values of α lead to faster but less accurate responses. The drift rate v reflects the quality of the evidence, such that greater values of v give rise to more accurate and faster responses. Note that in this model v is constant and not affected by the learning process. The

non-decision time τ models RT components related to motor and/or perceptual processing that are unrelated to the evidence accumulation process. The starting point parameter z models a bias towards one of the response boundaries. We fixed z at .5 as options were presented in randomized order on the left vs. right side of the screen.

Following earlier work[32] we then incorporated the learning process (Equations 1 and 2) in the DDM by setting trial-wise drift rates to be proportional to the difference in Q-values between chosen and unchosen options:

$$v_t = v_{coeff} * \left(Q_{optimal} - Q_{suboptimal}\right) \quad (5)$$

In this *linking function*[35], v_{coeff} models the degree to which trial-wise drift rates scale with the value difference between options. The intuition is that as Q-value differences increase, the accuracy increases and RTs decrease. Conversely, when cue-values are similar (and thus response conflict is high) choices are both noisier and slower. Note that we also examined a non-linear mapping scheme proposed earlier[36], but these models failed to converge in our data. This is likely attributable to the somewhat lower trial numbers in the present study compared to previous implementations of non-linear drift rate scaling [36,38,39]

Time-varying boundary separation and non-decision-times

We finally examined two further extensions of the RLDDM that might capture additional RT effects unrelated to the learning process. This was based on the observation that in the gamblers, RTs decreased over the course of the experiment, but this effect appeared to be only in part attributable to learning, such that models with constant α and τ did not reproduced RT changes in the gamblers (see posterior predictive checks below). Therefore, we examined whether allowing boundary separation[32,36] and/or non-decision-time to vary over the course of the experiment according to a power function could account for these effects, such that

$$\alpha_t = \alpha_0 * t^{\alpha_{exp}} \quad (6)$$

for the case of boundary separation that varies across trials t, and

$$\tau_t = \tau_0 * t^{\tau_{exp}} \quad (7)$$

for the case of non-decision times that vary across trials t. On trial 1, parameters correspond to α_0 or τ_0, and then change over trials according to a power function with exponents α_{exp} or τ_{exp}. The first case (Eq. 6) captures the idea that, over time, participants response caution might decrease due to increasing impatience, fatigue or boredom with the task, and they might therefore accept a decreasing amount of evidence before committing to a decision. The second case (Eq. 7) corresponds to the idea that motor and/or perceptual processes might improve over time, such that participants respond generally faster over the course of the

experiment. Conversely, it would also be conceivable that motor and/or perceptual processes slow down due to fatigue.

Table 2. Overview of priors for group means.

Parameter	Group-level prior (μ)
$\boldsymbol{\alpha}_0$	*Uniform (.01, 5)*
$\boldsymbol{\alpha}_{\text{exp}}$	*Uniform (-3, 3)*
$\boldsymbol{\tau}_0$	*Uniform (0.1, 2)*
$\boldsymbol{\tau}_{\text{exp}}$	*Uniform (-3, 3)*
$\boldsymbol{v}_{coeff}$	*Uniform (-100, 100)*
$\boldsymbol{\eta}_+, \boldsymbol{\eta}_-$	*Uniform (-3,3)*

Hierarchical Bayesian models

Models were fit to all trials from all participants using a hierarchical Bayesian modeling approach with separate group-level Gaussian distributions for all parameters for controls and gamblers. Posterior distributions were estimated using Markov Chain Monte Carlo as implemented in the JAGS software package[50] (Version 4.3) using the Wiener module for JAGS that implements the Wiener First Passage Time[51] distribution, in combination with Matlab (The MathWorks) and the *matjags* interface (https://github.com/msteyvers/matjags). For group-level means, we used uniform priors defined over numerically plausible parameter ranges (see Table 2). For group-level precisions, we used Gamma distributed priors with shape and scale parameters of .001.

For each model and group, we ran 2 chains with a burn-in period of 50.000 samples and thinning factor of 2. 10.000 additional samples were then retained for further analysis. Chain convergence was assessed by examining the Gelman-Rubinstein convergence diagnostic $\hat{R}$, and values of $1 \leq \hat{R} \leq 1.01$ were considered as acceptable for all group-level and individual-subject parameters. Relative model comparison was performed via the Widely Applicable Information Criterion (WAIC) and the estimated log pointwise predictive density (*elpd*) [52], an approximation of the leave-one-out cross-validation accuracy of the model.

Parameter recovery simulations

We ran parameter recovery simulations to ensure that known parameters underlying the data-generating process could be recovered using our modeling procedures. For this purpose, we simulated 10.000 full data sets from the posterior distribution of the best-fitting models, randomly selected 10 of these simulated data sets, and re-fit the models to these simulated data. Parameter recovery was then assessed via two procedures. For subject-level parameters,

we examined the correlation between generating and estimated parameters across all ten simulations. For group-level means and standard deviations, we examined whether the estimated 95% highest posterior density intervals contained the true generating parameter value.

Posterior predictive checks

We also performed posterior predictive checks to ensure that the best-fitting models captured key aspects of the data, again using data sets simulated from each model's posterior distributions. We then computed for each group mean RTs and accuracies for each of 10 trial bins (averaging across 1000 randomly selected simulated data sets), and compared these model-predicted values to the observed data per group.

FMRI data acquisition

MRI data were collected on a Siemens Trio 3T system using a 32-channel head coil. Participants performed a single run of 60 trials in total (following a short break after completion of our previously reported task[30]). Each volume consisted of 40 slices (2 x 2 x 2 mm in-plane resolution and 1-mm gap, repetition time = 2.47s, echo time 26ms). We tilted volumes by 30° from the anterior and posterior commissures connection line to reduce signal drop out in the ventromedial prefrontal cortex and medial orbitofrontal cortex[53]. Participants viewed the screen via a head-coil mounted mirror, and logged their responses via the index and middle finger of their dominant hand using an MRI compatible button box. High-resolution T1 weighted structural images were obtained following completion of the cognitive tasks.

FMRI preprocessing

All preprocessing and statistical analyses of the imaging data was performed using SPM12 (Wellcome Department of Cognitive Neurology, London, United Kingdom). As in our previous study in this sample[30], volumes were first realigned and unwarped to account for head movement and distortion during scanning. Second, slice time correction to the onset of the middle slice was performed to account for the shifted acquisition time of slices within a volume. Third, structural images were co-registered to the functional images. Finally, all images were smoothed (8mm FWHM) and normalized to MNI-space using the DARTEL tools included in SPM12 and the VBM8 template.

FMRI statistical analysis

Error trials were defined as trials were no response was made, or trials that were excluded from the computational modeling during RT-based trial filtering (see above, recall that for each participant, the fastest 5% of trials were excluded, as is common practice in DDM studies).

We then set up a first-level general linear model with the following regressors: 1) the onset of the decision option presentation, 2) the onset of the decision option presentation modulated by chosen – unchosen value, 3) the onset of the decision option presentation modulated by (chosen – unchosen value)^2, 4) the onset of the feedback presentation, 5) the onset of the feedback presentation modulated by model-based prediction error, 6) the onset of the decision option presentation for error trials and 7) the onset of the feedback presentation for error trials. Option values and prediction errors were calculated using each participants' mean posterior parameter estimates of the learning rate from the final hierarchical Bayesian model with a single learning rate and modulated non-decision time and boundary separation (RLDDM4). Single-subject contrast estimates were then taken to a second-level random effects analysis using the two-sample t-test model as implemented in SPM12. At the second level, the following z-scored covariates were include: age, depression as assessed via the Beck Depression Inventory II (BDI)[45], smoking behavior as assessed via the Fagerström Test for Nicotine Dependence (FTND)[43] and alcohol use as assessed via the Alcohol Use Disorders Identification Test (AUDIT)[44].

All contrasts are shown at $p<.005$ (uncorrected) with $k>=10$ voxels for display purposes, unless stated otherwise. We performed correction for multiple comparisons for our a-priori region of interest (ROI) in ventromedial prefrontal cortex (vmPFC) using small volume FWE correction based on a 10mm sphere centered at -9/39/-6, the peak of a previous conjunction analysis across value representations for a range of different reward categories [54].

Results

Behavioral data analysis and computational modeling proceeded in the following steps. We first analyzed model-free performance measures. Next, we carried out a detailed model comparison of a set of candidate reinforcement learning drift diffusion models (RLDDMs) and identified the best-fitting models per group. We then ran parameter recovery analyses to ascertain that the true data-generating parameters could be recovered, and ran posterior predictive checks to ensure that the key patterns in the data could be reproduced by the best-fitting models. Finally, we examined the parameter posterior distributions and compared them between groups, before moving to the analysis of the fMRI data.

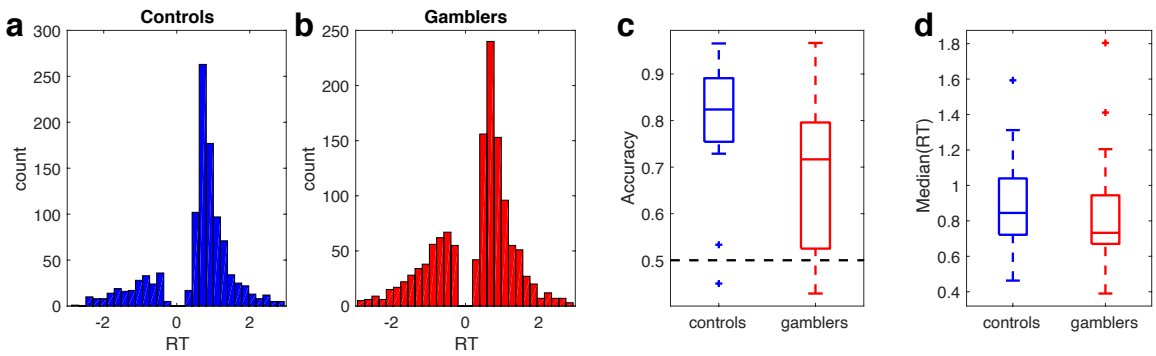


Figure 1. Response time (RT) distributions for controls (a) and gamblers (b). Here, choices of the suboptimal options are coded as negative RTs. c: Accuracy per group (chance level is 0.5). d: Median RT per group.

Model-free analysis

RT distributions per group are shown in Figure 1a and 1b, with choices of the suboptimal option coded as negative RTs. While controls selected the optimal stimulus around 80% of the time (Figure 1c, signed rank test vs. chance level: z=4.04, p=.00005, two-tailed), gamblers made around 68% correct choices (signed rank test vs. chance level: z=3.77, p=.00016, two-tailed) such that performance was significantly better in controls (rank sum test: z=2.60, p=.0092, two-tailed). The difference in median RTs (Figure 1d) between groups was not significant (rank sum test: z=1.11, p=.266, two-tailed).

Model comparison

We next compared a range of computational models (see methods section). As a reference, we first fit a null model (DDM_0) without a learning component. We then examined a set of

reinforcement learning DDMs (RLDDMs) that all included a linear mapping from Q-value differences to trial-wise drift rates [32,35] (see Eq. 5). This modeling scheme incorporates the intuition that, as learning progresses, accuracy should increase and RTs should decrease, and both effects should be greater when making easier choices, i.e. between options with larger Q-value differences. Furthermore, our model space included models with single vs. dual learning rates η (for positive vs. negative prediction errors), and models with fixed vs. modulated boundary separation α and fixed vs. modulated non-decision times τ (see Eq. 6 and 7), yielding a total of 8 RLDDMs (see Table 2).

Table 2. Model comparison of the different reinforcement learning drift diffusion models (RLDDMs), separately per group. We examined models with single vs. dual learning rates (η) and fixed vs. modulated non-decision times (τ) and boundary separation (α). In controls, the RLDDM8 fit the data best, whereas the RLDDM4 provided the best account of the data in gamblers, according to both the Widely Applicable Information Criterion (WAIC) the estimated log pointwise predictive density (elpd)[52]. These models both contained modulated terms for τ and α, but differed in whether a single (RLDDM4) or dual learning rates were included (RLDDM8).

Model	**η**	$\boldsymbol{\tau}$	$\boldsymbol{\alpha}$	**Controls**			**Gamblers**		
				WAIC	*-elpd*	*Rank*	*WAIC*	*-elpd*	*Rank*
DDM_0	-	Fixed	Fixed	1434.3	718.4	9	2227.9	1115.9	9
RLDDM 1	Single	Fixed	Fixed	1198.2	600.8	8	2104.5	1054.7	7
RLDDM 2	Single	Fixed	Power	1148.0	576.0	5	2037.5	1021.7	4
RLDDM 3	Single	Power	Fixed	1164.2	585.5	7	2045.1	1027.5	5
RLDDM 4	Single	Power	Power	1136.3	571.3	4	**2013.9**	**1011.4**	**1***
RLDDM 5	Dual	Fixed	Fixed	1158.8	581.8	6	2108.1	1056.7	8
RLDDM 6	Dual	Fixed	Power	1110.4	558.0	2	2031.1	1018.7	3
RLDDM 7	Dual	Power	Fixed	1122.7	562.2	3	2046.5	1028.7	6
RLDDM 8	Dual	Power	Power	**1093.9**	**550.8**	**1***	2019.0	1014.7	2

Model comparison was performed using the Widely Applicable Information Criterion (WAIC) and the estimated log pointwise predictive density (-elpd)[52]. As can be seen from Table 2, in both groups the data were generally better accounted for by models in which both α and τ were modulated according to a power function. However, the model ranking differed with respect to learning rates. In controls, the data were best accounted for by a model with separate learning rates for positive and negative prediction errors (RLDDM8), whereas in gamblers, the single learning rate model (RLDDM4) provided the best account of the data. This was the case for both the WAIC and the estimated log pointwise predictive density[52] (see Table 2).

Table 2. Single-subject parameter recovery results (Pearson correlations between generating and estimated parameters across 10 simulations) for the RLDDM4 and RLDDM8

Parameter	RLDDM4	RLDDM8
α_0	.70	.59
α_{exp}	.73	.78
τ_0	.88	.90
τ_{exp}	.65	.64
ν_{coeff}	.57	.69
η	.83	-
η_+	-	.71
η_-	-	.69

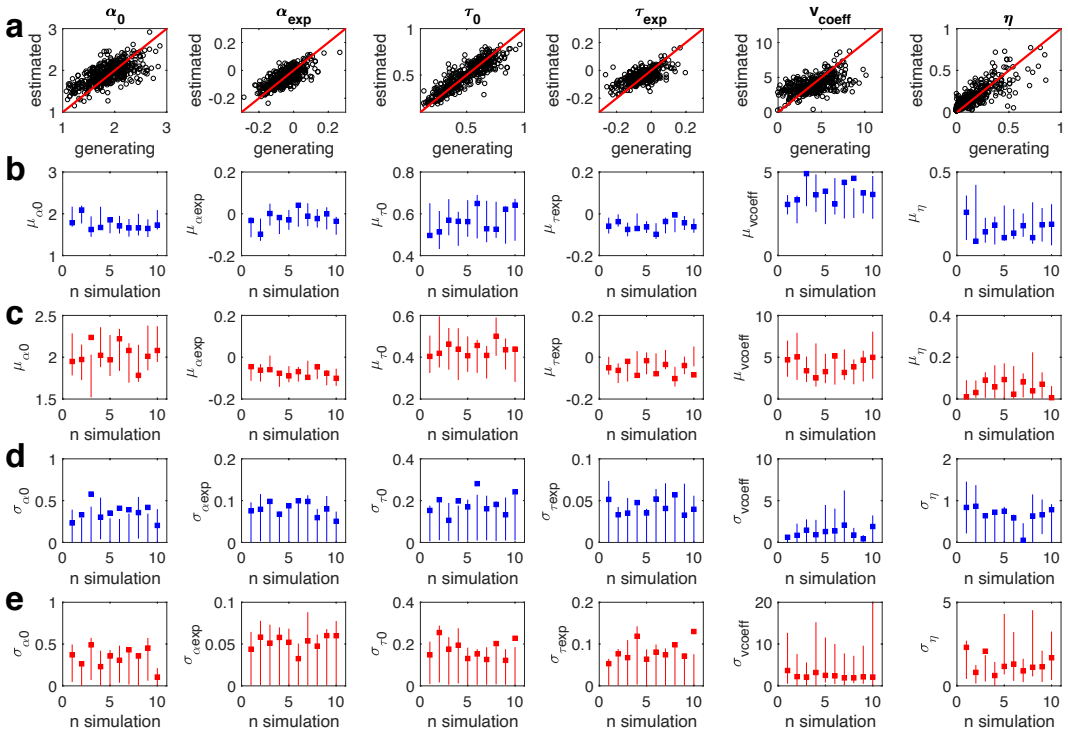


Figure 2. Parameter recovery simulation results for RLDDM4 (single learning rate, modulated boundary separation and modulated non-decision time). a) Generating vs. estimated single-subject parameters across all 10 simulations. Panels b-e show generating group-level parameters (means and standard deviations) plotted as squares, and estimated 95% highest posterior density intervals as vertical lines. b) Control group parameter means. c) Gambling group parameter means. d) Control group parameter standard deviations. e) Gambling group parameter standard deviations.

Parameter recovery simulations

Parameter recovery results from 10 simulations of the RLDDM4 are shown in Figure 2. The corresponding results from the RLDDM8 are provided in Supplemental Figure 1. As can be seen from the scatter plots in the top row of Figure 2, subject-level parameters recovered reasonably well, such that the correlations between generating and estimated parameters were

.56 < r <.90. Likewise, group-level parameters recovered well, such that the posterior distributions of group level means and standard deviations generally contained the generating parameters within their 95% highest density regions (lower panels in Figure 2). This was the case for the RLDDM8 as well (Supplemental Figure 1).

Posterior predictive checks

As a model comparison is always relative to a given set of candidate models, we next performed posterior predictive checks to examine the degree to which the best-fitting RLDDMs accounted for key patterns in the data, in particular with respect to changes in accuracy and RT over the course of learning. To this end, we simulated data from five models. For comparison, we included the DDM_0. For both single and dual learning rate models, we included both the models with constant α and τ (RLDDM1, RLDDM5) and the most complex models with modulated α and τ (RLDDM4, RLDDM8). We binned trials into ten bins over time, and computed mean accuracies and RTs per bin across a set of 10000 simulated full data sets (see methods section), separately for each group. Figure 3 (controls) and Figure 4 (gamblers) depicts the observed and model-predicted accuracies and RTs per trial bin for a subset of five models.

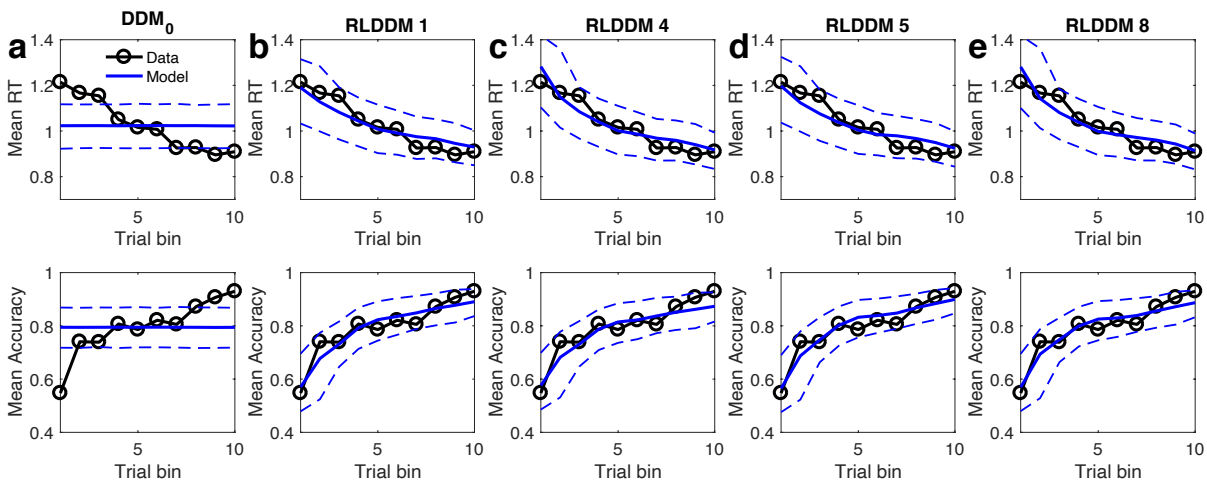


Figure 3. Posterior predictive checks in controls. Top row: observed RTs over time (black lines) and model predicted RTs (solid blue lines: means, dashed lines: +/- 95% percentiles). Bottom row shows observed accuracies over time (black lines) and model predicted accuracies (solid blue lines: means, dashed lines: +/- 95% percentiles). a) DDM_0 without reinforcement learning. b) RLDDM with a single learning rate, constant non-decision time and constant boundary separation. c) RLDDM with a single learning rate, modulated non-decision time and modulated boundary separation. d) RLDDM with dual learning rates, constant non-decision time and constant boundary separation. e) RLDDM with dual learning rates, modulated non-decision time and modulated boundary separation. Model comparison (see results) revealed that in controls, RLDDM8 fit the data best.

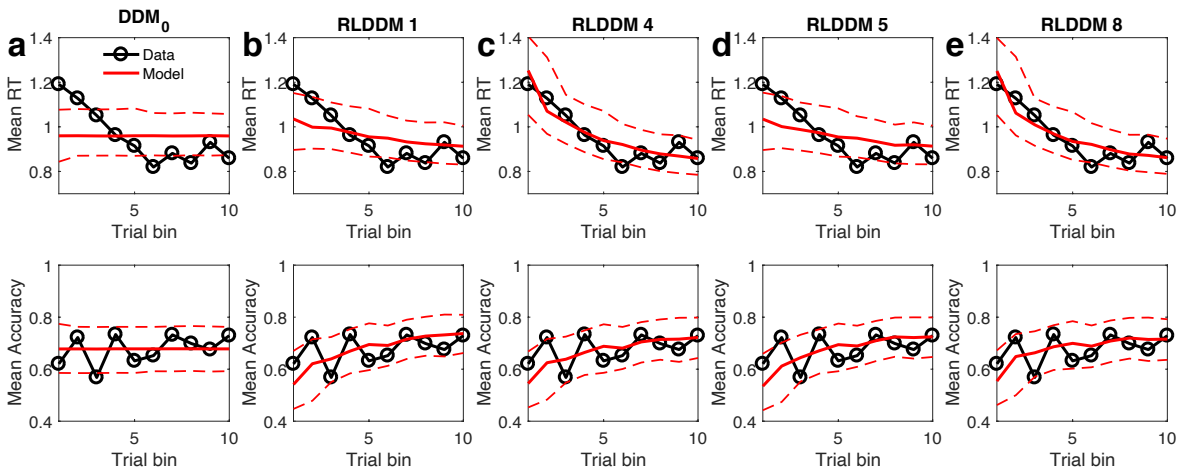


Figure 4. Posterior predictive checks in gamblers. Top row: observed RTs over time (black lines) and model predicted RTs (solid red lines: means, dashed lines: +/- 95% percentiles). Bottom row shows observed accuracies over time (black lines) and model predicted accuracies (solid red lines: means, dashed lines: +/- 95% percentiles). a) DDM_0 without reinforcement learning. b) RLDDM with a single learning rate, constant non-decision time and constant boundary separation. c) RLDDM with a single learning rate, modulated non-decision time and modulated boundary separation. d) RLDDM with dual learning rates, constant non-decision time and constant boundary separation. e) RLDDM with dual learning rates, modulated non-decision time and modulated boundary separation. Model comparison (see results) revealed that in gamblers RLDDM4 fit the data best.

The DDM_0 predicts constant accuracies and RTs over trials, and as can be seen in Figure 3a and 4a, it cannot reproduce the learning-related changes. In contrast, the RLDDMs predict learning-related increases in accuracy over time. These effects on accuracy are accounted for to a similar extent by all RLDDMs (lower panels in Figures 3b-e and 4b-e). Furthermore, the RLDDMs predict that RTs decrease over the course of learning. On the group level, all models in controls accounted for these RT effects in a largely similar way (but model comparison still favored the models with modulated α and τ). In gamblers, however, the differential account of the models with constant vs. modulated α and τ was visually quite striking (see Figure 4b-e upper panels): RLDDMs with constant α and τ (RLDDM1 and RLDDM5) did not fully reproduce the RT increases over time, whereas models with modulated α and τ (RLDDM4 and RLDDM8) reproduced these effects much more accurately. Together with the model comparison, this suggests that changes in RTs over time were only partially due to learning, and that modulations (decreases) in both α and τ over time additionally contributed to changes in RTs.

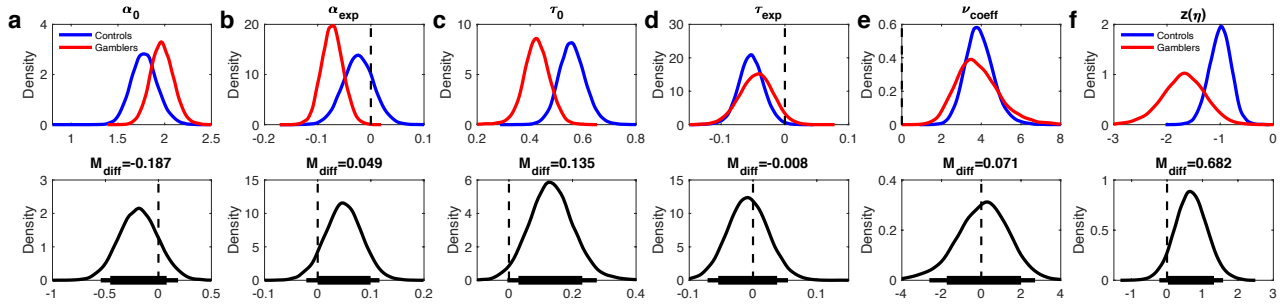


Figure 5. Results for the RLDDM4. Upper panels: posterior distributions of group mean parameters for controls (blue) and gamblers (red). Lower panels: posterior group differences per parameter (controls – gamblers). Solid (thin) horizontal lines in the lower panels denote 85% (95%) highest posterior density intervals.

Table 3. Group differences: Mean posterior group differences in model parameters (M_{diff}) and Bayes factors testing for directional effects (dBF). dBF values > 1 quantify the degree of evidence for a reduction in a parameter in gamblers vs. controls compared to the evidence for an increase. dBF values < 1 reflect the reverse. Directional tests: Bayes factors testing for directional effects on nested parameters.

RLDDM 4 parameter	**Group difference**		**Directional test (dBF)**	
	M_{diff}	***dBF***	***Controls***	***Gamblers***
α_0	-.187	.18	-	-
α_{exp}	.049	11.33	.25	.00016
τ_0	.135	33.24	-	
τ_{exp}	-.008	.69	.005	.044
v_{coeff}	.071	1.18	15819	12323
η	.68	13.47	-	

Group differences in model parameters

We next examined group differences in the parameters of the best-fitting RLDDMs in detail. Given that we observed a different model ranking in the two groups (recall that the model with dual learning rates fit the control data better, while the model with a single learning rate fit the gambler data better), we first focused on the simpler model that accounted for the data in the gamblers best (RLDDM4). Posterior distributions for all parameters per group as well as for group differences are depicted in Figure 4, and mean posterior group differences as well as Bayes Factors testing for directional group differences and effects (dBF) are listed in Table 3. First, gamblers showed a slightly higher initial boundary separation α_0 (Figure 5a, dBF=.18), that then attenuated substantially more rapidly in gamblers vs. controls, such that α_{exp} was substantially smaller in gamblers vs. controls (dBF=11.33). In line with this finding, a

directional test of α_{exp} revealed that in gamblers a value of < 0 was substantially more likely given the data, than a value of > 0 (dBF = .00016). In contrast, in controls this dBF was .25. Likewise, non-decision times were substantially reduced in gamblers (Figure 5b, dBF = 33.24), amounting to on average 135ms, but here no group differences in τ_{exp} were observed (Figure 5d). A directional test of τ_{exp} revealed that in gamblers a value of < 0 was substantially more likely, given the data, than a value of > 0 (dBF = .044). In controls, this dBF was .005. The value-dependency of the drift rate (v_{coeff}) was overall positive in both groups (dBFs >12000) and of similar magnitude (Figure 4e). In contrast, learning rates were higher in controls (Figure 5e, Table 3). The corresponding analyses using the RLDDM8 (the best fitting model in the controls) generally reproduced these effects (see Supplemental Figure 2 and Supplemental Table 1).

Model parameters showed the expected differential correlations with accuracy (see Supplemental Figure 3 for RLDDM4 and Supplemental Figure 4 for RLDDM8). We also combined the clinical questionnaires KFG[46] and SOGS[47], which showed a high positive correlation (r=.95), by averaging their respective Z-scores. However, none of the model parameters were significantly correlated with this measure of addiction severity in the gamblers (all |r| < .30).

Further visualizations of modeling results are provided in the Supplement. First, to illustrate the modulation of α and τ, Supplemental Figure 5 plots the trial-wise values of α and τ per each participant (according to Equations 6 and 7), computed using the mean posterior parameters per subject, as well as mean group effects. Supplemental Figure 6 then illustrates how the differences in learning rates resulted in a differential evolution of Q-values in the two groups.

Control analyses

We performed a number of additional control analyses. First, we carried out all group comparisons in posterior distributions also for the RLDDM8, which showed the best fit in the controls. As can be seen in Supplemental Figure 2 and Supplemental Table 1, results were highly similar to the analysis of the single learning rate model RLDDM4. Second, for comparison we also fit the data with a standard softmax choice rule (Equation 3). Here, model comparison revealed that a model with a single learning rate accounted for the data best in both groups. Results for this model (see Supplemental Figure 7 and Supplemental Table 2) revealed that the inverse temperature parameter was substantially reduced in gamblers vs.

controls, indicating higher levels of decision noise. Note that this is consistent with the effects observed for the RLDDMs, as both slower learning and a smaller boundary separation would contribute to noisier (more stochastic) decisions.

Gamblers show attenuated vmPFC value effects

We next examined neural valuation effects during the choice phase, i.e. during presentation of the decision options (see Figure 1). In controls, the Q-value difference between the chosen and unchosen option was correlated with activity in the ventromedial PFC (-10/38/-8, t=5.27, $p(FWE)_{SVC}$=.001, see Figure 6a). Chosen value representations in the same region were significantly attenuated in gamblers vs. controls (-8/34/-8, t=3.96, $p(FWE)_{SVC}$=.015). We confirmed the overlap of these effects both visually (see Figure 6a) and by computing a formal conjunction analysis across the two effects. The only region surviving this conjunction was again the vmPFC (-8/34/-8, t=3.96, $p(FWE)_{SVC}$=.015).

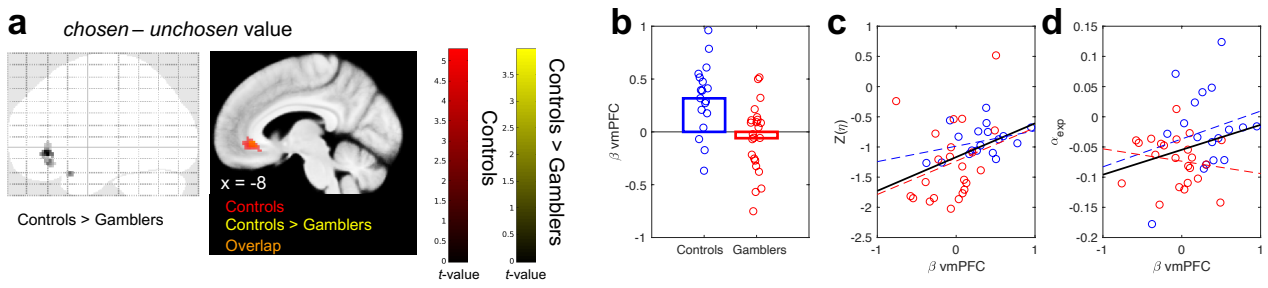


Figure 6. a) Positive linear parametric effects of chosen-unchosen value during option presentation were localized in vmPFC in controls (red, $p(PWE)_{SVC}$=.001). The response in this same region was attenuated in gamblers vs. controls (glass brain & yellow contrast on the section, overlap between contrasts in orange, $p(PWE)_{SVC}$=.015). b) Mean and single-subject parameter estimates per group for the effect depicted in a) for the peak voxel from the conjunction analysis. c) Scatter plot and regression lines of vmPFC parameter estimates against single-subject learning rates (red: gamblers, blue: controls, black: all subjects). d) Scatter plot and regression lines of vmPFC parameter estimates against single-subject boundary separation exponents (red: gamblers, blue: controls, black: all subjects). Maps are thresholded at p<.005, *uncorrected* with k>10 voxels for display purposes.

We next examined the degree to which individual model-parameters were associated with these vmPFC value representations. Indeed, reduced vmPFC chosen - unchosen value representations were associated with lower learning rates across all participants (Figure 6c, r_{all}=.39, p=.012). Correlations examined separately in the two groups were of similar direction and magnitude but not individually significant (see Figure 6c: $r_{controls}$=.33, p=.18, $r_{gamblers}$=.28,

p=.20). Likewise, reduced vmPFC chosen value representations tended to be associated with a more rapid reduction in boundary separation, that is, smaller values α_{exp} (Figure 6d, r_{all}=.26, p=.10). Correlations in the individual groups were of different direction and magnitude (see Figure 6d: $r_{controls}$=.22, p=.37, $r_{gamblers}$=-.17, p=.45).

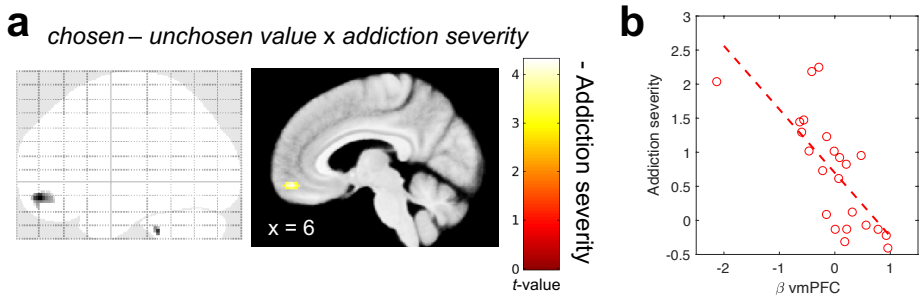


Figure 7. a) A negative correlation between a compound score of addiction severity (average z-score of KFG and SOGS gambling questionnaires) and chosen-unchosen value representations was observed in vmPFC. b) Scatter plot of the effect shown in a) at the peak voxel 6/52/-14. Maps are thresholded at p<.005, *uncorrected* with k>10 voxels for display purposes.

In a last step, we explored the link between alterations in value representations in the gamblers and clinical symptom severity. Since scores on the clinical questionnaires KFG[46] and SOGS[47] were highly correlated in the gamblers (r=.95), we combined them into a single measure of addiction severity by averaging their respective z-scores. An exploratory whole-brain regression of this addiction severity measure in the gamblers with the parametric chosen – unchosen value contrast revealed a negative association again in a region of ventromedial PFC (see Figure 7a, 6/52/-14, t=4.30). However, this cluster was located more anterior than the peak of the group difference (Figure 6) and fell outside our a-priori vmPFC region-of-interest.

Discussion

Here we comprehensively examined reinforcement learning impairments in gambling disorder using a combination of computational modeling and functional magnetic resonance imaging (fMRI). Modeling revealed that in both gamblers and controls, a reinforcement learning drift diffusion model (RLDDM) with modulated non-decision time and boundary separation accounted for the data best, showed good parameter recovery, and reproduced the observed accuracies and response times (RTs) over the course of learning well. Gamblers showed shorter non-decision times, reduced learning rates, and a more rapid attenuation of boundary separation over time. Effects were linked to attenuated chosen value representations in the vmPFC.

Model comparison showed that an extension of a RLDDM with modulated non-decision times and boundary separation parameters accounted for the data best in both groups. We performed extensive analyses to verify the performance of this extended RLDDM. First, we ran a series of parameter recovery simulations, which revealed that both subject-level and group-level parameters recovered well, which was the case for both the single (RLDDM4) and dual learning rates model (RLDDM8). Parameter recovery essentially determines the upper bound of reliability. It is therefore reassuring that estimated subject-level parameters showed a correlation between .59 and .90 with the true generating parameters. Likewise, estimated posterior distributions of group-level parameters generally contained the true generating parameters within their 95% highest posterior density intervals [36]. Second, in a series of posterior predictive checks, we simulated data from the model's posterior distributions. In both groups, the best-fitting RLDDMs reproduced both the increases in accuracy and the decreases in RTs over trials well. In particular in the gamblers, the difference in how well models with vs. without modulated boundary separation and non-decision times reproduced the observed RTs was visually apparent. Together, these analyses confirm that the modified RLDDMs performed well.

Interestingly, we observed a slightly different model ranking in the two groups. While generally, models with modulated boundary separation and non-decision times outperformed models were these parameters were held constant, groups differed with regard to learning rates: In controls the model with separate learning rates for positive and negative prediction errors performed best (RLDDM8), in line with previous work[32,36,48], whereas in gamblers, the model with a single learning rate showed the highest predictive accuracy (RLDDM4), with the dual learning rates model ranking 2nd. We addressed this group difference in model ranking in the following way. In order to err on the side of simplicity, we initially carried out

all main analyses using the single learning rate model (RLDDM4). However, we then also ran all modeling analyses (parameter recovery, posterior predictive checks, group comparisons) for the dual learning rate model (RLDDM8). These analyses confirmed that parameters of the RLDDM8 likewise recovered well and reproduced the observed group-level data to a similar degree as the RLDDM4. Comparisons of group-level posterior distributions of the RLDDM8 reproduced the effects observed for the single learning rate model (see Supplemental Figure 2 and Supplemental Table 1). Group differences in η_+ and η_- in the RLDDM8 were of similar magnitude as the learning rate differences in RLDDM4, but Bayes Factors were somewhat attenuated due to larger variances in group means in the gamblers.

Analysis of model parameters then allowed us to examine group differences in computational processes underlying task performance. While non-decision times showed a similar decay over time in both groups, the offset τ_0 was substantially shorter in gamblers. In contrast, while the boundary separation offset α_0 was slightly higher in gamblers, boundary separation then showed a much more rapid decay in gamblers vs. controls (α_{exp} was more negative). That is, over the course of the experiment, gamblers more than controls increasingly shifted their focus from accuracy to speed. These findings converge with previous observations of increased motor impulsivity[55], higher urgency/reduced premeditation[56] and higher levels of temporal discounting in gamblers[11]. In addition to these alterations in the speed-accuracy trade-off, learning impairments in gamblers were due to substantially reduced learning rates, consistent with previous work in other tasks [16,22,57]. Group differences in the drift rate scaling parameter v_{coeff}, on the other hand, were numerically small and of inconclusive directionality. Overall median RTs, although numerically shorter in gamblers, were not significantly different between groups (Figure 1d). Therefore, while shorter non-decision times and an overall more liberal threshold drove down RTs in the gamblers, this was counteracted by the reduction in learning rates. Recall that smaller learning rates, in combination with a drift rate coefficient v_{coeff} of similar magnitude, yields overall reduced drift rates over the course of the experiment. While the present work delineates computational mechanisms underlying reinforcement learning in gambling disorder, clinical factors contributing to these effects remain unclear. For example, overall clinical symptom severity [46,47] was not significantly associated with RLDDM parameters, but showed an association with vmPFC signals in an exploratory analysis (see below). Taken together, our findings highlight the power of computational analyses using combined RLDDMs[35]: model-based decomposition of RT distributions revealed substantial group differences in component

processes underlying reinforcement learning, despite the fact that overall RTs were similar between groups.

Our results might provide insights into potential mechanisms underlying the development and maintenance of gambling behavior. In animal models, exposure to uncertainty gives rise to behavioral and neural effects similar to those observed during repeated exposure to drugs of abuse[58–61], conceptually linking behavioral addictions[62] and theories of drug addiction such as incentive sensitization theory[63]. In structured environments, overall experienced uncertainty is inversely related to learning performance, and midbrain dopamine neurons fire maximally during uncertain reward prediction[64]. Likewise, human subcortical dopaminergic structures encode risk[65], and striatal dopamine release in gamblers is highest under conditions of maximum uncertainty[66]. Therefore, once could speculate that reduced learning might (via an increase in overall uncertainty and concomitant dopamine release[64]) increase the attribution of motivational value (incentive salience) to gambling cues [58,59]. This mechanism might contribute to the development and/or maintenance of gambling behavior[61,62].

In previous work[30], we examined exploration during reinforcement learning using a restless four-armed bandit task[29] in the same participants. This previous task differs from the present reinforcement learning task in a number of important respects: First, average payoffs of each bandit changed continuously according to gaussian random walk processes, whereas in the present task, reinforcement rates were stable. Second, reward feedback consisted of points in the range of 0-100, whereas in the present task, participants received probabilistic binary (win / no win) feedback. Third, 300 trials in total were performed in the bandit task, whereas the present task was substantially shorter. Finally, in our previous task, a minimum waiting period was included before button presses could be logged, which precluded us from comprehensively analyzing RTs and RLDDMs. Interestingly, however, in the four-armed bandit task, gamblers' general performance was comparable to that of controls[30], such that they earned a similar number of points across the experiment. Yet, computational modeling revealed that they relied less on a "directed exploration" strategy favoring selection of highly uncertain options[30] for information gain. It is nonetheless striking that in the arguable more complex task in a volatile environment, impairments in gamblers were more subtle, wheras in the present simple stationary reinforcement learning task, group differences in overall accuracy were substantial. What could account for these relative differences in performance? One possibility is that the required degree of temporal integration plays a role. In the restless bandit task, reward feedback on any given trial provides (almost) complete information on the

current value of the chosen bandit ("almost" because outcomes are corrupted by gaussian noise). In contrast, in order to accurately approximate the underlying reinforcement rates, the present task requires integration of outcomes across a number of consecutive trials. However, working memory deficits, which might contribute to impairments in feedback integration across trials, are not typical neuropsychological characteristics of gambling disorder [19,67].

FMRI analyses then confirmed a positive correlation between activity in vmPFC and chosen – unchosen value in controls, which is in line with a wealth of previous imaging findings[3,4,54,68]. Interestingly, a group comparison of these value signals then revealed a significant attenuation in gamblers vs. controls in the same vmPFC cluster. Furthermore, individual differences in learning rates and (to a lesser degree) boundary separation decay were correlated with signals in that same cluster. Valuation signals in a more anterior vmPFC cluster that fell outside our a priori vmPFC ROI were modulated by addiction severity in gamblers, as observed in a number of previous studies [6,69]. Alterations in regions of the reward system, in particular ventral striatum and vmPFC, have frequently been reported in gambling disorder, as outlined in a number of reviews [1,2,70]. However, the directionality of these changes in gambling disorder has long puzzled researchers, as both increases and reductions have been reported[2,5,8–10], e.g. depending on task phases [2,71], contextual factors [5,7] or reinforcer categories [6,72]. While the present results do not resolve these inconsistencies, they nonetheless contribute additional insights. We provide first evidence that during a simple stationary reinforcement learning task, chosen value representations in gamblers are attenuated, which adds to previous work that mostly focused on other tasks[2]. Nonetheless, similar effects have been observed in the Iowa Gambling Task [73] (but see[74]), whereas a previous fMRI study on reward learning did not report decision-related activity, but rather focused on outcome processing [22]. Another study found *enhanced* anticipatory activity in the ventral striatum during a guessing game[71] without a learning component. Furthermore, a different vmPFC subregion (outside our a priori vmPFC region-of-interest) showed parametric chosen – unchosen value activity that was negatively modulated by addiction severity in the gamblers. This complements previously reported modulations of reward responses in gambling disorder by clinical symptom severity observed using different tasks [2].

A number of limitations of the present study need to be acknowledged. First, the sample size was relatively small, and findings thus require replication in larger samples. Second, as is often the case in studies on gambling disorder, due to the greater prevalence of problem gambling behavior in males[75], only male participants were tested, limiting the generalizability of our results. Third, in contrast to the original study according to which the

task was set up[42], we only included a gain condition, and no loss condition. The degree to which the reported impairments in reinforcement learning in gamblers extend to tasks with an explicit loss condition therefore remain to be examined in future studies. Finally, classifications of gamblers into different subtypes according to clinical characteristics, disorder trajectories and gambling motivations have been proposed [76,77]. These factors are likely differentially associated with the impairments observed here. However, given the small sample size, examination of such subtypes was beyond the scope of the present study.

Taken together, here we provide a comprehensive drift diffusion model-based analysis of reinforcement learning performance in gambling disorder in a stationary environment, complementing our previous work on exploration in volatile environments[30]. Model-based decomposition of RTs revealed that, although overall RTs were similar between groups, the underlying processes differed considerably. In particular, gambling disorder was associated with reduced learning rates, faster non-decision times, and an increasing focus on speed vs. accuracy over the course of the experiment. Our findings highlight that reinforcement learning impairments in gambling disorder can be attributable to alterations in multiple component processes, and support the view that the vmPFC is a central hub in the neural circuitry underlying this disorder.

References

1. Fauth-Bühler, M., Mann, K. & Potenza, M. N. Pathological gambling: a review of the neurobiological evidence relevant for its classification as an addictive disorder. *Addict. Biol.* **22**, 885–897 (2017).

2. Clark, L., Boileau, I. & Zack, M. Neuroimaging of reward mechanisms in Gambling disorder: an integrative review. *Mol. Psychiatry* **24**, 674–693 (2019).

3. Clithero, J. A. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014).

4. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage* **76**, 412–427 (2013).

5. Leyton, M. & Vezina, P. Striatal ups and downs: their roles in vulnerability to addictions in humans. *Neurosci. Biobehav. Rev.* **37**, 1999–2014 (2013).

6. Miedl, S. F., Peters, J. & Büchel, C. Altered neural reward representations in pathological gamblers revealed by delay and probability discounting. *Arch. Gen. Psychiatry* **69**, 177–186 (2012).

7. Miedl, S. F., Buchel, C. & Peters, J. Cue-Induced Craving Increases Impulsivity via Changes in Striatal Value Signals in Problem Gamblers. *J. Neurosci.* **34**, 4750–4755 (2014).

8. Leyton, M. & Vezina, P. On cue: striatal ups and downs in addictions. *Biol Psychiatry* **72**, e21-2 (2012).

9. Balodis, I. M. *et al.* Attending to striatal ups and downs in addictions. *Biol Psychiatry* **72**, e25-6 (2012).

10. van Holst, R. J., Veltman, D. J., van den Brink, W. & Goudriaan, A. E. Right on cue? Striatal reactivity in problem gamblers. *Biol Psychiatry* **72**, e23-4 (2012).

11. Wiehler, A. & Peters, J. Reward-based decision making in pathological gambling: the roles of risk and delay. *Neurosci. Res.* **90**, 3–14 (2015).

12. Alessi, S. M. & Petry, N. M. Pathological gambling severity is associated with impulsivity in a delay discounting procedure. *Behav. Processes* **64**, 345–354 (2003).

13. Holt, D. D., Green, L. & Myerson, J. Is discounting impulsive?. Evidence from temporal and probability discounting in gambling and non-gambling college students. *Behav Process.* **64**, 355–367 (2003).

14. Dixon, M. R., Marley, J. & Jacobs, E. A. Delay discounting by pathological gamblers. *J Appl Behav Anal* **36**, 449–58 (2003).

15. Ligneul, R., Sescousse, G., Barbalat, G., Domenech, P. & Dreher, J. C. Shifted risk preferences in pathological gambling. *Psychol Med* 1–10 (2012).

16. Boog, M. *et al.* Cognitive Inflexibility in Gamblers is Primarily Present in Reward-Related Decision Making. *Front. Hum. Neurosci.* **8**, 569 (2014).

17. Goudriaan, A. E., Oosterlaan, J., de Beurs, E. & van den Brink, W. Decision making in pathological gambling: a comparison between pathological gamblers, alcohol dependents, persons with Tourette syndrome, and normal controls. *Brain Res Cogn Brain Res* **23**, 137–51 (2005).

18. Zhou, Z., Zhou, H. & Zhu, H. Working memory, executive function and impulsivity in Internet-addictive disorders: a comparison with pathological gambling. *Acta Neuropsychiatr.* **28**, 92–

100 (2016).

19. Ledgerwood, D. M. *et al.* Executive function in pathological gamblers and healthy controls. *J. Gambl. Stud.* **28**, 89–103 (2012).

20. Hur, J.-W. *et al.* Do pathological gambling and obsessive-compulsive disorder overlap? a neurocognitive perspective. *CNS Spectr.* **17**, 207–213 (2012).

21. Alvarez-Moya, E. M. *et al.* Subtyping study of a pathological gamblers sample. *Can. J. Psychiatry Rev. Can. Psychiatr.* **55**, 498–506 (2010).

22. de Ruiter, M. B. *et al.* Response perseveration and ventral prefrontal sensitivity to reward and punishment in male problem gamblers and smokers. *Neuropsychopharmacology* **34**, 1027–38 (2009).

23. van Timmeren, T., Daams, J. G., van Holst, R. J. & Goudriaan, A. E. Compulsivity-related neurocognitive performance deficits in gambling disorder: A systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* **84**, 204–217 (2018).

24. Harris, A. & Griffiths, M. D. A Critical Review of the Harm-Minimisation Tools Available for Electronic Gambling. *J. Gambl. Stud.* **33**, 187–221 (2017).

25. Yücel, M., Carter, A., Harrigan, K., van Holst, R. J. & Livingstone, C. Hooked on gambling: a problem of human or machine design? *Lancet Psychiatry* **5**, 20–21 (2018).

26. Breen, R. B. & Zimmerman, M. Rapid onset of pathological gambling in machine gamblers. *J. Gambl. Stud.* **18**, 31–43 (2002).

27. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).

28. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).

29. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).

30. Wiehler, A., Chakroun, K. & Peters, J. Attenuated directed exploration during reinforcement learning in gambling disorder. *bioRxiv* 823583 (2019) doi:10.1101/823583.

31. O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, Reward, and Decision Making. *Annu. Rev. Psychol.* **68**, 73–100 (2017).

32. Pedersen, M. L., Frank, M. J. & Biele, G. The drift diffusion model as the choice rule in reinforcement learning. *Psychon. Bull. Rev.* **24**, 1234–1251 (2017).

33. Forstmann, B. U., Ratcliff, R. & Wagenmakers, E.-J. Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annu. Rev. Psychol.* **67**, 641–666 (2016).

34. Ratcliff, R. & McKoon, G. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).

35. Miletić, S., Boag, R. J. & Forstmann, B. U. Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia* **136**, 107261 (2020).

36. Fontanesi, L., Gluth, S., Spektor, M. S. & Rieskamp, J. A reinforcement learning diffusion decision model for value-based decisions. *Psychon. Bull. Rev.* **26**, 1099–1121 (2019).

37. Shahar, N. *et al.* Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Comput. Biol.* **15**, e1006803 (2019).

38. Peters, J. & D'Esposito, M. The drift diffusion model as the choice rule in inter-temporal and risky choice: A case study in medial orbitofrontal cortex lesion patients and controls. *PLOS Comput. Biol.* **16**, e1007615 (2020).

39. Wagner, B., Clos, M., Sommer, T. & Peters, J. Dopaminergic modulation of human inter-temporal choice: a diffusion model analysis using the D2-receptor-antagonist haloperidol. *bioRxiv* 2020.02.13.942383 (2020) doi:10.1101/2020.02.13.942383.

40. Peters, J., Vega, T., Weinstein, D., Mitchell, J. & Kayser, A. Dopamine and risky decision-making in gambling disorder. *eNeuro* (2020) doi:10.1523/ENEURO.0461-19.2020.

41. Ballard, I. C. & McClure, S. M. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *J. Neurosci. Methods* **317**, 37–44 (2019).

42. Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**, 1042–1045 (2006).

43. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C. & Fagerstrom, K. O. The Fagerstrom Test for Nicotine Dependence: A revision of the Fagerstrom Tolerance Questionnaire. *Br. J. Addict.* **86**, 1119–27 (1991).

44. Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R. & Grant, M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption--II. *Addiction* **88**, 791–804 (1993).

45. Beck, A. T., Steer, R. A. & Brown, G.K. *Manual for the Beck Depression Inventory-II*. (Psychological Corporation, 1996).

46. Petry, J. *Psychotherapie der Glücksspielsucht*. (Beltz/Psychologie Verlags Union, 1996).

47. Lesieur, H. R. & Blume, S. B. The South Oaks Gambling Screen (SOGS): a new instrument for the identification of pathological gamblers. *Am J Psychiatry* **144**, 1184–8 (1987).

48. Frank, M. J., Seeberger, L. C. & O'Reilly R, C. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–3 (2004).

49. Maia, T. V. & Frank, M. J. From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.* **14**, 154–162 (2011).

50. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. in *Proceedings of the 3rd international workshop on distributed statistical computing* vol. 124 125 (Technische Universit at Wien, 2003).

51. Wabersich, D. & Vandekerckhove, J. Extending JAGS: a tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behav. Res. Methods* **46**, 15–28 (2014).

52. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).

53. Deichmann, R., Gottfried, J. A., Hutton, C. & Turner, R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *NeuroImage* **19**, 430–441 (2003).

54. Chib, V. S., Rangel, A., Shimojo, S. & O'Doherty, J. P. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *J Neurosci* **29**, 12315–20 (2009).

55. Chowdhury, N. S., Livesey, E. J., Blaszczynski, A. & Harris, J. A. Pathological Gambling and Motor Impulsivity: A Systematic Review with Meta-Analysis. *J. Gambl. Stud.* **33**, 1213–1239 (2017).

56. Kräplin, A. *et al.* Dysfunctional decision-making in pathological gambling: pattern specificity and the role of impulsivity. *Psychiatry Res.* **215**, 675–682 (2014).

57. Vanes, L. D. *et al.* Contingency learning in alcohol dependence and pathological gambling: learning and unlearning reward contingencies. *Alcohol. Clin. Exp. Res.* **38**, 1602–1610 (2014).

58. Anselme, P., Robinson, M. J. F. & Berridge, K. C. Reward uncertainty enhances incentive salience attribution as sign-tracking. *Behav. Brain Res.* **238**, 53–61 (2013).

59. Robinson, M. J. F., Anselme, P., Suchomel, K. & Berridge, K. C. Amphetamine-induced sensitization and reward uncertainty similarly enhance incentive salience for conditioned cues. *Behav. Neurosci.* **129**, 502–511 (2015).

60. Robinson, M. J. F., Anselme, P., Fischer, A. M. & Berridge, K. C. Initial uncertainty in Pavlovian reward prediction persistently elevates incentive salience and extends sign-tracking to normally unattractive cues. *Behav. Brain Res.* **266**, 119–130 (2014).

61. Zack, M., Featherstone, R. E., Mathewson, S. & Fletcher, P. J. Chronic exposure to a gambling-like schedule of reward predictive stimuli can promote sensitization to amphetamine in rats. *Front. Behav. Neurosci.* **8**, 36 (2014).

62. Robinson, M. J. F., Fischer, A. M., Ahuja, A., Lesser, E. N. & Maniates, H. Roles of 'Wanting' and 'Liking' in Motivating Behavior: Gambling, Food, and Drug Addictions. *Curr. Top. Behav. Neurosci.* **27**, 105–136 (2016).

63. Robinson, T. E. & Berridge, K. C. The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Res. Brain Res. Rev.* **18**, 247–291 (1993).

64. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–902 (2003).

65. Preuschoff, K., Bossaerts, P. & Quartz, S. R. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* **51**, 381–390 (2006).

66. Linnet, J. *et al.* Striatal dopamine release codes uncertainty in pathological gambling. *Psychiatry Res.* **204**, 55–60 (2012).

67. Kapsomenakis, A., Simos, P. G., Konstantakopoulos, G. & Kasselimis, D. S. In Search of Executive Impairment in Pathological Gambling: A Neuropsychological Study on Non-treatment Seeking Gamblers. *J. Gambl. Stud.* **34**, 1327–1340 (2018).

68. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal Cortex Encodes Willingness to Pay in Everyday Economic Transactions. *J. Neurosci.* **27**, 9984–9988 (2007).

69. Reuter, J. *et al.* Pathological gambling is linked to reduced activation of the mesolimbic reward system. *Nat Neurosci* **8**, 147–148 (2005).

70. Potenza, M. N. Neurobiology of gambling behaviors. *Curr. Opin. Neurobiol.* **23**, 660–667 (2013).

71. van Holst, R. J., Veltman, D. J., Büchel, C., van den Brink, W. & Goudriaan, A. E. Distorted Expectancy Coding in Problem Gambling: Is the Addictive in the Anticipation? *Biol. Psychiatry* **71**, 741–748 (2012).

72. Sescousse, G., Barbalat, G., Domenech, P. & Dreher, J.-C. Imbalance in the sensitivity to different types of rewards in pathological gambling. *Brain* **136**, 2527–2538 (2013).

73. Tanabe, J. *et al.* Prefrontal cortex activity is reduced in gambling and nongambling substance users during decision-making. *Hum. Brain Mapp.* **28**, 1276–1286 (2007).

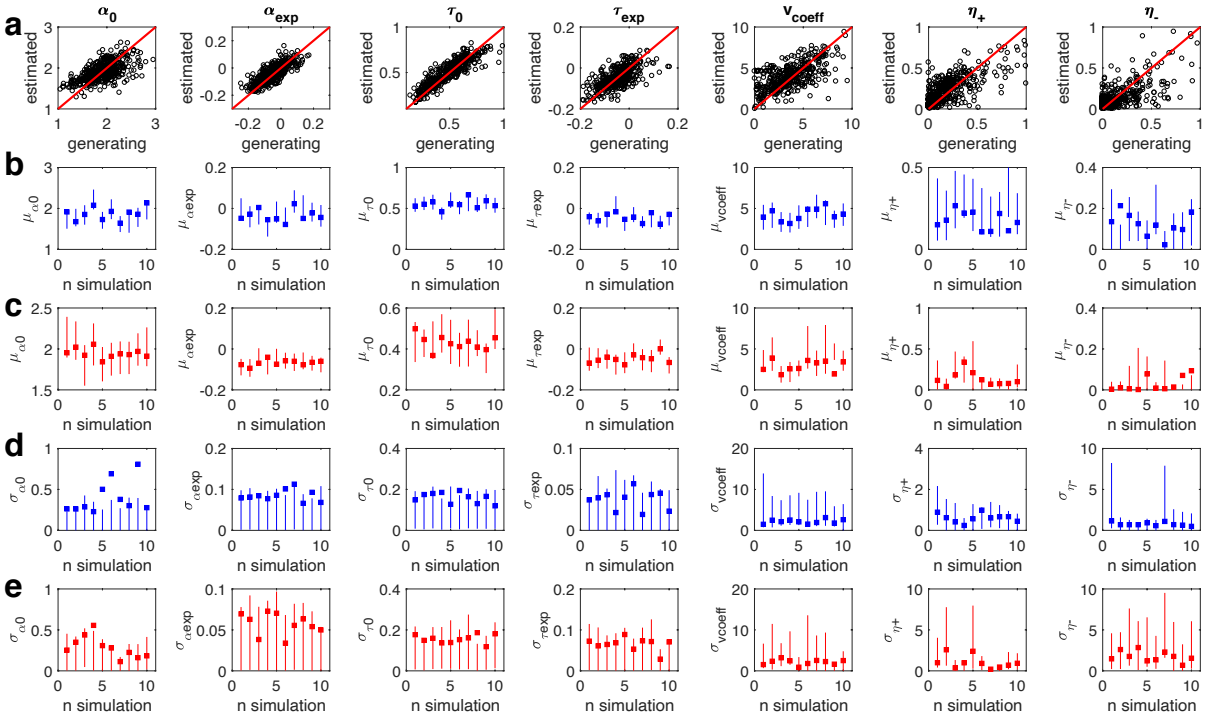
74. Power, Y., Goodyear, B. & Crockford, D. Neural correlates of pathological gamblers preference for immediate rewards during the iowa gambling task: an fMRI study. *J. Gambl. Stud.* **28**, 623–636 (2012).

75. Hing, N., Russell, A., Tolchard, B. & Nower, L. Risk Factors for Gambling Problems: An Analysis by Gender. *J. Gambl. Stud.* **32**, 511–534 (2016).

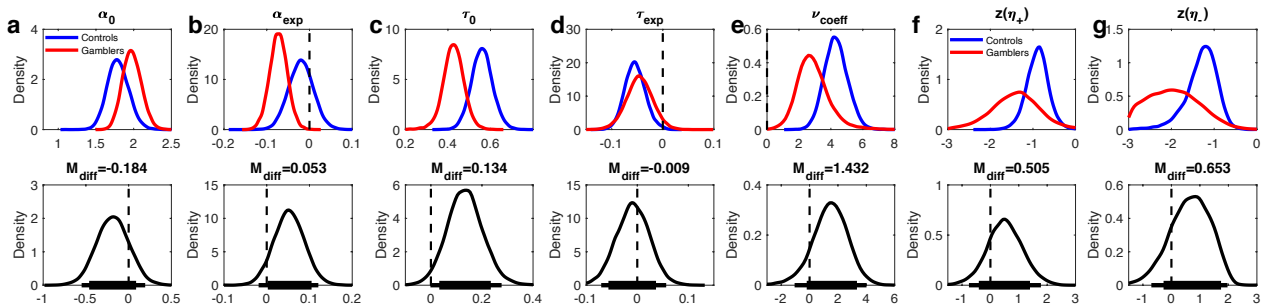
76. Milosevic, A. & Ledgerwood, D. M. The subtyping of pathological gambling: a comprehensive review. *Clin. Psychol. Rev.* **30**, 988–998 (2010).

77. Blaszczynski, A. & Nower, L. A pathways model of problem and pathological gambling. *Addict. Abingdon Engl.* **97**, 487–499 (2002).

Supplemental material



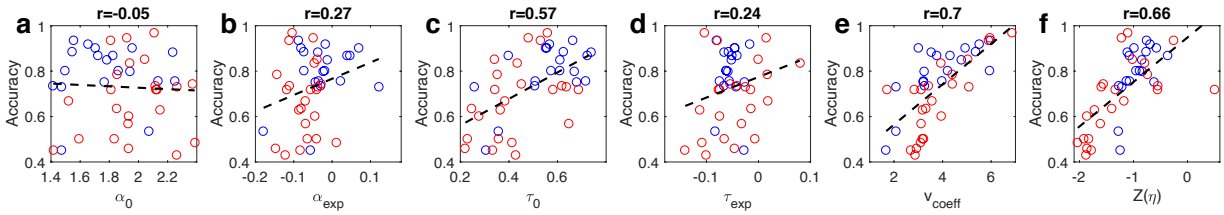
Supplemental Figure 1. Parameter recovery simulation results for RLDDM8 (dual learning rates, modulated boundary separation and non-decision time). a) Generating vs. estimated single-subject parameters across all 10 simulations. b) Control group parameter means. c) Gambling group parameter means. d) Control group parameter standard deviations. e) Gambling group parameter standard deviations. In b-e, squares denote the generating parameter, and vertical lines denote the 95% highest posterior density of the parameter estimation.



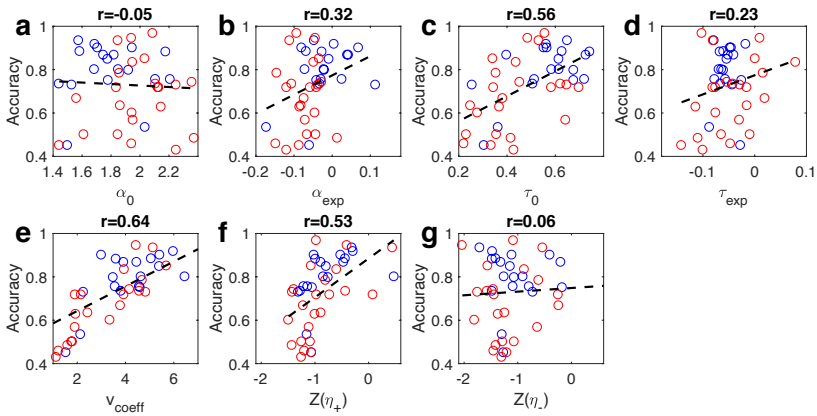
Supplemental Figure 2. RLDDM8 results. Upper panels: posterior distributions of group mean parameters for controls (blue) and gamblers (red). Lower panels: posterior group differences per parameter (controls – gamblers). Solid (thin) horizontal lines in the lower panels denote 85% (95%) highest posterior densities.

Supplemental Table 1. Group differences: Mean posterior group differences in model parameters (M_{diff}) and Bayes factors testing for directional effects (dBF) for the RLDDM8. dBF values > 1 quantify the degree of evidence for a reduction in a parameter in gamblers vs. controls compared to the evidence for an increase. dBF values < 1 reflect the reverse. Directional tests: Bayes factors testing for directional effects on nested parameters (i.e. parameters were a value of 0 indicates the absence of the corresponding effect).

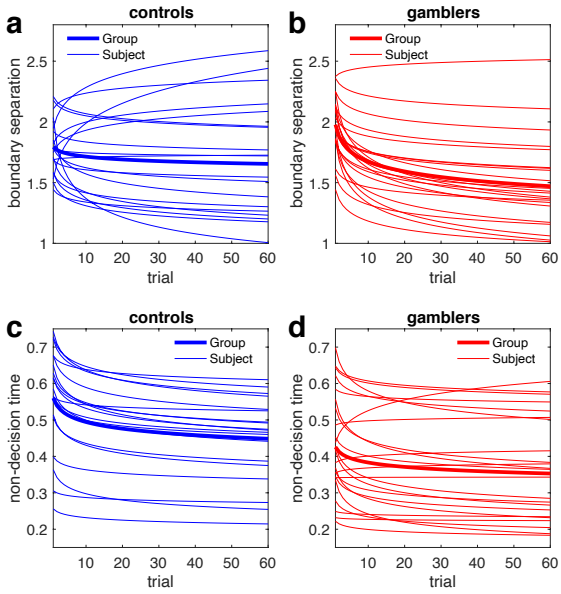
RLDDM 8 parameter	**Group difference**		**Directional test (dBF)**	
	M_{diff}	***dBF***	***Controls***	***Gamblers***
α_0	-.184	.20	-	
α_{exp}	.053	14.17	.31	.00063
τ_0	.134	34.13	-	
τ_{exp}	-.009	.67	.0059	.039
v_{coeff}	1.43	7.47	15914	711.1
η_+	.51	3.69	-	
η_-	.65	4.37	-	



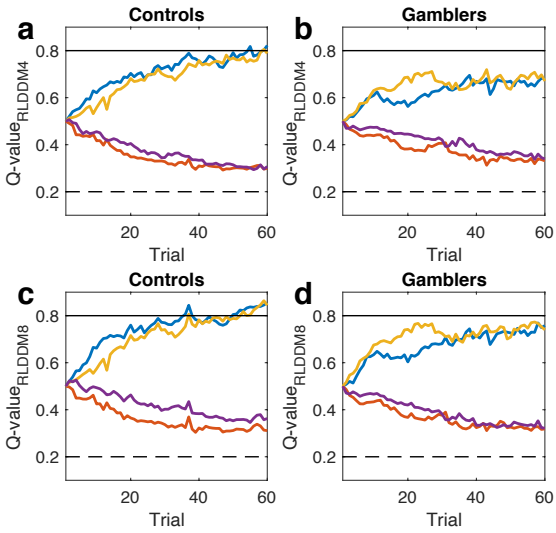
Supplemental Figure 3. Between-subject correlations between single-subject mean posterior model parameters from the RLDDM4 and overall accuracy (red data points: gamblers, blue data points: controls).



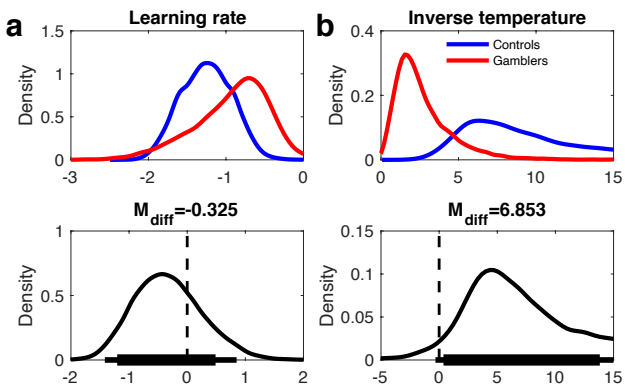
Supplemental Figure 4. Between-subject correlations between single-subject mean posterior model parameters from the RLDDM8 and overall accuracy (red data points: gamblers, blue data points: controls).



Supplemental Figure 5. Estimated trial-wise changes in boundary separation (a,b) and non-decision time (c,d) for controls (blue) and gamblers (red) according to RLDDM4. The solid lines plot parameter changes based on the mean group-level posteriors, whereas the thin lines depict individual subject curves.



Supplemental Figure 6. Evolution of mean Q-values for the RLDDM4 (a: controls, b: gamblers) and the RLDDM8 (c: controls, d: gamblers) for the two optimal (yellow, blue) and suboptimal (red, purple) options. The true reinforcement rates (.8, .2) are indicated by solid and dashed black lines.



Supplemental Figure 7. Upper panels: Softmax model posterior distributions of group mean learning rates (a) and softmax inverse temperatures (b) for controls (blue) and gamblers (red). Lower panels: posterior group differences per parameter (controls – gamblers). Solid (thin) horizontal lines in the lower panels denote 85% (95%) highest posterior density intervals.

Supplemental Table 2. Group differences: Mean posterior group differences in model parameters (M_{diff}) and Bayes factors testing for directional effects (dBF). dBF values > 1 quantify the degree of evidence for a reduction in a parameter in gamblers vs. controls compared to the evidence for an increase. dBF values < 1 reflect the reverse.

SM parameter	**Group difference**	
	M_{diff}	***dBF***
η	-.325	.373
β	6.85	22.89