nature COMMUNICATIONS

ARTICLE

Received 10 Dec 2014 | Accepted 14 Jul 2015 | Published 25 Aug 2015

DOI: 10.1038/ncomms9096

OPEN

Contextual modulation of value signals in reward and punishment learning

Stefano Palminteri[1,2], Mehdi Khamassi[3,4], Mateus Joffily[4,5] & Giorgio Coricelli[2,4,6]

Compared with reward seeking, punishment avoidance learning is less clearly understood at both the computational and neurobiological levels. Here we demonstrate, using computational modelling and fMRI in humans, that learning option values in a relative—context-dependent—scale offers a simple computational solution for avoidance learning. The context (or state) value sets the reference point to which an outcome should be compared before updating the option value. Consequently, in contexts with an overall negative expected value, successful punishment avoidance acquires a positive value, thus reinforcing the response. As revealed by post-learning assessment of options values, contextual influences are enhanced when subjects are informed about the result of the forgone alternative (counterfactual information). This is mirrored at the neural level by a shift in negative outcome encoding from the anterior insula to the ventral striatum, suggesting that value contextualization also limits the need to mobilize an opponent punishment learning system.

[1] Institute of Cognitive Neuroscience (ICN), University College London (UCL), London WC1N 3AR, UK. [2] Laboratoire de Neurosciences Cognitives (LNC), Département d'Etudes Cognitives (DEC), Institut National de la Santé et Recherche Médical (INSERM) U960, École Normale Supérieure (ENS), 75005 Paris, France. [3] Instintut des Systèmes Intelligents et Robotique (ISIR), Centre National de la Recherche Scientifique (CNRS) UMR 7222, Université Pierre et Marie Curie (UPMC), 70013 Paris, France. [4] Interdepartmental Centre for Mind/Brain Sciences (CIMeC), Università degli study di Trento, 38060 Trento, Italy. [5] Groupe d'Analyse et de Théorie Economique, Centre National de la Recherche Scientifique (CNRS) UMR 5229, Université de Lyon, 69003 Lyon, France. [6] Department of Economics, University of Southern California (USC), 90089-0253 Los Angeles, California, USA. Correspondence and requests for materials should be addressed to S.P. (email: stefano.palminteri@gmail.com).

In the past decades, significant advances have been made in the understanding of the computational and neural bases of reward-based learning and decision making. On the other hand, computations and neural mechanisms mediating punishment-based learning and decision making remain more elusive[1,2].

The first problem is computational. In fact, avoidance learning faces an apparent paradox: once a punishment is successfully avoided, the instrumental response is no longer reinforced. As a consequence, basic learning models predict better performance on reward learning (in which the extrinsic reinforcements are frequent, because they are sought) compared with punishment learning (in which the extrinsic reinforcements are infrequent, because they are avoided), despite the fact that human subjects have been shown to learn equally well in both domains[3–6].

The second problem is neuroanatomical: a debate in cognitive neuroscience concerns whether the same brain areas (namely the ventral striatum and the ventromedial prefrontal cortex) represent positive as well as negative values or, alternatively, aversive value encoding and learning are organized in an opponent system (namely the insula and the dorsomedial prefrontal cortex)[7–12].

We hypothesized that the two questions could be resolved in the framework of value context dependence. Recently, context dependence of option values has provided a formal framework to understand adaptive coding and range adaptation of value-responsive neurons and brain areas[13–16]. Concerning punishment learning, operationalizing the principle behind the two-factor theory, we propose that successful avoidance, which is a neutral outcome in an absolute scale, acquires a positive value because it is computed relative to the value of its choice context, which is negative[17–19]. In other words, successful avoidance is 'reframed' as a positive outcome[20]. On the other side, divergent functional magnetic resonance imaging (fMRI) findings could be reconciled assuming that, in absence or limited contextual information, punishments and rewards are implemented in opponent channels; subsequently, if contextual information is acquired or provided, outcome representations converge to the ventral frontostriatal system. This is supported by the fact that ventral striatal and prefrontal responses to punishment avoidance were observed in situations in which the value of the context was made explicit by instruction or overtraining[21–23].

To test these hypotheses, healthy subjects underwent fMRI scanning while performing an instrumental learning task, involving multiple two-armed bandits (choice contexts) and followed by a post-learning assessment of option values. Two features of the task served our purposes: first, the task contrasted reward seeking with punishment avoidance learning; second, in specific choice contexts, we provided the information about the outcome of the foregone alternative—counterfactual information—to enhance relative value encoding[24–26]. We reasoned that presenting subjects with both the outcomes of the chosen and the unchosen options would facilitate the learning of the average value of the choice context (that is, the context value).

We found behavioural and neural evidence consistent with the idea that providing both the outcomes of the chosen and the unchosen options favoured the learning of a context-specific reference point. Behavioural results indicated that subjects learn similarly well reward seeking and punishment avoiding: a result that was efficiently captured by a computational model that embodies the idea of relative value learning. The same model was able to account for context dependence-induced valuation biases, as revealed by the post-learning test, specifically for options learnt in the presence of counterfactual feedback. fMRI analyses served two purposes. First, we used neural data to provide further experimental support to the computational analyses. Crucially model-based and model-free fMRI analyses concordantly indicated that neural activity in the brain valuation system was better explained assuming relative, than absolute value learning. Second, fMRI permitted us to reconcile previous discordant findings advocating for anatomical overlap or dissociation between reward seeking and punishment avoidance neural systems. In fact, the observed increase in contextual discrimination in the complete feedback conditions was followed by a shift in the neural encoding of negative outcomes from the insula to the ventral striatum.

Results

Experimental design. Healthy subjects performed a probabilistic instrumental learning task with monetary gains and losses, adapted from those used in previous imaging, pharmacological and lesion studies[3,6,27,28]. The novel task presented a 2 × 2 factorial design with outcome valence (reward or punishment) and feedback information (partial or complete) as factors (Fig. 1a,b). In the learning task, options (materialized as abstract symbols) were always presented in fixed pairs. The fixed pairs of options represented stable choice contexts, with different overall expected value. In each context the two options were associated with different, but stationary, outcome probabilities, so that the subjects' task was learning to choose the options associated either with highest reward probability or those associated with lowest punishment probability (correct options: G_{75} and L_{25} in the reward and the punishment context, respectively; incorrect options: G_{25} and L_{75} in the reward and the punishment context, respectively). Subjects performed four sessions of the task during fMRI scanning, each involving novel pairs of options. After the last session, subjects performed a post-learning test in which they were asked to indicate the option with the highest value, in choices involving all possible binary combinations—that is, including pairs of options that had never been associated during the task (Fig. 1c). As in previous studies, post-learning test choices were not followed by feedback, to not interfere with subjects' final estimates of option values[29,30].

Instrumental performance. We found significant evidence of instrumental learning (that is, participants sought rewards or avoided punishments; Table 1). Indeed, average correct response rate was significantly higher than chance level (that is, 0.5) in all contexts ($T > 7.0$, $P < 0.001$; Fig. 2a). A two-way analysis of variance (ANOVA) showed no effect of outcome valence ($F = 1.4$, $P > 0.2$), a significant effect of feedback information ($F = 30.7$, $P < 0.001$), and no significant interaction ($F = 0.7$, $P > 0.7$). Accordingly, *post hoc* investigation showed performances in complete feedback contexts as significantly higher compared with the partial feedback contexts (reward and punishment contexts: $T > 3$, $P < 0.01$). Thus, as in previous studies, healthy subjects learnt similarly from reward and punishments[3,29], and efficiently integrated counterfactual information in instrumental learning[31–33] (see Supplementary Note 1 and Supplementary Fig. 1 for reaction times data analysis).

Post-learning choices. We found significant evidence of value retrieval during the post-learning test (Table 1)[29,30]. Indeed, a three-way ANOVA showed a significant effect of outcome valence ($F = 53.0$, $P < 0.001$) and a significant effect of option correctness ($F = 170.1$, $P < 0.001$), but no effect of feedback information ($F = 0.0$, $P > 0.5$; Fig. 2b). The only interaction that reached statistical significance was the correctness *x* feedback information ($F = 11.9$, $P < 0.01$). As for other interactions (double or triple), none reached statistical significance (all: $F < 2.0$, $P > 0.1$). A two-way ANOVA limited on the intermediate value

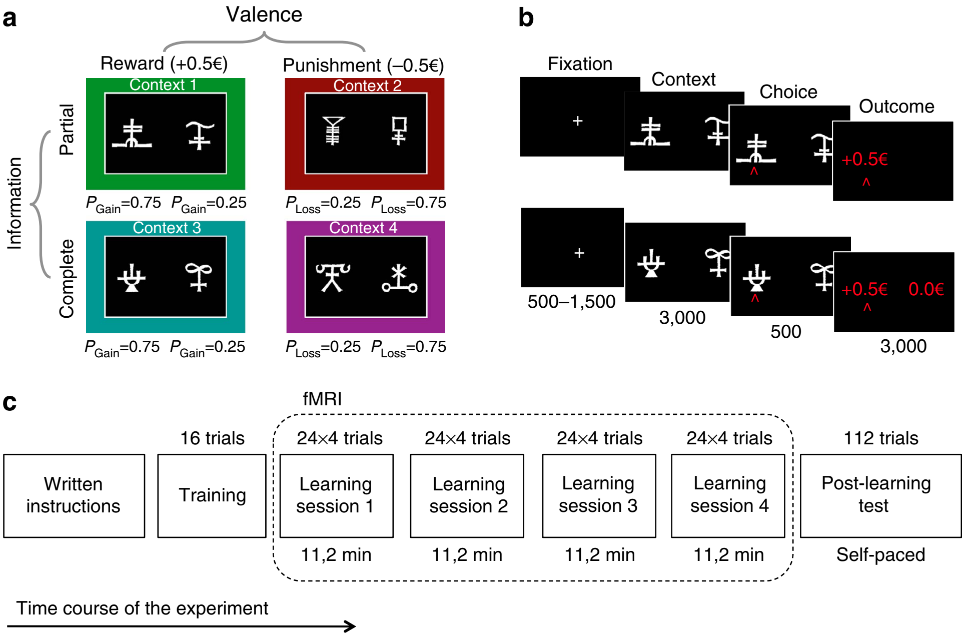


Figure 1 | Experimental task and design. (**a**) Learning task 2 × 2 factorial design with 4 different contexts: reward/partial, punishment/partial, reward/complete, and punishment/complete. $P_{Gain} =$ probability of winning 0.5€; $P_{Loss} =$ probability of losing 0.5€. Note that the coloured frames are introduced in the figure for illustrative purposes, but were not present in the original task. (**b**) Successive screens of typical trials in the reward partial (top) and complete (bottom) contexts. Durations are given in milliseconds. (**c**) Time course of the experiment. Note that the post-learning test was uniquely based on the eight options of the last learning session.

Table 1 | Experimental and computational model-derived variables.

Dependent variables	DATA	ABSOLUTE	RELATIVE
Learning test: correct choice rate			
Reward partial (% correct)	0.73 ± 0.03	0.75 ± 0.02	0.74 ± 0.02
Punishment partial (% correct)	0.74 ± 0.03	0.70 ± 0.02	0.72 ± 0.02
Reward complete (% correct)	0.83 ± 0.02	0.83 ± 0.02	0.84 ± 0.02
Punishment complete (% correct)	0.86 ± 0.02	0.81 ± 0.02*	0.84 ± 0.02
Post-learning test: choice rate			
G_{75} partial (% choices)	0.78 ± 0.04	0.86 ± 0.01	0.85 ± 0.01
G_{25} partial (% choices)	0.51 ± 0.06	0.58 ± 0.01	0.54 ± 0.01
L_{25} partial (% choices)	0.45 ± 0.04	0.37 ± 0.01	0.45 ± 0.01
L_{75} partial (% choices)	0.25 ± 0.03	0.17 ± 0.01	0.16 ± 0.01
G_{75} complete (% choices)	0.83 ± 0.03	0.91 ± 0.01	0.83 ± 0.02
G_{25} complete (% choices)	0.40 ± 0.04	0.66 ± 0.01*	0.38 ± 0.03
L_{25} complete (% choices)	0.61 ± 0.03	0.37 ± 0.01*	0.62 ± 0.03
L_{75} complete (% choices)	0.17 ± 0.03	0.08 ± 0.01	0.16 ± 0.02

ABSOLUTE, absolute value learning model; DATA, experimental data; RELATIVE, relative value learning model (best-fitting model).
The table summarizes for both tasks their experimental and model-derived dependent variables. Data are expressed as mean ± s.e.m.
*$P < 0.05$, *t*-test, comparing the model-derived values with the actual data after correcting for multiple comparisons ($N = 28$).

options (that is, the less rewarding option in the reward contexts and the less-punishing option in the punishment contexts G_{25} and L_{25}) with valence and feedback information as factors, crucially showed no significant effect of valence ($F = 1.6$, $P > 0.2$) nor of feedback information ($F = 0.2$, $P > 0.2$), but a significant interaction ($F = 9.4$, $P < 0.01$), thus reflecting an inversion in the evaluation of intermediate options, when moving from the partial to the complete feedback information contexts. More precisely, *post hoc* tests revealed that the percentage of choices towards the correct option of the punishment/complete context (L_{25}) was higher compared with that towards the incorrect option in the reward/complete context (G_{25}; $T = 3.2$, $P < 0.01$), despite their absolute expected value (EV; Probability(outcome) × Magnitude (outcome)) suggesting the opposite. ($EV(L_{25}) = -12.5¢$; $EV(G_{25}) = +12.5¢$). *Post hoc* analysis also showed a significantly different choice rates for the correct options in the reward compared with the punishment context (G_{75} versus L_{25} in both feedback information contexts: $T > 4.6$, $P < 0.001$), despite similar choice rate in the learning task (see also Supplementary Table 1). This indicated that post-learning choices could be explained neither by assuming that option values were encoded in an absolute manner, nor by assuming that they were merely reflecting past choice propensity (policy), but that they laid somehow halfway between these two extremes: a phenomenon that is parsimoniously explained by context-dependent option-value learning.

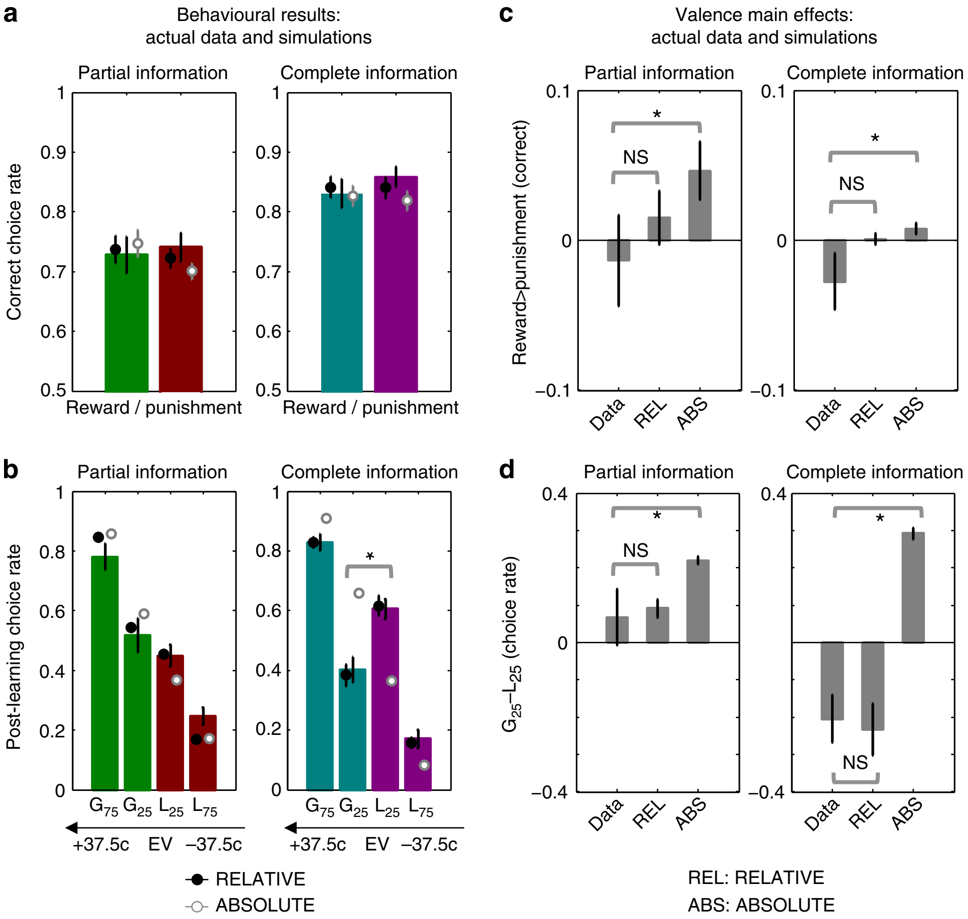


Figure 2 | Behavioural results and model simulations. (**a**) Correct choice rate during the learning test. (**b**) Choice rate in the post-learning test. G_{75} and G_{25}: options associated with 75% and 25% per cent of winning 0.5€, respectively; L_{75} and L_{25}: options associated with 75% and 25% per cent of losing 0.5€, respectively. EV: absolute expected value (Probability(outcome) × Magnitude(outcome)) in a single trial. The values +37.5¢ and −37.5¢ correspond G_{75} and the L_{75} options, respectively. In **a** and **b** coloured bars represent the actual data and black (RELATIVE) and white (ABSOLUTE) dots represent the model simulated data. (**c**) Reward minus punishment correct choice rate during the learning test. (**d**) G_{25} minus L_{25} choice rate during the post learning test. $^*P<0.05$ one sample t-test; NS, not significant ($N=28$). Error bars represent s.e.m.

Computational models. We fitted the behavioural data with model-free reinforcement-learning models (see Methods)[34]. The tested models included a standard Q-learning (thereafter referred to as ABSOLUTE), adapted to account for learning from counterfactual feedback, which has been most frequently used with this kind of task and we therefore consider as the reference model (hypothesis zero)[3,6,27,28,33]. We also considered a modified version of the ABSOLUTE model, which, similarly to other theories assumes that choice context (or state) values are separately learnt and represented[35,36]. The crucial feature of this model (thereafter referred to as RELATIVE) is that the context value sets the reference point to which an outcome should be compared before updating the option value; option values are therefore no longer encoded in an absolute, but in a relative scale (Fig. 3). The context value (V(s)) is defined as a 'random-policy' state value, aimed at capturing the overall expected value of a given pair of options, independent from subjects' choice propensity. Note that the RELATIVE model shares a crucial feature (that is, relative option value encoding) with previous computational formulations, such as actor–critic and advantage learning models, that inspired its conception (see Supplementary Note 2 for additional model comparison including these preceding models and a discussion of their differences)[37,38].

Bayesian model selection. For each model, we estimated the free parameters by likelihood maximization (to calculate the Akaike Information Criterion, AIC, and the Bayesian Information Criterion, BIC) and by Laplace approximation of the model evidence (to calculate the exceedance probability; Tables 2 and 3). After *post hoc* analyses we found that the RELATIVE model better accounted for the data, both at fixed and random effect analysis (compared with the ABSOLUTE LL: $T=4.1$, $P<0.001$). This was also true when accounting (penalizing) for the different number of free parameters (AIC: $T=3.4$, $P<0.001$; BIC: $T=2.1$, $P<0.05$)[39]. We also calculated the exceedance probability (XP) of the model based on an approximate posterior probability of the model, and we consistently found that our model significantly outperformed the others (XP = 1.0)[40]. Thus, context-dependent value encoding (RELATIVE) provided better account of learning test choices, even after correcting for its higher degrees of freedom (note that this conclusion was not affected by using different learning rates for the reward and the punishment contexts).

Relative value encoding explains instrumental performance. To characterize the effect of context-dependent over absolute value learning, we generated for each trial t the probability of choosing the best option according to the models, given the

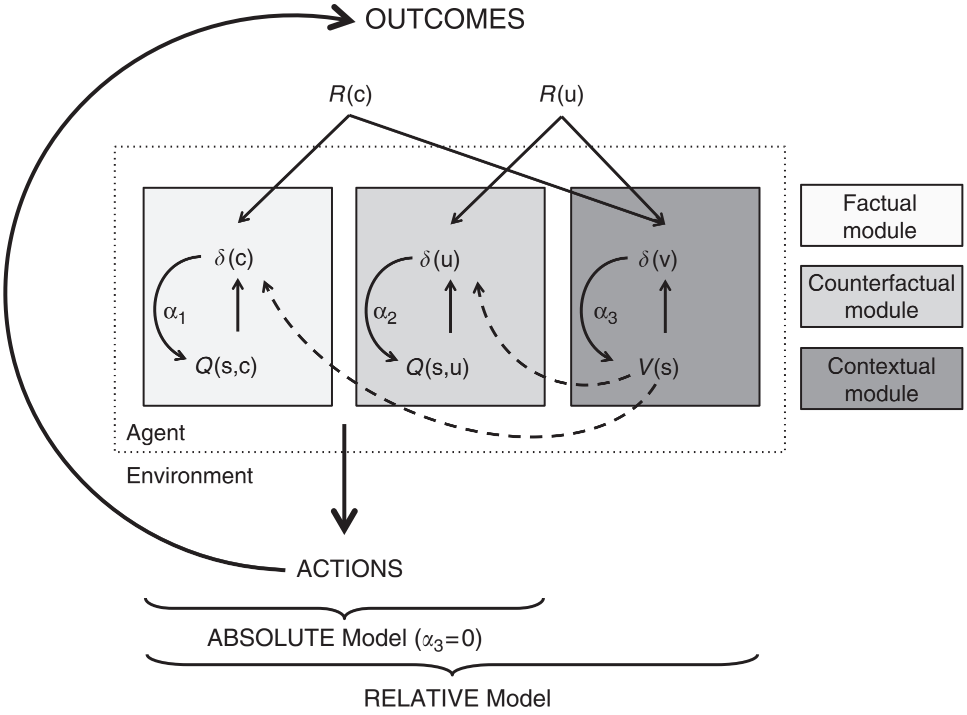


Figure 3 | Computational architecture. The schematic illustrates the computational architecture used for data analysis. For each context (or state) 's', the agent tracks option values (Q(s,:)), which are used to decide amongst alternative courses of action. In all contexts, the agent is informed about the outcome corresponding to the chosen option (R(c)), which is used to update the chosen option value (Q(s,c)) via a prediction error (δ(c)). This computational module ('factual learning') requires a learning rate (α_1). In the complete feedback condition, the agent is also informed about the outcome of the unselected option (*R*(u)), which is used to update the unselected option value (*Q*(s,u)) via a prediction error (δ(u)). This computational module ('counterfactual learning') requires a specific learning rate (α_2). In addition to tracking option value, the agent also tracks the value of the context (*V*(s)), which is also updated via a prediction error (δ(v)), integrating over all available feedback information (*R*(c) and *R*(u), in the complete feedback contexts and *Q*(s,u) in the partial feedback contexts). This computational module ('contextual learning') requires a specific learning rate (α_3). The RELATIVE model can be reduced to the ABSOLUTE model by suppressing the contextual learning module (that is, assuming $\alpha_3 = 0$).

Table 2 | Model comparison criteria.

Model	DF	− 2*LLmax	2*AIC	BIC	− 2*LPP	PP	XP
ABSOLUTE	3	307 ± 20	319 ± 20	325 ± 20	314 ± 20	0.08 ± 0.03	0.0
RELATIVE	4	295 ± 22	311 ± 22	319 ± 22	304 ± 21	0.92 ± 0.03	1.0

AIC, Akaike Information Criterion (computed with LLmax); BIC, Bayesian Information Criterion (computed with LLmax); DF, degrees of freedom; LLmax, maximal log likelihood; LPP, log of posterior probability; PP, posterior probability of the model given the data; XP, exceedance probability (computed from LPP).
The table summarizes for each model its fitting performances.

Table 3 | Computational free parameters.

	LL maximization		LPP maximization	
Free parameter	**ABSOLUTE**	**RELATIVE**	**ABSOLUTE**	**RELATIVE**
Inverse temperature (β)	17.4 ± 5.92	21.52 ± 5.95	11.4 ± 0.97*	13.66 ± 1.32*
Factual learning rate (α_1)	0.28 ± 0.02	0.19 ± 0.02	0.29 ± 0.02*	0.20 ± 0.01*
Counterfactual learning rate (α_2)	0.18 ± 0.02	0.15 ± 0.02	0.20 ± 0.02*	0.16 ± 0.02*
Context learning rate (α_3)	—	0.33 ± 0.07	—	0.34 ± 0.07*

ABSOLUTE, absolute value learning model; RELATIVE, relative value learning model (best-fitting model); LL maximization, parameters obtained when maximizing the negative log likelihood; LPP maximization, parameters obtained when maximizing the negative log of the Laplace approximation of the posterior probability.
The table summarizes for each model the likelihood maximizing ('best') parameters averaged across subjects. Data are expressed as mean ± s.e.m.
The average values retrieved from the LL maximization procedure are those used to generate the parametric modulators of GLM1a and GLM1b.
*$P < 0.001$ when correlating the LPP-based with LL-based free parameters (robust regression, $N = 28$).

subjects' history of choices and outcomes at trial $t-1$ (Fig. 2a) and the individual best-fitting free parameters. We submitted model-simulated choice probabilities to the same statistical analyses reported above for their model-free counterpart. The RELATIVE model's choices showed no effect of outcome valence ($F = 0.7$, $P > 0.7$), a significant effect of feedback

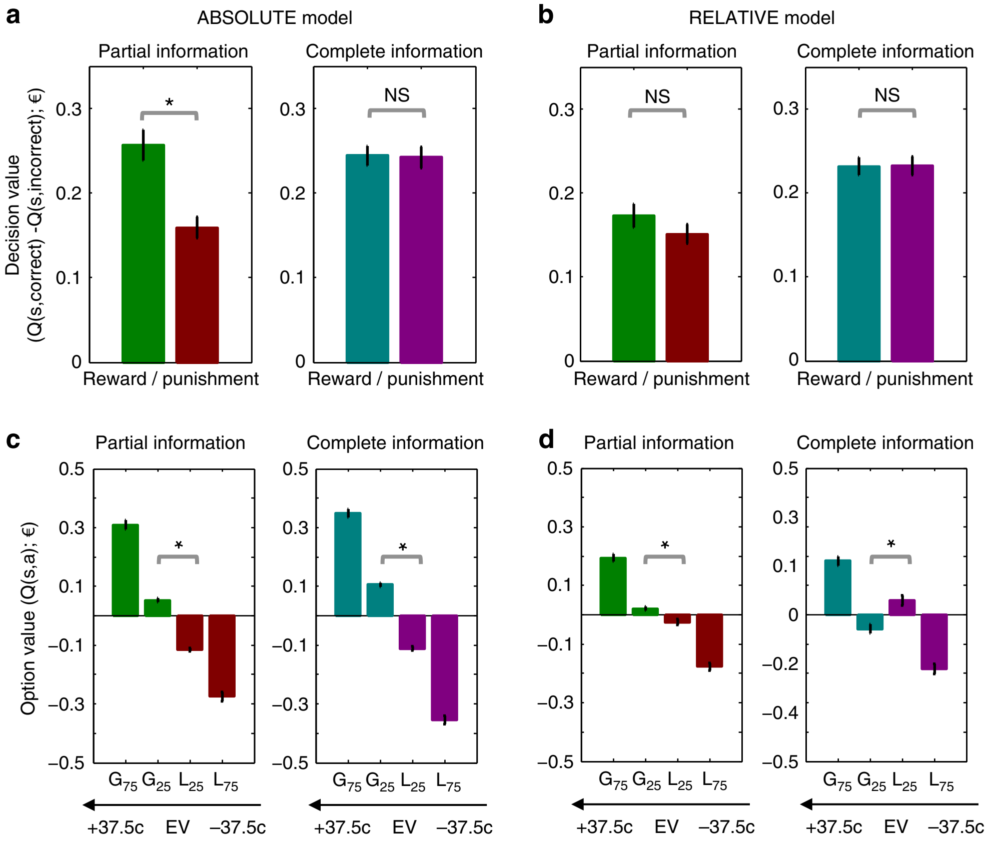


Figure 4 | ABSOLUTE and RELATIVE model final value estimates. (**a**,**b**) The bars represent, for each model, the final optimal decision value estimates (the value of the correct minus the value of the incorrect option). (**c**,**d**): the bars represent, for each model, the final option value estimates.. G_{75} and G_{25}: options associated with 75% and 25% per cent of winning 0.5€, respectively; L_{75} and L_{25}: options associated with 75 and 25% per cent of losing 0.5€, respectively. EV: absolute expected value (Probability(outcome) * Magnitude(outcome)) in a single trial. The values $+37.5$¢ and -37.5¢ correspond G_{75} and the L_{75} options, respectively. The estimates are generated from individual history of choices and outcomes and subject-specific free parameters. $^*P<0.05$ one sample t-test; ns: not significant ($N=28$). Error bars represent s.e.m.

information (F = 53.4, $P<0.001$), and no significant interaction (F = 0.7, $P>0.4$): the same statistical pattern as the actual data. The ABSOLUTE model choices displayed a significant effect of outcome valence (F = 7.0, $P<0.05$), a significant effect of feedback information (F = 43.1, $P<0.001$), and a significant interaction (F = 4.2, $P<0.05$): a different statistical pattern compared to the actual data (Fig. 2c). A *post hoc* test showed lower performances in punishment/partial compared with the reward/partial context ($T=2.4$, $P<0.05$; Table 1). In fact, in the ABSOLUTE model, the model's estimate of the decision value—defined as the difference between the correct and the incorrect option value—was significantly reduced in the punishment/partial compared with the reward/partial context ($+15.9\pm1.2$¢ versus $+25.6\pm1.8$¢; $T=6.5$, $P<0.001$; Fig. 4a). This naturally emerged from the reduced sampling of the G_{25} and L_{75} options respectively, induced by correct responding. This effect formally instantiates the computational problem inherent to punishment avoidance. This effect is not present in the RELATIVE model, in which, thanks to option value centring (that is, $R_{C,t}$—$V_t(s)$ in δ_C; and $R_{U,t}$—$V_t(s)$ in δ_U), decision values were similar in the reward and punishment domains (final decision values: $+17.3\pm1.4$¢ versus $+15.1\pm1.2$¢; $T=1.7$, $P>0.1$; Fig. 4b). Thus, as predicted from the analysis of model-derived option values, absolute value learning suffers from not being able to adequately fit symmetrical performances in the reward and punishment domains. The introduction of value contextualization proved sufficient to obviate this deficiency (Table 1 and Fig. 2c).

Relative value encoding explains post-learning choices. To further probe the explanatory power of context-dependent (relative) over absolute value learning, we assessed and compared their ability to explain post-learning test choices (Fig. 2b). First, we found that the cumulative log-likelihood of the post-learning test was significantly higher assuming choices based on final option values obtained by the RELATIVE, compared with those by the ABSOLUTE model (-172.1 ± 11.5 versus -220.3 ± 16.7; $T=7.0$, $P<0.001$; predictive performances). Second, the post-learning choices simulated with the ABSOLUTE option values, produced a different behavioural pattern than the actual choices, specifically failing to capture the value inversion between intermediate value options (G_{25} and L_{25}) in the complete feedback contexts (generative performances). Indeed, a two-way ANOVA on the RELATIVE simulated choices limited to the intermediate value options, with valence and feedback information as factors, showed, no significant main effect of valence (F = 2.5, $P>0.1$), in line with actual data. The same analysis applied to ABSOLUTE simulated choices produced a significant effect of valence (F = 660.2, $P<0.001$), contrary to actual data (Fig. 2d). *Post hoc* tests showed that the RELATIVE model fitted significantly higher choice rate for the complete L_{25} option compared with the complete G_{25} as observed in the behavioural data ($T=3.4$, $P<0.001$), whereas the ABSOLUTE model generated a significant opposite effect ($T=19.2$, $P<0.001$; Table 1). In fact, because of the additional (counterfactual) information provided to subjects, choice context values were better resolved in the complete compared with the partial feedback information contexts (final

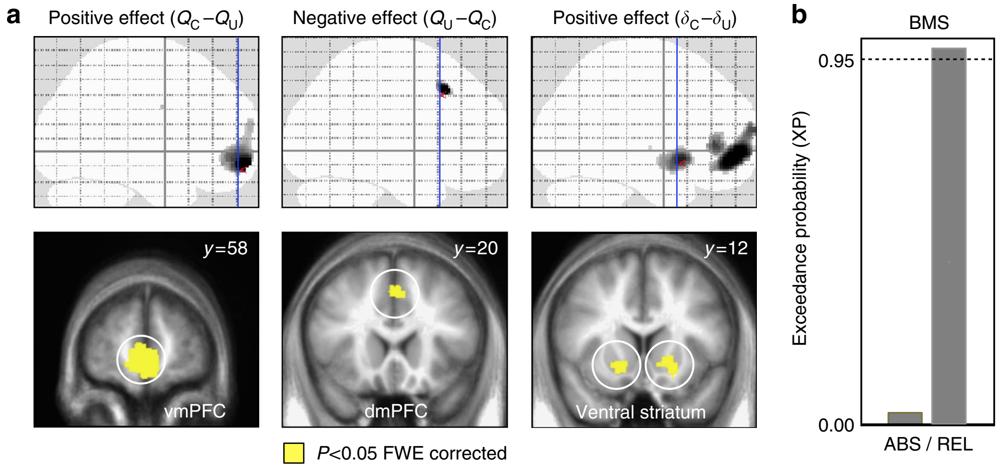

a
Positive effect (Q_C–Q_U)
Negative effect (Q_U–Q_C)
Positive effect (δ_C–δ_U)
y=58
y=20
y=12
vmPFC
dmPFC
Ventral striatum
🟨 P<0.05 FWE corrected
b
BMS
0.95
0.00
Exceedance probability (XP)
ABS / REL

Figure 5 | Neural model comparison. (**a**) Brain areas correlating positively and negatively with the difference between chosen and unchosen option value (Q_C-Q_U; left and central column), and correlating positively with the difference between chosen and unchosen prediction error (δ_C-δ_U; right column). Significant voxels are displayed on the glass brains (top) and superimposed to slices of the between-subjects averaged anatomical T1 (bottom). Coronal slices correspond to the blue lines on sagittal glass brains. Areas coloured in gray-to-black gradient on glass brains and in yellow on slices showed a significant effect ($P<0.05$, voxel level FWE corrected). Y coordinates are given in the MNI space. The results are from the GLM using the ABSOLUTE model parametric modulators (GLM1a). (**b**) Bayesian model comparison (BMS) of GLMs regressing ABSOLUTE (ABS) and RELATIVE (REL) option values and prediction errors (GLM1a and GLM1b). BMS is performed within the functional ROIs, presented on the left in yellow on the brain slices. Note that ROI selection avoids double dipping in favour of the hypothesis we aimed to validate, since the ROIs were defined from GLM1a (ABS) and GLM1a (ABS) was the hypothesis we aimed to reject.

reward minus punishment context values: $\Delta V_{\mathrm{Complete}} = +33.6 \pm 2.3$¢ versus $\Delta V_{\mathrm{Partial}} = +22.4 \pm 3.4$¢; $T=6.9$, $P<0.001$; Supplementary Fig. 3A and 4A). As a direct consequence, contextual influences on option values were more pronounced in the complete feedback contexts. Indeed, intermediate value options (G_{25} and L_{25}) in the complete feedback contexts displayed a more pronounced deviation from absolute expected value encoding (Fig. 4c,d). More precisely G_{25} options acquired a negative value (-4.8 ± 1.6¢; $T=2.9$, $P<0.01$), whereas L_{25} a positive one ($+4.9 \pm 1.7$¢; $T=3.0$, $P<0.01$). Thus, as predicted from the analysis of model-derived option values, absolute value learning and encoding suffers from not being able to adequately fit the value inversion between intermediate value options in the complete context. Again, the introduction of value contextualization proved sufficient to obviate this deficiency (Table 1 and Fig. 2d).

Neural Bayesian model selection. After showing at multiple behavioural levels that option value contextualization occurs, we turned to corroborate this claim using model-based fMRI[41]. To achieve this, we devised a general linear model (GLM1) in which we modelled as separated events the choice onset and the outcome onset, each modulated by different parametric modulators: chosen and unchosen option values (Q_C and Q_U) and prediction errors (δ_C and δ_U). In a first GLM (GLM1a) we regressed the computational variables derived from the ABSOLUTE model. In a second GLM (GLM1b) we used the estimates from the RELATIVE model. We used the GLM1a to generate second level contrasts and, replicating previous findings, we found brain areas significantly correlating with the decision value (Q_C–Q_U) both positively (vmPFC) and negatively (dmPFC,), and brain areas correlating with the decision prediction error (δ_C–δ_U; vmPFC and ventral striatum: VS; $P<0.05$, whole brain family-wise error (FWE) corrected; Fig. 5a and Table 4; see also Supplementary Fig. 5)[3,27,42,43]. In a second step, we estimated within this prefrontal and striatal areas the same GLMs using Bayesian statistics. We found that the context-dependent value encoding (GLM1b) provided a significantly better account of the network's neural activity (1,511 voxels; XP = 0.97; Fig. 5b)[44]. Importantly this result also held true for each region of interest (ROI) separately (vmPFC: 936 voxels, XP = 0.87; dmPFC: 71 voxels, XP = 0.97; VS: 505 voxels, XP = 0.93; for the RELATIVE model. Thus, replicating previous imaging findings implicating the medial prefrontal cortex and the striatum in value learning and decision making, we found that neural activity in these areas supports context-dependent (GLM1b) as opposed to absolute (GLM1a) value signals. Note that the ROIs were selected to favour the hypothesis that we want to reject (GLM1a)[45].

vmPFC activity is consistent with relative value encoding. Model-based Bayesian fMRI analyses corroborated the RELATIVE model. To further support relative value encoding from the neural perspective, we also devised a categorical GLM (GLM2), in which choice events were modelled separately for each context and learning phase (early: first eight trials; late: last eight trials). For this analysis we focused on the vmPFC: the region that has been more robustly implicated in value encoding[11,12]. To avoid double dipping, we used a literature-based independent vmPFC ROI. On the basis of the model predictions and the assumption that the vmPFC represents values signals, we expected higher activation in the punishment/complete late trials (once the correct option L_{25} of the punishment/complete context has acquired a positive value), compared with reward/complete early trials (when the option values are not yet very different from zero). On the other side, we expected no such a difference in the partial contexts. To test this hypothesis we submitted the choice-related regression coefficients to a three-way ANOVA with valence (reward and punishment), feedback information (partial and complete) and learning phase (early and late) as factors (Fig. 6a). We found a significant main effect of phase (F = 11.6, $P<0.01$), reflecting an overall learning-induced increase of vmPFC signal. We also found a significant main effect of valence (F = 11.4, $P<0.01$), and a significant valence x information interaction (F = 17.3, $P<0.001$), indicating that valence did not affect choice activations similarly in the

Table 4 | Brain activations.

Contrast	Label	[*x y z*]	BA	AAL	*T*	*S*	GLM
Q_C-Q_U							
	vmPFC	[4 58 − 12]	10,11	Medial frontal gyrus, pars orbitalis	6.57	939	1
Q_U-Q_C							
	dmPFC	[− 6 20 42]	8,32	Superior medial frontal gyrus	5.63	71	1
δ_C-δ_U							
	vmPFC	[− 6 54 − 4]	10,11	Medial frontal gyrus (pars orbitalis)	7.00	1226	1
	vlPFC	[− 52 36 2]	45,47	Inferior frontal gyrus (pars triangularis)	5.79	119	1
	Left-VS	[− 16 12 − 8]	—	Putamen, pallidum	4.78	271	1
	Right-VS	[14 10 − 8]	—	Putamen, pallidum	4.57	234	1
+ 0.5€ > 0.0€ (reward/partial)							
	Right-VS	[16 12 − 6]	—	Putamen, pallidum	3.89	21	3
− 0.5€ > 0.0€ (punishment/partial)							
	AI	[16 12 − 6]	48	Insula	4.02	43	3

AAL, automatic anatomic labelling; AI, anterior insula; BA, Brodmann area; dmPFC; dorso-medial prefrontal cortex; GLM, general linear model; *S*, size of the activation (voxels); *T*, *t*-values of the maxima; vmPFC, ventro-medial prefrontal cortex; VS, ventral striatum; [*x y z*], MNI coordinates.
The table summarizes brain activations reported in Fig 5a and Supplementary Fig. 6a,b, significant at $P<0.05$ FWE whole brain-level (GLM1a) or anatomic mask-level (GLM3) FWE corrected (one-sample *t*-test; $N=28$).

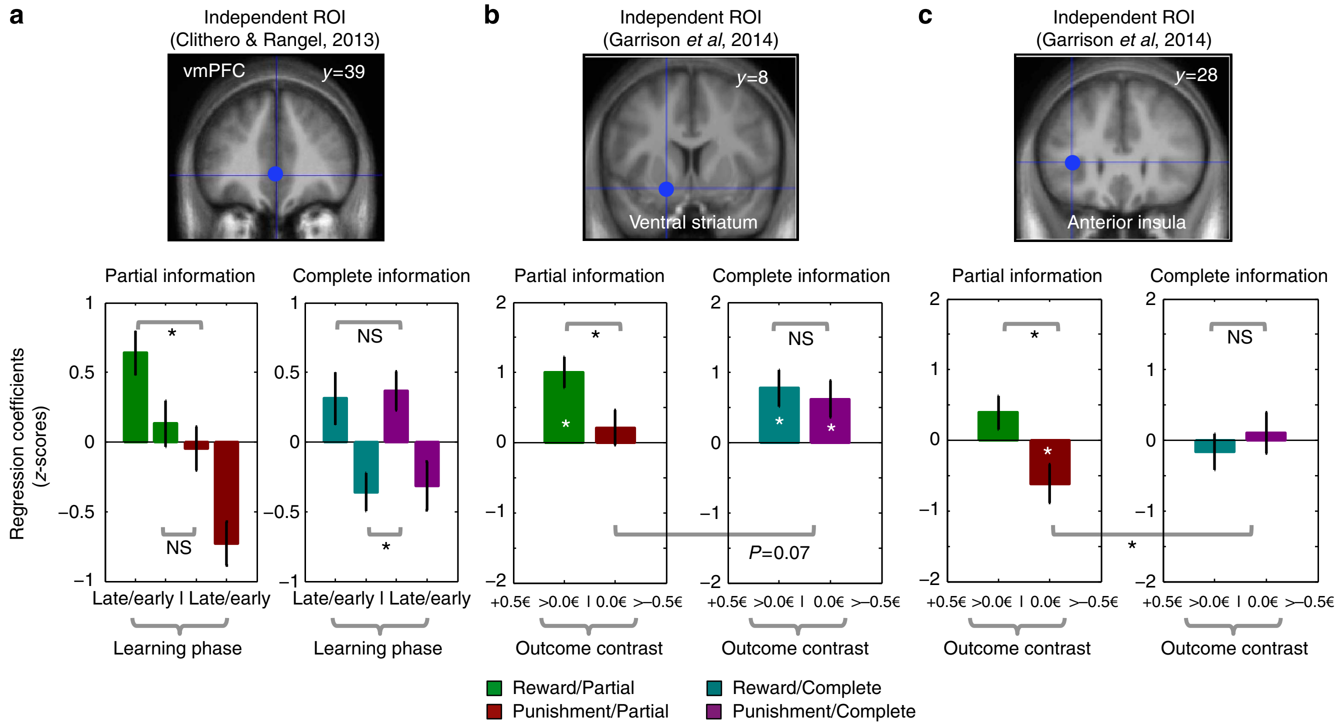


Figure 6 | Model-free neural evidence of value contextualization. (**a**) Bars represent the regression coefficients extracted in the ventromedial prefrontal cortex, as a function of the task contexts (represented by different colours) and leaning phase (early: first eight trials; late: last eight trials). Regression coefficients are extracted from the model-free GLM2 within a sphere centered on literature-based coordinates of the ventromedial prefrontal cortex[11]. (**b**) & (**c**) Bars represent the regression coefficients for best > worst outcome contrast as a function of the task contexts. ('+ 0.5€ > 0.0€': best > worst outcome contrast in the reward contexts; '0.0€ > − 0.5€': best > worst outcome contrast in the punishment contexts). Regression coefficients are extracted from the model-free GLM3 within spheres centered on literature-based coordinates of the striatum and anterior insula[8]. *Y* coordinates are given in the MNI space. Note that ROI selection avoids double dipping, since the ROIs were defined from independent studies (metanalyses). *$P<0.05$ one sample *t*-test comparing between regressors (black '*') or to zero (white '*'; $N=28$); NS: not significant. Error bars represent s.e.m.

partial and complete contexts, respectively. Consistent with this valence *x* information interaction, *post hoc* test indicated significant higher activations in the reward compared to the punishment late trials in the partial contexts ($T=3.5$, $P<0.01$), but no such difference in the complete contexts ($T=0.7$, $P>0.4$). Crucially and consistent with our predictions, *post hoc* test also indicated significant higher activations in the punishment early trials compared to the reward late trials in the complete contexts ($T=3.8$, $P<0.001$), but no such difference in the partial contexts ($T=0.2$, $P>0.8$). This result closely resembles that of option value inversion in the post-learning test. In summary, in addition to model-based fMRI analyses, we found that the activation pattern of the vmPFC is still consistent with relative, rather than absolute value encoding, also when analysed in a model-free manner.

Outcome encoding is modulated by contextual discrimination. Previous fMRI and lesions studies, using similar behavioural tasks, suggest a role for the anterior insula (AI) in punishment learning, in contrast to that of the VS in the reward domain[3,6,20,46–48]. To challenge this hypothesis, we analysed outcome encoding within an anatomic mask including the insular cortex and the basal ganglia (Supplementary Fig. 6E). In the GLM (GLM3) used for this analyses, outcome events were modelled separately for each context and factual feedback value (R_C). GLM3 was also 'model-free', since the categories were not derived from a computational model, but from the observable outcomes. We computed for each context separately a best > worst outcome contrast. Consistent with the neural opponency hypothesis and replicating previous findings, we found voxels in the VS significantly activated by the +0.5€ > 0.0€ contrast in the reward/partial context, thus encoding obtained rewards, and voxels in the AI significantly deactivated by the 0.0€ > −0.5€ contrast in the punishment/partial context, thus encoding obtained punishments ($P < 0.05$ FWE mask-level corrected; Supplementary Fig. 6A,B and Table 4). This functional dissociation still held at more permissive threshold of $P < 0.001$ uncorrected, and after literature-based independent ROIs test[8]. In fact, to simultaneously and formally assess this functional dissociation as well as the effect of contextual information on outcome encoding, we submitted the outcome related contrasts to a three-way ANOVA with valence (reward and punishment) feedback information (partial and complete) and brain area (VS and AI) as factors (Fig. 6b,c). Indeed, ANOVA indicated a significant main effect of brain system (VS versus AI; F = 45.8, $P > 0.001$), which confirms the fact that outcomes are encoded with opposite signs in the two neural systems. We also found a significant main effect of feedback information (F = 4.2, $P < 0.05$) and a significant valence *x* information interaction (F = 4.7, $P < 0.05$), indicating that valence did not affect outcome signals similarly in the partial and complete contexts, respectively. *Post hoc* testing revealed significant differences in outcome encoding between the reward/partial and the punishment/partial contexts in both the AI ($T = 2.9$, $P < 0.01$), and the VS ($T = 2.3$, $P < 0.05$). Such differences were not observed when comparing the reward/complete to the punishment/complete contexts ($T < 0.7$, $P > 0.4$). Interestingly, *post hoc* tests also revealed that, in the complete feedback contexts, VS significantly encoded avoidance ($T = 2.4$, $P < 0.05$) and, concomitantly, the AI stopped responding to punishments (compared with the partial/punishment context: $T = 2.8$, $P < 0.01$). Finally, the triple valence *x* information *x* brain area interaction was not significant, reflecting the fact that the signal increases similarly in both areas when moving from the partial to the complete feedback contexts (in the striatum, from zero, it becomes positive; in the insula, from negative, it becomes zero; F = 1.9; $P > 0.1$).To further check that the result was not dependent on the (independent) ROI selection, we explored outcome related activations at an extremely permissive threshold ($P < 0.01$ uncorrected), confirming no detectable implication of the AI in the punishment/complete context (Supplementary Fig. 6C,D). Altogether these results show that when additional information is provided (that is, complete feedback contexts), and therefore context value is better identified, punishment avoidance signals converge to the VS allowing the opponent system to 'switch off'.

Discussion

Healthy subjects performed an instrumental conditioning task, involving learning to maximize rewards and minimize punishments. Orthogonally to outcome valence, complete feedback information (the outcome of the chosen and the unchosen option) was provided to the subjects, in order to promote relative value encoding. The data displayed convergent evidence of option value contextualization at two independent behavioural levels: instrumental choices, and post-learning choices. First, punishment avoidance performances were matched to reward seeking ones, a result that cannot be explained by absolute value encoding; second, post-learning evaluation of the instrumental options, especially for those of the complete feedback contexts, displayed significant biases that can be parsimoniously explained assuming relative value encoding.

All these behavioural effects were submitted to computational model-based analyses. More specifically our analyses compared models representing two opposite views of the signals that drive decision-making: context-independent absolute value signals (that is, Q-learning) and context-dependent relative value signals (RELATIVE)[25]. We made a deliberate effort to keep these models as simple and parsimonious as possible. The RELATIVE model essentially tracks the mean of the distribution of values of the choice context (that is, the reference point) and uses it to centre option values. Notably, this model represents a minimal departure from a standard reinforcement learning algorithms that imply context or option values are updated with a delta rule, such as the Q-learning and actor–critic[34]. On the other side, the RELATIVE model can be seen as the most parsimonious algorithm implementation of a model that, departing from experienced raw values, learns to identify, for each situation (context) the 'best' and the 'worst' possible outcomes, based on an explicit representation of the underlying generative process (the task structure)[49,50].

Punishment avoidance is computationally challenging. Simply stated: how can the instrumental response (avoid a punishment) be maintained despite the absence of further extrinsic reinforcement (punishment)? As already known and replicated here, absolute value learning methods are structurally not capable to cope with this problem[37,38]. In fact, the ABSOLUTE model predicted significant higher performances in the reward compared with the punishment context. Psychological models, such as the two-factor theory, suggested that a successful punishment avoidance could acquire a positive value and therefore act as intrinsic reinforcement to sustain learning[17–20,22]. The RELATIVE model embodies this idea by considering outcomes relative to the context in which they were delivered (R_C–V). As a consequence of this feature, successful punishment avoidance (the neutral outcome 0.0€), acquired a positive value in the punishment avoidance context (where V is negative), providing a substrate for reinforcing the response. By doing so, it managed to equalize the performances between the reward and punishment context, as observed in human subjects.

We probed relative value encoding with an additional, and independent, behavioural measure. As in previous studies, we asked subjects to retrieve the value of the options after learning[29,30]. In this last task, options were presented in all possible combinations and were therefore extrapolated from their original choice context. Post-learning choices showed clear signs of value encoding. In fact, G_{75} choice rate was higher compared to L_{25} choice rate, despite the fact that their instrumental choice rate was similar. However, more in-depth analyses indicated that the behavioural pattern was more consistent with relative, rather than absolute value encoding. Subjects indeed failed to correctly retrieve the value of intermediate value options, to the point of preferring a lower value option (L_{25}) to a higher value option (G_{25}) in the complete feedback information, where relative value encoding was enhanced. Importantly, only the RELATIVE model was able to precisely capture this choice pattern (out-of-sample validation). The across task stability of relative value further corroborated our assumptions regarding the model, namely that value contextualization occurs within the learning rule and not within the policy. This effect is reminiscent of choice irrationality induced by context dependency (that is, preference reversal or ‘less is more’ effect) as if the adaptive function of value contextualization (in our case coping with punishment avoidance in the learning tasks) was traded against a bias of value estimates in the post-learning test[51,52]. Thus, as far as we were able to infer option values from choice data, they showed signs of context-dependency.

Replicating previous results, we found neural correlates of option values and prediction errors in a well-established reinforcement learning and decision-making network, including cortical and subcortical brain regions[3,27,42,43]. Relative and absolute value regressors shared a significant part of their variance due to the fact that both depend on the same history of choices and outcomes, and that the two models are structurally similar (nested) and similarly affected by task factors. Given these premises, to overcome this issue and corroborate relative value encoding, we implemented, as in recent studies, a neural model comparison analysis[53,54]. Bayesian model comparison showed that, within this prefrontal-striatal network, the context-dependent value-learning model (RELATIVE) provided a better explanation for BOLD responses than the ABSOLUTE model, which was used to generate the ROIs. Everything else being constant (individual history of choices and outcomes), the difference in model evidence could only be attributed to the value contextualization process itself, and therefore corroborates behavioural model selection. This model-based fMRI result has been backed up by a model-free analysis showing that signal changes in vmPFC (the core of the brain valuation system), once decomposed as a function of learning phase and task factors, displayed a pattern fully compatible with relative value encoding. More precisely we found that late vmPFC signal in punishment/complete context, was higher compared to the early signal reward/partial context: an effect that closely resembles that of the post-leaning value inversion.

Our finding of a functional double dissociation between the striatum and the insula in positive and negative outcome encoding perfectly replicates our previous results, and adds to a now critical mass of studies suggesting the existence of an opponent insular system dedicated to punishment-based learning and decision making[3,6,20,46–48,55]. Indeed, we found that the AI represented received negative outcomes in the punishment/partial context, in opposition to the pattern of activity in the ventral striatum, which represented received positive outcomes in the reward/partial condition. Strikingly, we found that in the punishment/complete context, negative outcome encoding in the AI faded, while the ventral striatum was concomitantly taking over. Globally, these results suggest that, by default, positive and negative values are represented in opposite directions by two opponent channels to ensure optimal outcome encoding in face of the impossibility of negative firing rates[56–58]. They also indicate that when relative valuation is facilitated (here in presence of complete feedback information), the ventral system is tuned to respond to ‘successful avoidance’ (intrinsic reinforcement) as it does for rewards[20,22]. This suggests that value contextualization can limit the need for simultaneously mobilizing multiple neural systems and therefore promotes neural parsimony. In our design, this effect was achieved in presence of the complete feedback information. Accordingly counterfactual outcome processing has been tightly associated to context-dependent (that is, relative) decision-making models, such as the regret theory[24–26]. However, it is nonetheless possible that in previous studies other task features, such as blocked design or explicit information about the outcome contingencies, could have concurred to reframe punishment avoidance tasks in order to induce the striatum to respond to successful avoidance[36,59–62].

To summarize, our data suggest that as soon as an agent is engaged in a utility maximization-learning task, (s)he learns concomitantly the value of the available options and the value of the choice context in which they are encountered (the reference point). These quantities, option and context values, do not remain segregated but are rather integrated, so that the option value, originally encoded in an absolute scale, becomes relative to their choice context. Our study shows how value contextualization has the adaptive function of permitting efficient avoidance learning. Nevertheless, option value, being learned in a context-dependent manner, can produce suboptimal preferences (value inversion: irrational behaviour) when the options are extrapolated from their original choice context (for example, post learning). In the brain, value updating, supposedly achieved via prediction errors, is originally implemented by two different systems for the reward (reward system: ventral striatum) and the punishment (opponent system: anterior insula) domains, respectively, to obviate the difficulty to efficiently encode a large range of negative values. As a result of value contextualization, the reward responds to successful avoidance (*per se* a neutral outcome) and concomitantly the activity in the opponent system is suppressed.

Methods

Subjects. We tested 28 subjects (16 females; age 25.6 ± 5.4 years). Power calculation studies suggested that a statistically valid sample size for fMRI study should be comprised of between 16 and 24 subjects[63]. We included $N=28$ subjects based on a pessimistic drop-out rate of 15%. We experienced no technical problems, so we were able to include all 28 subjects. Subjects were screened for the absence of any history of neurological and psychiatric disease or any current psychiatric medication, for right handedness and for normal or correct to normal vision. The research was carried out following the principles and guidelines for experiments including human participants provided in the declaration of Helsinki (1964). The local Ethical Committee of the University of Trento approved the study and subjects provided written informed consent prior to their inclusion. To sustain motivation throughout the experiment, subjects were remunerated according to the exact amount of money won in the experiment plus a fixed amount for their travel to the MRI center.

Behavioural tasks. Subjects performed a probabilistic instrumental learning task adapted from previous imaging and patient studies[3,6,27,28]. Subjects were first provided with written instructions, which were reformulated orally if necessary (see Supplementary Note 3). They were informed that the aim of the task was to maximize their payoff, that reward seeking and punishment avoidance were equally important and that only factual (and not counterfactual) outcomes counted. Prior to entering the scanner, subjects performed a shorter (training) session, aimed to familiarize them with the task’s timing and responses. In the scanner subjects performed four learning sessions. Options were abstract symbols taken from the Agathodaimon alphabet. Each session contained eight novel options divided into four novel fixed pairs of options. The pairs of options were fixed so that a given option was always presented with the same other option. Thus, within each session, pairs of options represented stable choice contexts. Within sessions, each pair of options was presented 24 times for a total of 96 trials. The four option pairs corresponded to the four contexts (reward/partial, reward/complete, punishment/partial and punishment/complete), which were associated with different pairs of outcomes (reward contexts: winning 0.5€ versus nothing; punishment contexts: losing 0.5€ versus nothing) and a different quantity of information being given at feedback (partial and complete). In the partial feedback contexts, only the outcome about the chosen option was provided, while in the complete feedback contexts both the outcome of the chosen and the unchosen option were provided. Within each pair, the two options were associated to the two possible outcomes with reciprocal probabilities (0.75/0.25 and 0.25/0.75). During each trial, one option was randomly presented on the left and one on the right side of a central fixation cross. Pairs of options were presented in a pseudorandomized and unpredictable manner to the subject (intermixed design). The side on which a given option was presented was also pseudorandomized, such that a given option was presented an equal number of times in the left and the right of the central cross. Subjects were required to select between the two options by pressing one of the corresponding two buttons with their left or right thumb to select the leftmost or the rightmost option, respectively, within a 3,000 ms time window. After the choice window, a red pointer appeared below the selected option for 500 ms. At the end of the trial the options disappeared and the selected one was replaced by the outcome (‘+0.5€’, ‘0.0€’ or ‘−0.5€’) for 3,000 ms. In the complete information contexts, the outcome corresponding to the unchosen option (counterfactual) was also displayed. Note that between cues the outcome probability was truly independent on a trial-by-trial basis, even if it was anti-correlated in average. Thus, in a complete feedback trial, subjects could observe the same outcome from both cues on 37.5% of trials and different outcomes from each cue on 62.5% of trials. A novel trial started after a fixation screen (1,000 ms, jittered between 500–1,500 ms). During the anatomical scan and after the four sessions subjects performed a post-learning assessment of option value. This task involved only the 8 options (2 × 4 pairs) of the last session, which were presented in all possible pair-wise combinations (28, not including pairs formed by the same option)[29,30]. Each pair of options was presented 4 times, leading to a total of 112 trials. Instructions were provided orally after the end of the last learning session. Subjects were informed that they would be presented pairs of options taken from the last session, and that all pairs had not necessarily been displayed together before. During each trial, they had to indicate the option with the highest value by pressing on the buttons as they had done during the learning task. Subjects were also advised that there was no money at stake, but encouraged to respond as they would have if that were the case. In order to prevent explicit memorizing strategies, subjects were not informed that they would have performed this task until the end of the fourth (last) session of the learning test. Timing of the post-test differed from the learning test in that the choice was self-paced and in the absence of the outcome phase.

Behavioural analyses. From the learning test, we extracted the choice rate as dependent variable. Statistical analyses were performed on the percentage of correct choices, i.e., choices directed toward the most advantageous stimulus (most rewarding or the less punishing), sorted as a function of the context (see Behavioral tasks). Statistical effects were assessed using two-way repeated-measures ANOVA with (1) feedback information and (2) feedback valence as factors. Between-context differences in correct responses were also tested *post hoc* using a two-sided, one-sample *t*-test. Reaction times were also extracted from the learning test and submitted to the same factorial analyses used for the correct choice rate (see Supplementary Note 1 and Supplementary Fig. 1). Choice rate was also extracted from the post-learning test and sorted for each option separately, as the percentage of choice toward a given stimulus taking into account all possible comparisons. Post-learning choice rate was submitted to three-way repeated-measures ANOVA, to assess the effects of (1) feedback information, (2) feedback valence and (3) option correctness. We also performed a two-way repeated-measures ANOVA focused on the intermediate value options, assessing the effect of (1) feedback information and (2) valence. Between-option differences in post-learning choices were tested *post hoc* using a two-sided, one-sample *t*-test. As a control analysis, the percentage of direct choices involving the G_{25} and the L_{25} cues (that is, the intermediate value cues) has also been analysed separately for each comparison (see Supplementary Note 1 and Supplementary Fig. 2). All statistical analyses were performed using Matlab (www.mathworks.com) with the addition of the Statistical toolbox and other free-download functions (rm_anova2.m, RMAOV33.m).

Computational models. We analysed our data with model-free reinforcement learning algorithms[34]. The goal of all models was to find in each choice context (state: s) the option that maximizes the cumulative reward R. We compared two alternative computational models: a Q-learning model, extended to account for counterfactual learning (ABSOLUTE), which instantiates 'absolute value-based' learning and decision making by learning option values independently of the choice context in which they are presented[25,34]; the RELATIVE model which learns option values relative to the choice context in which they are presented[35–38,64] (Fig. 3).

(1) ABSOLUTE model

At trial t the chosen (c) option value of the current context (s) is updated with the Rescorla-Wagner rule (also called delta-rule)[65]:

$$Q_{t+1}(\mathrm{s,c}) = Q_t(\mathrm{s,c}) + \alpha_1\delta_{\mathrm{C,t}}$$

and

$$Q_{t+1}(\mathrm{s,u}) = Q_t(\mathrm{s,u}) + \alpha_2\,\delta_{\mathrm{U,t}},$$

where α_1 is the learning rate for the chosen option and α_2 the learning rate for the unchosen (u) option (counterfactual learning rate). δ_{C} and δ_{U} are prediction error terms calculated as follows:

$$\delta_{\mathrm{C,t}} = R_{\mathrm{C,t}} - Q_t(\mathrm{s,c})$$

(update in both the partial and complete feedback contexts) and

$$\delta_{\mathrm{U,t}} = R_{\mathrm{U,t}} - Q_t(\mathrm{s,u})$$

(in the complete feedback contexts only).

(2) RELATIVE model

We also devised a new model (RELATIVE), which, instantiates the 'relative value-based' learning and decision-making. The key idea behind RELATIVE model is that it separately learns and tracks the choice context value (V(s)), used as the reference point to which an outcome should be compared before updating option values. Previous algorithms, such the actor-critic and the advantage learning model, inspired the RELATIVE model (see Supplementary Note 2, Supplementary Fig. 3 and Supplementary Table 3 for additional model comparison analyses including the actor-critic model). All these models implement relative value learning of option values, based on V(s) estimates. The RELATIVE model differs in that it is extended to account for counterfactual feedback and that V(s) is learnt in an 'random-policy' manner (that is, the state value is independent from the policy followed by the subject. (see Supplementary Note 2, Supplementary Fig. 4 and Supplementary Table 4 for additional model comparison analyses supporting these assumptions). Crucially V(s) is not merely the choice-probability weighted sum of options' value, but rather affects (controls) them.

In fact V(s) is used to centre option prediction errors as follows:

$$\delta_{\mathrm{C,t}} = \mathrm{R_{C,t}} - V_{\mathrm{t}}(\mathrm{s}) - Q_t(\mathrm{s,c})$$

and

$$\delta_{\mathrm{U,t}} = R_{\mathrm{U,t}} - V_{\mathrm{t}}(\mathrm{s}) - Q_{\mathrm{t}}(\mathrm{s,u})$$

(in the complete feedback contexts only). As a consequence the option values are no longer calculated in an absolute scale, but relatively to their choice context value V(s). Context value is also learned via a delta rule:

$$V_{\mathrm{t+1}}(s) = V_{\mathrm{t}}(s) + \alpha_3\delta_{\mathrm{V,t}}$$

Where α_3 is the context value learning rate and δ_{V} is a prediction error-term calculated as follows:

$$\delta_{\mathrm{V,t}} = R_{\mathrm{V,t}} - V_{\mathrm{t}}(\mathrm{s})$$

where *t* is the number of trials and R_{V} is the context-level outcome at trial *t*: a global measure that encompasses both the chosen and unchosen options. In the complete feedback contexts the average outcome trial (R_{V}) is calculated as the average of the factual and the counterfactual outcomes as follows:

$$R_{\mathrm{V,t}} = \left(R_{\mathrm{C,t}} + R_{\mathrm{U,t}}\right)/2.$$

Given that the average outcome trial (R_{V}) is meant to be a context-level measure, in order to incorporate unchosen option value information in R_{V} also in the partial feedback contexts, we considered $\mathrm{Q_t(s,u)}$ a good proxy of $R_{\mathrm{U,t}}$ and calculated $R_{\mathrm{V,t}}$ as follows (see Supplementary Note 2 and Supplementary Table 2 for model comparison justifications of these assumptions):

$$R_{\mathrm{V,t}} = \left(R_{\mathrm{C,t}} + \mathrm{Q_t(s,u)}\right)/2.$$

To sum up, our model space included 2 models: the ABSOLUTE model (Q-learning) and the RELATIVE model. In all models decision-making relied on a softmax function:

$$P_{\mathrm{t}}(\mathrm{s,a}) = \left(1 + \exp(\beta(\mathrm{Q_t(s,b)} - \mathrm{Q_t(s,a)}))\right)^{-1},$$

where β is the inverse temperature parameter. The Matlab codes implementing the computational models are available upon request to the corresponding author.

Parameters optimization and model selection procedures. We optimized model parameters, the temperature (β), the factual (α_1), the counterfactual (α_2) and the contextual (α_3) learning rates (in the RALATIVE model only), by minimizing the negative log likelihood ($\mathrm{LL_{max}}$) and (in a separate optimization procedure) the negative log of posterior probability (LPP) of the data given different parameters settings using Matlab's fmincon function, initialized at multiple starting points of the parameter space, as previously described[66,67]. Negative log-likelihoods ($\mathrm{LL_{max}}$) were used to compute classical model selection criteria. The LPP was used to compute the exceedance probability and the expected frequencies of the model.

We computed at the individual level (random effects) the Akaike's information criterion (AIC),

$$\mathrm{AIC} = 2df + 2\times\mathrm{LL_{max}},$$

the Bayesian information criterion (BIC),

$$\mathrm{BIC} = \log(\mathrm{ntrials})\times df + 2\times\mathrm{LL_{max}}$$

and the Laplace approximation to the model evidence (LPP);

$$\mathrm{LPP} = \log(P(D\,|\,M,\theta))),$$

where *D, M* and θ represent the data, model and model parameters respectively. $P(\theta_{\mathrm{n}})$ is calculated based on the parameters value retrieved from the parameter optimization procedure, assuming learning rates beta distributed (betapdf(parameter,1.1,1.1)) and softmax temperature gamma-distributed (gampdf(parameter,1.2,5))[68]. The present distributions have been chosen to be relatively flat over the range of parameters retrieved in the previous and present studies. The LPP increases with the likelihood (which measures the accuracy of the fit) and is penalized by the integration over the parameter space (which measures the complexity of the model). The LPP, as the BIC or AIC, thus represent a trade-off between accuracy and complexity and can guide model selection. Individual LPPs were fed to the mbb-vb-toolbox (https://code.google.com/p/mbb-vb-toolbox/)[40]. This procedure estimates the expected frequencies of the model (denoted PP) and the exceedance probability (denoted XP) for each model within a set of models, given the data gathered from all subjects. Expected frequency quantifies the posterior probability, i.e., the probability that the model generated the data for any randomly selected subject. This quantity must be compared to chance level (one over the number of models in the search space). Exceedance probability quantifies the belief that the model is more likely than all the other models of the set, or in other words, the confidence in the model having the highest expected frequency. We considered the 'best model', the model which positively fulfilled all the criteria.

Model simulation analyses. Once we had optimized models' parameters, we analysed their generative performance by analysing the model simulation of the data[69]. Model estimates of choice probability were generated trial-by-trial using the best individual parameters in the individual history of choices and outcomes. Model choice probability was then submitted to the same statistical analysis as the actual choices. The evaluation of generative performances involved two steps: first, the assessment of the model's ability to reproduce the key statistical effects of the data; second, the assessment of the model's ability to match subjects' choices. The first step essentially involved within-simulated data comparisons, in both the form of ANOVA and *post hoc* one-sample *t*-test. The second step involved comparison between simulated and actual data with a one-sample *t*-test, and adjusting the significance level for the multiple comparisons (see the results reported in Table 1). We also tested models' performances out of the sample by assessing their ability to account for post-learning test choices. Concerning the post-learning test analysis, under the assumption that choices in the post-learning test were dependent on the final option values, we calculated the probability of choice in the post-learning test using a softmax, using the same individual choice temperature optimized during the learning test (note that similar results have been obtained when optimizing a β specific to the post-learning test). On the basis of model-estimate choice probability, we calculated the log-likelihood of post-learning choices that we compared between computational models. Finally, we submitted the model-estimate post-learning choice probability to the same statistical analyses as the actual choices (ANOVA and *post hoc* *t*-test; within-simulated data comparison) and we compared modelled choices to the actual data (pair-wise comparisons, corrected for multiple comparisons; Table 1).

fMRI data acquisition and preprocessing. A 4T Bruker MedSpec Biospin MR scanner (CiMEC, Trento, Italy) and an eight-channel head coil were used to acquire both high resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1 mm^3 voxel and T2*-weighted Echo planar imaging (EPI). The parameters of the acquisition were the following, 47 slices acquired in ascending interleaved order, the in-plan resolution was 3 mm^3 voxels, the repetition time 2.2 s, and the echo time was 21 ms. A tilted plane acquisition sequence was used to optimize functional sensitivity to the orbitofrontal cortex[70]. The acquisition started from the inferior surface of the temporal lobe. This implicated that, in most subjects, the acquired volume did not include the inferior part of the cerebellum. Preprocessing of the T1-weighted structural images consisted in coregistration with the mean EPI, segmentation and normalization to a standard T1 template, and average across all subjects to allow group-level anatomical localization. Preprocessing of EPI consisted in spatial realignment, normalization using the same transformation as structural images, and spatial smoothing using a Gaussian kernel with a full width a half-maximum of 8 mm. Final voxel size was 2 mm^3. Preprocessing was realized using SPM8 (www.fil.ion.ucl.ac.uk).

fMRI data analyses. EPI images were analysed in an event-related manner within the general linear model (GLM) framework, using SPM8 software. In GLM1, each trial was modelled as having two time points, corresponding to choice and outcome display onsets, modelled by two separate regressors. Choice onset and outcome onset were then modulated with different parametric regressors. In order to account for irrelevant motor or visual activations, the first parametric modulators for GLM1 were: (1) the response (coded as 1 and -1, for the right or left response, respectively) for the choice onset, and (2) the number of outcomes on the screen (codes as 1 and 2 for the partial and complete feedback context, respectively) for the outcome onset. These control parametric modulators generated motor and visual activations (data not shown). To correct for motion artifact, all GLMs also included the subject/session specific realignment parameters as nuisance covariates. The GLM1a and GLM1b differed in the computational model used to generate the parametric modulators. In addition to motor and visual regressors, in GLM1 the choice onsets were modulated by the trial-by-trial estimates of Q_C and Q_U, whereas the outcome onsets by the trial-by-trial estimates of δ_C and δ_U. In the partial feedback trials, the unchosen prediction error regressor (δ_U) was systematically set at zero. Computational regressors were generated for each subject using the group level mean of the best individual parameters and the individual history of choices and outcomes. Regressors were z-scored before regression in order to ensure between-model, between-subject and between-modulator commensurability of the regression coefficients (Table 3). The computational variables of the GLM1a were derived from the ABSOLUTE computational model. GLM1b was structurally identical to GLM1a, except for the fact that the computational variables were derived from the RELATIVE model. All activations concerning GLM1 reported in the figure 5 survived a threshold of $P<0.05$ with voxel level whole brain FWE correction for multiple comparisons. In GLM2, each trial was modelled as having one-time points, corresponding to the stimulus display onsets. The choice onsets were split into eight different events (categories) as a function of task factors (feedback information x outcome valence) and the position of the trial within the learning curve (early: first eight trials; late: last eight trials; we did not include the mid eight trials so as to only include in each category trials belonging as clearly as possible to the 'incremental' versus the 'plateau' phase of the learning curves). In GLM3, each trial was modelled as having one-time points, corresponding to the outcome display onsets. The outcome onsets were split into eight different events (categories) as a function of the task factors (feedback information x outcome valence) and obtained outcome (R_C). We computed at the first level a best > worst outcome contrast for each context separately ('$+0.5€>0.0€$': best > worst outcome contrast in the reward contexts; '$0.0€>-0.5€$': best > worst outcome contrast in the punishment contexts). All GLMs were estimated using classical statistical techniques and linear contrast of parametric modulators were computed at the individual level and then taken to a group-level random effect analysis (one-sample *t*-test). Based on our hypotheses, second level contrasts of GLM3 were estimated within an anatomic mask encompassing bilaterally the insula and the basal ganglia (caudate, putamen and pallidum; $>9\times10^5$ voxels of 2 mm^3) (Supplementary Fig. 6E). The mask has been designed using MARINA software (http://www.fil.ion.ucl.ac.uk/spm/ext/). Activations concerning GLM3 and reported in yellow in Supplementary Fig. 6 survived a threshold of $P<0.05$ with voxel level anatomic mask FWE correction, that is, the multiple comparison accounted for the number of voxels in the mask rather than the whole brain (small volume correction). Activations are reported in the coordinates space of the Montreal Neurology Institute (MNI). Activations were anatomically labelled using the Brodmann and the automatic anatomical labelling template implemented by the software MRIcron (www.mccauslandcenter.sc.edu/mricro).

Region of interest analyses. ROI analyses served three purposes: (1) assess and compare the goodness of fit of the neural data between the RELATIVE and the ABSOLUTE computational model parametric modulators (GLM1); (2) assess choice related brain activity in the vmPFC as a function of the task contexts (GLM2); (3) assess outcome encoding in the VS and the AI as a function of the task contexts (GLM3). All ROI analyses were designed to avoid double dipping in favour of the hypothesis we aimed to validate[45]. To assess goodness of fit (neural model selection), we first defined from GLM1a (ABSOLUTE's regressors) a 'task network' mask including all the voxels which survived cluster level $P<0.05$ (FWE corrected) in the following contrasts: positive and negative correlation with 'Q_C–Q_U' (decision value) and 'δ_C—δ_U' (decision prediction error) (see Fig. 5a). Within this mask (total voxels number = 1,511), we estimated GLM1a and GLM1b (best model regressors) using Bayesian statistics, which provided log evidence for each GLM. Log evidence was then fed to BMS random effects analysis, which computed the exceedance probability of each GLM within the mask[44]. This analysis indicates which GLM better explained the neural data. To avoid double dipping in favour of the hypothesis that we wanted to support, we selected the ROIs, which favoured the hypothesis we wanted to reject (GLM1a, ABSOLUTE model)[45]. The second ROI analysis was devoted to study how task factors (contexts) affected choice related activity. A spherical ROI of 4 mm diameter was centred on ventromedial prefrontal coordinates reported to be significantly associated with decision value in a recent meta-analysis[11]. Regression coefficients from the GLM2 were submitted to a repeated measure three-way ANOVA analysis with valence (reward and punishment), feedback information (partial and complete) and learning phase (early, late) as factors. The third ROI analysis was devoted to study how task factors (contexts) affected outcome encoding. Spherical ROIs of 4 mm were centered on striatal (VS) and insular (AI) coordinates reported to be significantly associated with reward and punishment prediction errors in a recent meta-analysis[8]. Regression coefficients were submitted to a repeated measure three-way ANOVA analysis with neural system (VS or AI) and valence (reward and punishment) and feedback information (partial and complete) as factors. In the second and third ROI analyses the *post hoc* significance assessed with two-sided one-sample *t*-test.

References

1. Dayan, P. Twenty-five lessons from computational neuromodulation. *Neuron* **76,** 240–256 (2012).
2. Daw, N. D. *Advanced Reinforcement Learning* (Academic Press, 2014).
3. 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).
4. Guitart-Masip, M. *et al.* Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage* **62,** 154–166 (2012).
5. Pessiglione, M. *et al.* Subliminal instrumental conditioning demonstrated in the human brain. *Neuron* **59,** 561–567 (2008).
6. Palminteri, S. *et al.* Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* **76,** 998–1009 (2012).
7. 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).
8. Garrison, J., Erdeniz, B. & Done, J. Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* **37,** 1297–1310 (2013).
9. Knutson, B., Katovich, K. & Suri, G. Inferring affect from fMRI data. *Trends Cogn. Sci.* **18,** 422–428 (2014).
10. Liu, X., Hairston, J., Schrier, M. & Fan, J. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* **35,** 1219–1236 (2011).

11. Clithero, J. a. & Rangel, A. Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9,** 1289–1302 (2013).
12. Pessiglione, M. & Lebreton, M. in *Handb Biobehav Approaches to Self-Regulation*. (eds Gendola, G., Mattie, T. & Koole, S.) 157–173 (Springer, 2015).
13. Louie, K. & Glimcher, P. W. Efficient coding and the neural representation of value. *Ann. NY Acad. Sci.* **1251,** 13–32 (2012).
14. Seymour, B. & McClure, S. M. Anchors, scales and the relative coding of value in the brain. *Curr. Opin. Neurobiol.* **18,** 173–178 (2008).
15. Rangel, A. & Clithero, J. a. Value normalization in decision making: theory and evidence. *Curr. Opin. Neurobiol.* **22,** 970–981 (2012).
16. Padoa-schioppa, C. & Rustichini, A. Rational attention and adaptive coding. *Am. Econ. Rev. Pap. Proc.* **104,** 507–513 (2014).
17. Grey, J. A. *The Psychology of Fear and Stress* Vol. 5 (Cambridge Univ. Press, Cambridge, UK, 1991).
18. Solomon, R. L. & Corbit, J. D. An opponent-process theory of motivation. I. Temporal dynamics of affect. *Psychol. Rev.* **81,** 119–145 (1974).
19. Mowrer, O. H. *Learning theory and behavior* (John Wiley & Sons Inc, 1960).
20. Kim, H., Shimojo, S. & O'Doherty, J. P. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biol.* **4,** e233 (2006).
21. Winston, J. S., Vlaev, I., Seymour, B., Chater, N. & Dolan, R. J. Relative Valuation of Pain in Human Orbitofrontal Cortex. *J. Neurosci.* **34,** 14526–14535 (2014).
22. Seymour, B. *et al.* Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat. Neurosci.* **8,** 1234–1240 (2005).
23. Nieuwenhuis, S. *et al.* Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage* **25,** 1302–1309 (2005).
24. Loomes, G. & Sugden, R. Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *Econ. J.* **92,** 805–824 (1982).
25. Vlaev, I., Chater, N., Stewart, N. & Brown, G. D. a. Does the brain calculate value? *Trends Cogn. Sci.* **15,** 546–554 (2011).
26. Coricelli, G. *et al.* Regret and its avoidance: a neuroimaging study of choice behavior. *Nat. Neurosci.* **8,** 1255–1262 (2005).
27. Palminteri, S., Boraud, T., Lafargue, G., Dubois, B. & Pessiglione, M. Brain hemispheres selectively track the expected value of contralateral options. *J. Neurosci.* **29,** 13465–13472 (2009).
28. Worbe, Y. *et al.* Reinforcement Learning and Gilles de la Tourette Syndrome. *Arch. Gen. Psychiatry* **68,** 1257–1266 (2011).
29. Frank, M. J., Seeberger, L. C., Reilly, R. C. O. & O'Reilly, R. C. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* **306,** 1940–1943 (2004).
30. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338,** 270–273 (2012).
31. Li, J. & Daw, N. D. Signals in human striatum are appropriate for policy update rather than value prediction. *J. Neurosci.* **31,** 5504–5511 (2011).
32. Boorman, E. D., Behrens, T. E. & Rushworth, M. F. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol.* **9,** e1001093 (2011).
33. Fischer, A. G. & Ullsperger, M. Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron* **79,** 1243–1255 (2013).
34. Sutton, R. S. R. S. & Barto, A. G. A. G. *Reinforcement Learning: An Introduction. IEEE Trans Neural Networks* 9 (MIT Press, 1998).
35. Niv, Y., Joel, D. & Dayan, P. A normative perspective on motivation. *Trends Cogn. Sci.* **10,** 375–381 (2006).
36. Guitart-Masip, M., Duzel, E., Dolan, R. & Dayan, P. Action versus valence in decision making. *Trends Cogn. Sci.* **18,** 194–202 (2014).
37. Moutoussis, M., Bentall, R. P., Williams, J. & Dayan, P. A temporal difference account of avoidance learning. *Network.* **19,** 137–160 (2008).
38. Maia, T. V. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn. Behav.* **38,** 50–67 (2010).
39. Pitt, M. a. & Myung, I. J. When a good fit can be bad. *Trends Cogn. Sci.* **6,** 421–425 (2002).
40. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS. Comput. Biol.* **10,** e1003441 (2014).
41. O'Doherty, J. P., Hampton, A. & Kim, H. Model-based fMRI and its application to reward learning and decision making. *Ann. NY Acad. Sci.* **1104,** 35–53 (2007).
42. Burke, C. J., Tobler, P. N., Baddeley, M. & Schultz, W. Neural mechanisms of observational learning. *Proc. Natl Acad. Sci. USA* **107,** 14431–14436 (2010).
43. Li, J., Delgado, M. R. & Phelps, E. a. How instructed knowledge modulates the neural systems of reward learning. *Proc. Natl Acad. Sci. USA* **108,** 55–60 (2011).
44. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46,** 1004–1017 (2009).
45. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12,** 535–540 (2009).
46. Kahnt, T. *et al.* Dorsal striatal-midbrain connectivity in humans predicts how reinforcements are used to guide decisions. *J. Cogn. Neurosci.* **21,** 1332–1345 (2009).
47. Samanez-Larkin, G. R., Hollon, N. G., Carstensen, L. L. & Knutson, B. Individual differences in insular sensitivity during loss: Anticipation predict avoidance learning: Research report. *Psychol. Sci.* **19,** 320–323 (2008).
48. Büchel, C., Morris, J., Dolan, R. J. & Friston, K. J. Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* **20,** 947–957 (1998).
49. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120,** 190–229 (2013).
50. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117,** 197–209 (2010).
51. Pompilio, L. & Kacelnik, A. Context-dependent utility overrides absolute memory as a determinant of choice. *Proc. Natl Acad. Sci. USA* **107,** 508–512 (2010).
52. Tversky, A. & Simonson, I. Context- dependent preferences. *Manage Sci.* **39,** 1179–1189 (2012).
53. Morris, R. W., Dezfouli, A., Griffiths, K. R. & Balleine, B. W. Action-value comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. *Nat. Commun.* **5,** 4390 (2014).
54. Lee, S. W. W., Shimojo, S., O'Doherty, J. P. & O'Doherty, J. P. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron* **81,** 687–699 (2014).
55. Skvortsova, V., Palminteri, S. & Pessiglione, M. Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. *J. Neurosci.* **34,** 15621–15630 (2014).
56. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47,** 129–141 (2005).
57. Daw, N. D., Kakade, S. & Dayan, P. Opponent interactions between serotonin and dopamine. *Neural Netw.* **15,** 603–616 (2002).
58. Grossberg, S. & Schmajuk, N. A. Neural dynamics of attentionally-modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology* **15,** 195–240 (1987).
59. Brooks, A. M. & Berns, G. S. Aversive stimuli and loss in the mesocorticolimbic dopamine system. *Trends Cogn. Sci.* **17,** 281–286 (2013).
60. Seymour, B., Singer, T. & Dolan, R. The neurobiology of punishment. *Nat. Rev. Neurosci.* **8,** 300–311 (2007).
61. Delgado, M. R., Li, J., Schiller, D. & Phelps, E. a. The role of the striatum in aversive learning and aversive prediction errors. *Philos. Trans R. Soc. Lond. B Biol. Sci.* **363,** 3787–3800 (2008).
62. Jessup, R. K. & O'Doherty, J. P. Distinguishing informational from value-related encoding of rewarding and punishing outcomes in the human brain. *Eur. J. Neurosci.*. n/a–n/a **39,** 2014–2026 (2014).
63. Desmond, J. E. & Glover, G. H. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *J. Neurosci. Methods.* **118,** 115–128 (2002).
64. Baird, L. C. Reinforcement learning in continuous time: advantage updating. in *Proc 1994 IEEE Int Conf Neural Networks* 4, 2448–2453 (IEEE, 1994).
65. Rescorla, R. A. & Wagner, A. R. in *Class Cond II Curr Res theory* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Applenton-Century-Crofts, 1972).
66. Daw, N. D. in *Decision Making, Affect, and Learning: Attention and Performance XXIII* **23,** 3–38 (2011).
67. Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F. & Procyk, E. Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. *Cereb. Cortex.* doi: 10.1093/cercor/bhu114 (2014).
68. Worbe, Y. *et al.* Valence-dependent influence of serotonin depletion on model-based choice strategy. *Mol. Psychiatry.* doi: 10.1038/mp.2015.46 (2015).
69. Corrado, G. S., Sugrue, L. P., Brown, J. R. & Newsome, W. T. in *Neuroeconomics Decis Mak Brain* (eds Glimcher, P. W., Fehr, E., Camerer, C. F. & Poldrack, R. a.) 463–480 (Academic Press, 2009).
70. Weiskopf, N., Hutton, C., Josephs, O., Turner, R. & Deichmann, R. Optimized EPI for fMRI studies of the orbitofrontal cortex: compensation of susceptibility-induced gradients in the readout direction. *MAGMA* **20,** 39–49 (2007).

Acknowledgements

We are thankful to Alexandre Salvador for helping during behavioural piloting and the staff of the Laboratorio di Neuroimmagine Funzionale (LNiF) at Mattarello (Italy), for helping during fMRI experiments. We are grateful to Etienne Koechlin, Maël Lebreton, Florent Meyniel and Liane Schmidt and the anonymous reviewers for insightful conversations or comments on early versions of this manuscript. We thank Muriel Ekovich, Emily Garret and Charles Finding for checking the English. SP was supported by the *Agence National de la Rercheche* (ANR- 11-EMCO-010) and by a Marie Sklodowska-Curie Individual European Fellowship (PIEF-GA-2012 Grant 328822). The study was funded by the European Research Council (ERC Consolidator Grant 617629) and the Agence National de la Rercheche (ANR-11-EMCO-010) grants to GC. MK was supported by the Agence National de la Recherche (Grants ANR-11-BSV4-006, ANR-12-CORD-0030, ANR-11-LABX-65, ANR-11-IDEX-0004-02), by Sorbonne Universités (Grant SU-15-R-PERSU-14) and by the EU Horizon 2020 Programme (Grant 640891). The Département d'Etudes Cognitives (Ecole Normale Supérieure) was supported by the Agence National de la Recherche grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*.

Author contributions

S.P. and G.C. designed the research. S.P. and M.J. acquired the data. S.P. analysed the data. M.K. provided tools for the computational analyses. S.P., M.K. and C.G. prepared the manuscript. All authors discussed the interpretation of the results.

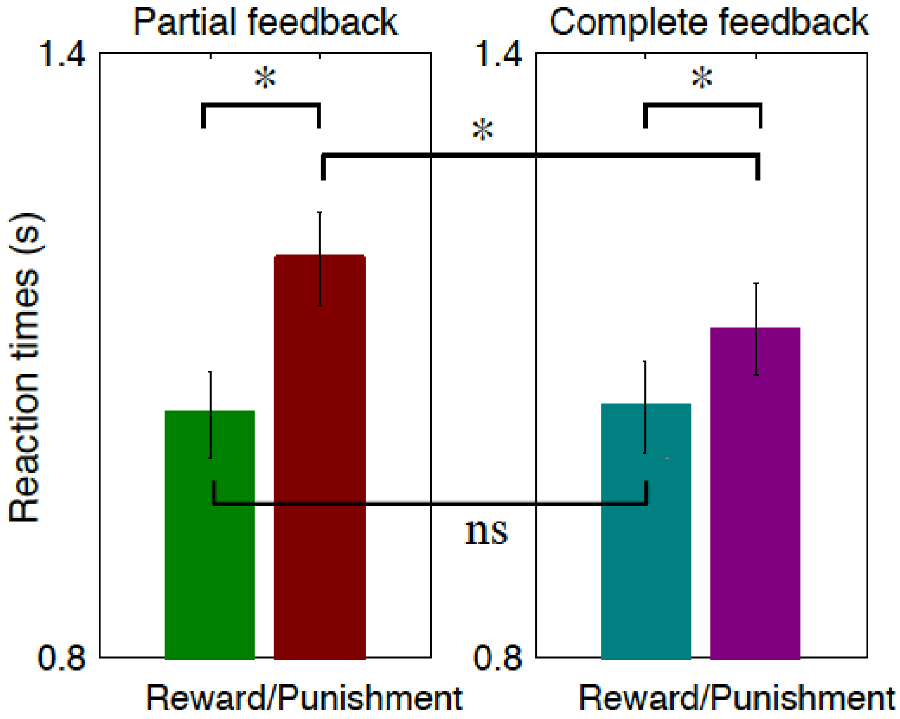
Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

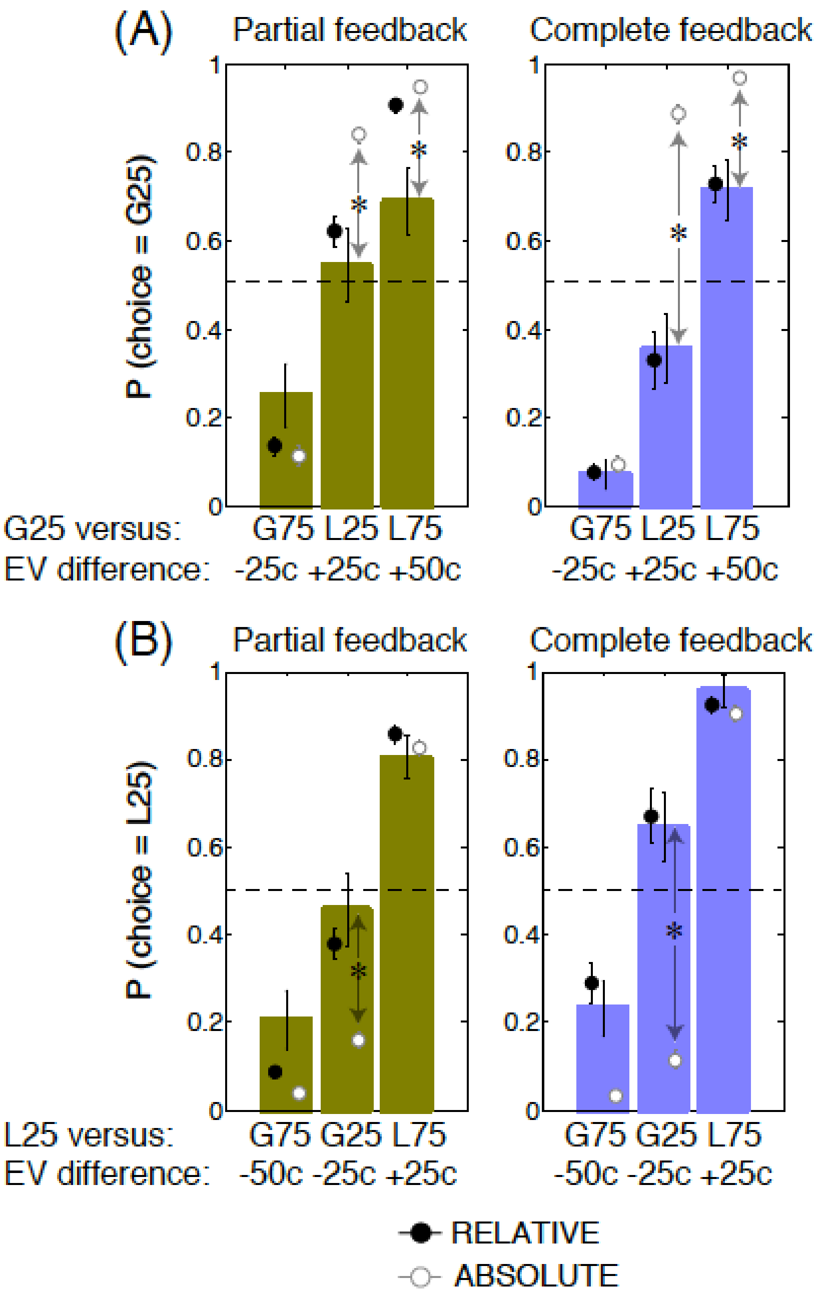
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Palminteri, S. *et al.* Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* 6:8096 doi: 10.1038/ncomms9096 (2012).



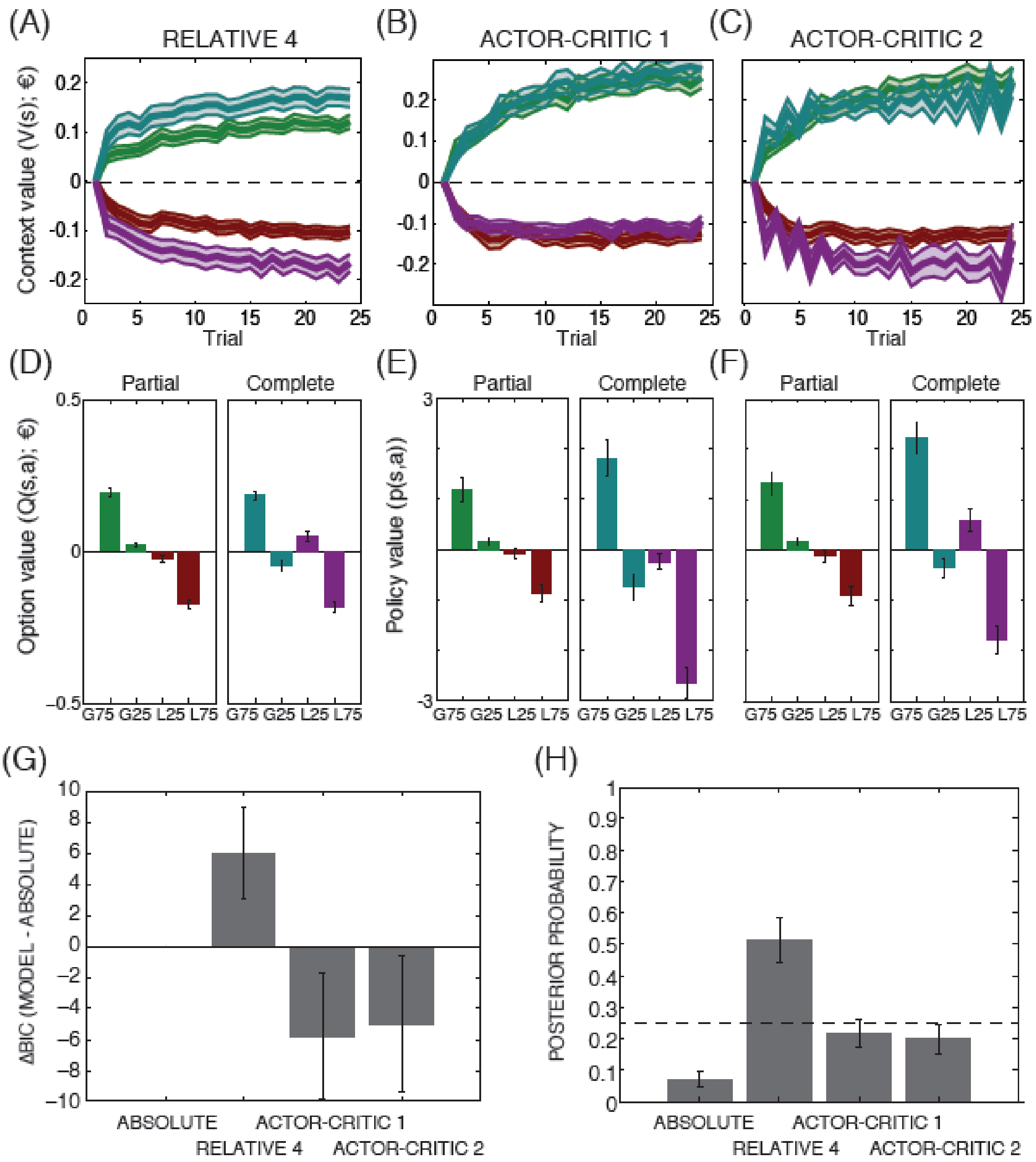
Supplementary Figure 1: reaction times.

Average reaction times during the learning test as a function of the choice contexts. *P<0.05 one sample t-test; ns: not significant. Error bars represent s.e.m.



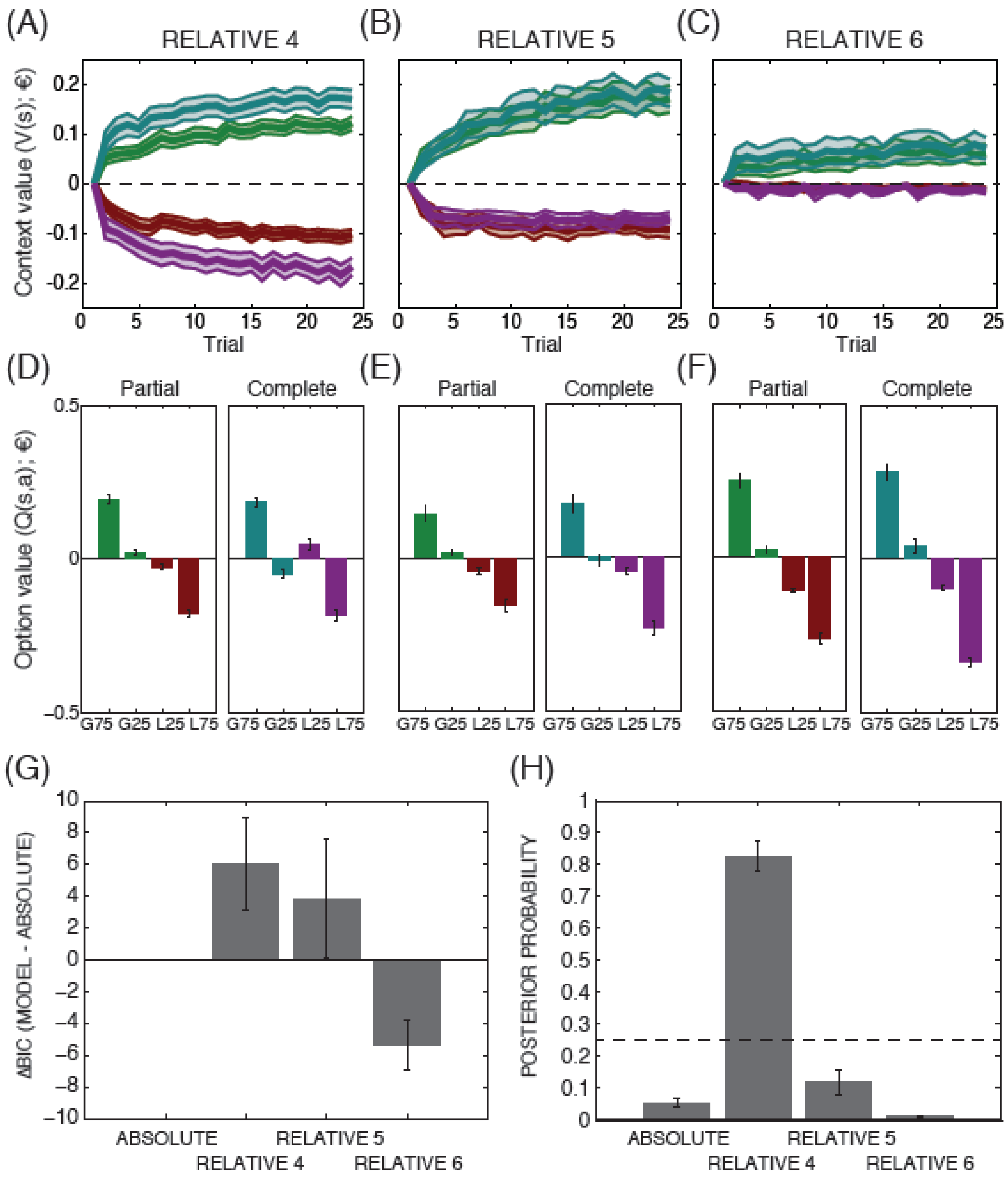
Supplementary Figure 2: intermediate value cues post learning choice rates

(A) and (B) Post-learning choice rate for comparisons involving the incorrect option in reward conditions (G_{25}: options associated with 25% percent of winning 0.5€) and the correct option in the punishment conditions (L_{25}: options associated with 25% percent of losing 0.5€), respectively. EV difference: difference in the absolute expected value (Probability(outcome) * Magnitude(outcome)) for a given cues comparison. Negative "EV difference" values indicate lower EV in the intermediate value cue (G_{25} or L_{25}) compared to the cue to which it is compared. Positive "EV difference" values indicate the opposite. Colored bars represent the actual data and black (RELATIVE) and white (ABSOLUTE) dots represent the model-simulated data. *P<0.05 one sample t-test corrected for multiple (twelve) comparisons. Error bars represent s.e.m.



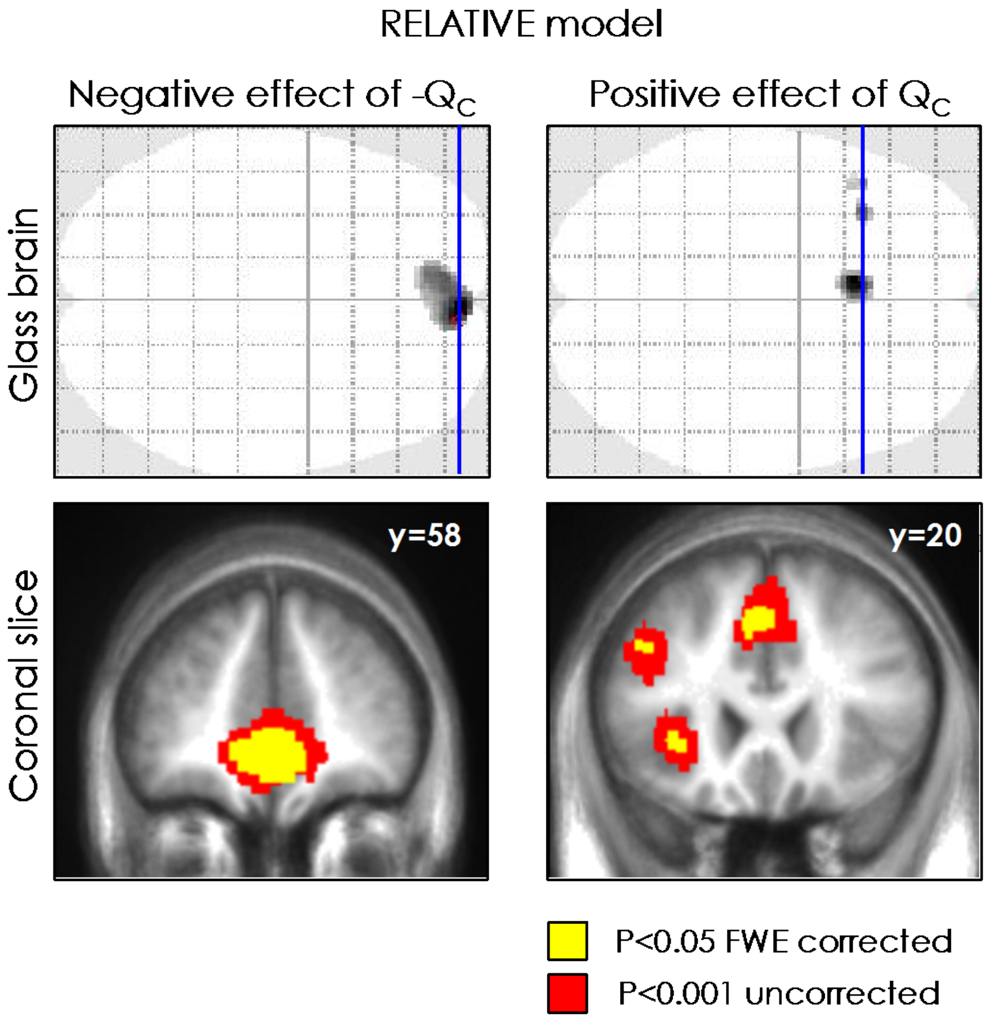
Supplementary Figure 3: comparison between the RELATIVE 4 and the actor-critic models

(A), (B) and (C): the graphs represent the model estimate of the context (state) values as a function of trial and the task context (the color scheme is the same used in the main text). Bold lines represent the mean; the shaded areas represent the s.e.m. (D), (E) and (F): the bars represent the final option or policy value estimates. . G$_{75}$ and G$_{25}$: options associated with 75% and 25% percent of winning 0.5€, respectively; L$_{75}$ and L$_{25}$: options associated with 75% and 25% percent of losing 0.5€, respectively. The estimates are generated from individual history of choices and outcomes and subject-specific free parameters. (G): the bars represent the difference in BIC between a model and the ABSOLUTE model (Q-learning). Positive values indicate better fit, negative values worst fit, compared to the ABSOLUTE model. (H): the bars represent the posterior probability of the model given the data and the parameters values (calculated based on the LPP; see supplementary Table 1). The dotted line represents chance level (0.25). Errors bars represent s.e.m.



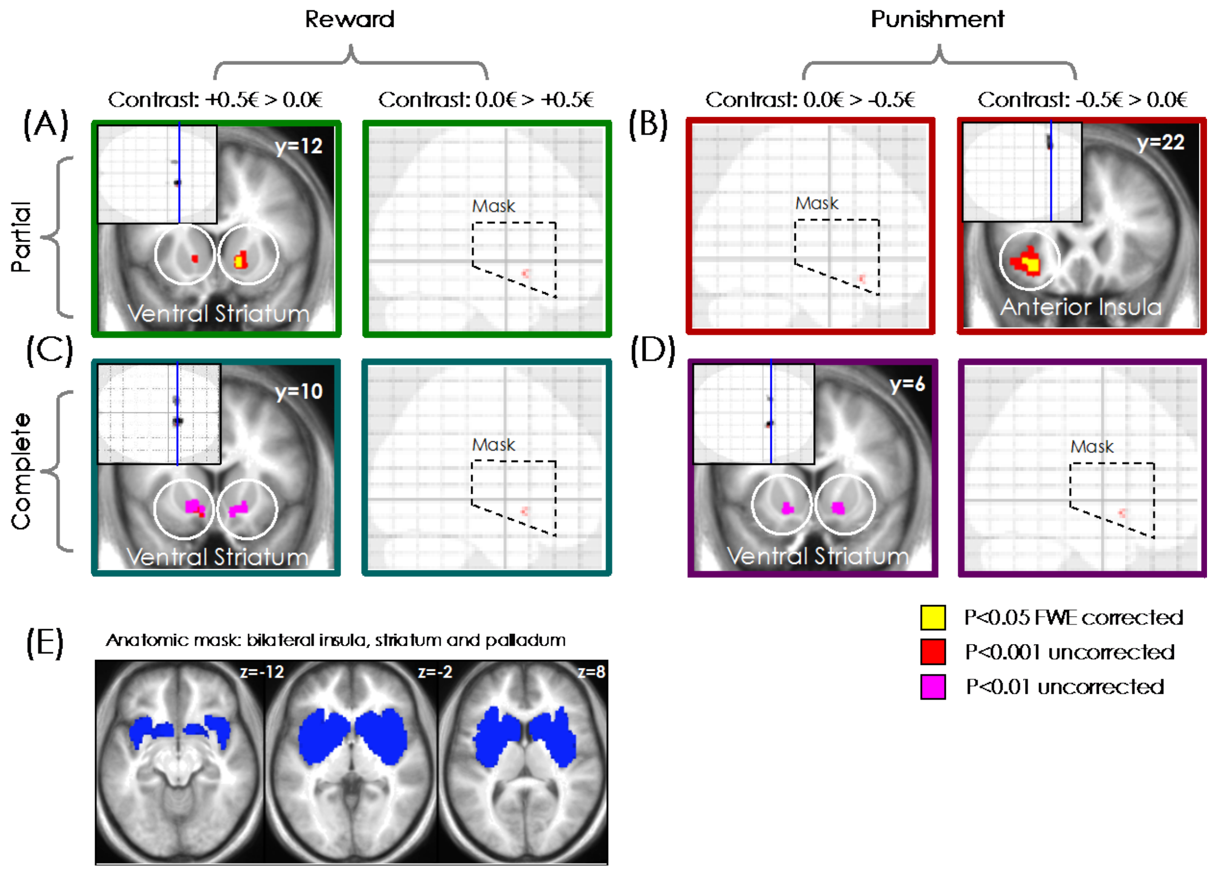
Supplementary Figure 4: comparison between the RELATIVE 4, 5 and 6 models

(A), (B) and (C): the graphs represent the model estimate of the context (state) values as a function of trial and the task context (the color scheme is the same used in the main text). Bold lines represent the mean; the shaded areas represent the s.e.m. (D), (E) and (F): the bars represent the final option value estimates. G_{75} and G_{25}: options associated with 75% and 25% percent of winning 0.5€, respectively; L_{75} and L_{25}: options associated with 75% and 25% percent of losing 0.5€, respectively. The estimates are generated from individual history of choices and outcomes and subject-specific free parameters. (G): the bars represent the difference in BIC between a model and the ABSOLUTE model (Q-learning). Positive values indicate better fit, negative values worst fit, compared to the ABSOLUTE model. (H): the bars represent the posterior probability of the model given the data and the parameters values (calculated based on the LPP; see supplementary Table 1). The dotted line represents chance level (0.25). Errors bars represent s.e.m.



Supplementary Figure 5: chosen option value representation in the RELATIVE model

Brain areas correlating positively and negatively with the chosen option value (Q_C; left and right column). Significant voxels are displayed on the glass brains (top) and superimposed to slices of the between-subjects averaged anatomical T1 (bottom). Coronal slices correspond to the blue lines on sagittal glass brains. Areas colored in gray-to-black gradient on glass brains and in yellow on slices showed a significant effect at $P<0.05$, voxel level FWE corrected). Areas colored in red on the slices showed a significant effect at $P<0.001$, uncorrected. Y coordinates are given in the MNI space. The results are from the GLM using the RELATIVE model parametric modulators (GLM1b).



Supplementary Figure 6: outcome encoding and anatomic mask

The figure presents the brain activations, concerning the outcome contrasts between the best and the worst outcomes (R$_{C}$; reward contexts contrast: +0.5€>0.0€; punishment contexts contrast: 0.0€>-0.5€), obtained from the categorical GLM3. Significant voxels are displayed on axial glass brains and superimposed to coronal slices of the between-subjects averaged anatomical T1. Coronal slices correspond to the blue lines on axial glass brains. Y coordinates are given in the MNI space. The results are from the categorical GLM2. Areas colored in gray-to-black gradient on glass brains and in red on slices showed a significant effect (P<0.001, uncorrected in A & B, and P<0.01, uncorrected in C & D). Areas colored in yellow on slices showed a significant effect (P<0.05, FWE mask-level corrected). (A) Significant activations by the best>worst outcome (+0.5€>0.0€) contrast in the reward/partial condition. (B) Significant activations by the best>worst outcome (0.0€>-0.5€) contrast in the punishment/partial condition. (C) Significant activations by the best>worst outcome (+0.5€>0.0€) contrast in the reward/complete condition. (D) Significant activations by the best>worst outcome (0.0€>-0.5€) contrast in the punishment/complete condition. (E) The blue voxels correspond to the anatomic mask used for the study of outcome related activations. The mask includes all voxels classified as striatum, pallidum and insula in the Automatic Anatomic Labeling (AAL) atlas. The mask is superimposed to axial slices of the between-subjects averaged anatomical T1.

Supplementary Tables

Supplementary Table 1: intermediate value cues post learning choice rates

The table summarizes for both intermediate value cues (G_{25} and L_{25}) and feedback information (partial and complete) their experimental and model-derived dependent post-learning choice rate. DATA: experimental data; RELATIVE 4: relative value learning model with delta rule update and context-specific heuristic (best fitting model in all model comparison analyses); ABSOLUTE: absolute value learning model (Q-learning). Data are expressed as mean ± s.e.m. *P<0.05 t-test, comparing the model-derived values to the actual data after correcting for multiple comparisons.

Comparison	DATA	ABSOLUTE	RELATIVE 4
G25 vs G75 (partial)	0.25±0.07	0.11±0.02	0.14±0.02
G25 vs L25 (partial)	0.54±0.08	0.84±0.02*	0.61±0.03
G25 vs L75 (partial)	0.69±0.07	0.95±0.01*	0.91±0.02
G25 vs G75 (complete)	0.07±0.03	0.09±0.02	0.08±0.02
G25 vs L25 (complete)	0.36±0.08	0.88±0.02*	0.33±0.06
G25 vs L75 (complete)	0.71±0.07	0.97±0.01*	0.73±0.04
L25 vs G75 (partial)	0.20±0.06	0.04±0.01	0.09±0.02
L25 vs G25 (partial)	0.45±0.08	0.16±0.02*	0.38±0.03
L25 vs L75 (partial)	0.80±0.05	0.83±0.02	0.86±0.02
L25 vs G75 (complete)	0.23±0.06	0.03±0.02*	0.29±0.05
L25 vs G25 (complete)	0.64±0.08	0.11±0.02	0.67±0.06
L25 vs L75 (complete)	0.95±0.04	0.90±0.02	0.92±0.02

Supplementary Table 2: model comparison of different algorithmic specifications of the RELATIVE model

The table summarizes for each model its fitting performances. DF: degrees of freedom; LLmax: maximal Log Likelihood; AIC: Akaike Information Criterion (computed with LLmax); BIC: Bayesian Information Criterion (computed with LLmax); LPP: Log of Posterior Probability; XP: exceedance probability (computed from LPP). PP: posterior probability of the model given the data. RELATIVE 4 is the model described in the main text.

Model	Update	Heuristic	DF	-2*LLmax	2*AIC	BIC	-2*LPP	PP	XP
ABSOLUTE	-	-	3	307±20	319±20	325±20	314±20	0.08±0.04	0.0
RELATIVE 1	Frequentist	Aspecific	3	306±22	318±22	324±22	315±21	0.00±0.01	0.0
RELATIVE 2	Frequentist	Specific	3	303±22	315±22	322±22	313±22	0.06±0.01	0.0
RELATIVE 3	Delta rule	Aspecific	4	298±22	315±22	323±21	307±21	0.02±0.03	0.0
RELATIVE 4	**Delta rule**	**Specific**	**4**	**295±22**	**311±22**	**319±22**	**304±21**	**0.84±0.05**	**1.0**

Supplementary Table 3: model comparison involving the actor-critic models

The table summarizes for each model its fitting performances. DF: degrees of freedom (number of free parameters). LLmax: maximal Log Likelihood; AIC: Akaike Information Criterion (computed with LLmax); BIC: Bayesian

Information Criterion (computed with LLmax); LPP: Log of Posterior Probability; XP: exceedance probability (computed from LPP). PP: posterior probability of the model given the data (computed from LPP).

Model	DF	2*LLmax	2*AIC	BIC	2*LPP	PP	XP
ABSOLUTE	3	307±20	319±20	324±20	314±20	0.07±0.02	0.00
RELATIVE 4	4	295±22	311±22	319±22	304±21	0.51±0.07	0.90
ACTOR-CRITIC1	4	307±22	323±22	331±22	310±22	0.22±0.04	0.09
ACTOR-CRITIC2	4	306±22	322±22	329±22	310±22	0.19±0.05	0.01

Supplementary Table 4: model comparison as a function of different ways to calculate the context value prediction error in the RELATIVE model (i.e. “random-policy”, “on-policy”, “best-policy”).

The table summarizes for each model its fitting performances. Partial R_V: calculation of the context-level outcome term used to update the context value V(s) in the partial feedback conditions. Complete R_V: calculation of context-level outcome term used to update the context value V(s) in the complete feedback conditions. LLmax: maximal Log Likelihood; AIC: Akaike Information Criterion (computed with LLmax); BIC: Bayesian Information Criterion (computed with LLmax); LPP: Log of Posterior Probability; XP: exceedance probability (computed from LPP). PP: posterior probability of the model given the data (computed from LPP).

Model	Partial R_V	Complete R_V	2*LLmax	2*AIC	BIC	2*LPP	PP	XP
ABSOLUTE	-	-	307±20	319±20	324±20	314±20	0.05±0.02	0.00
RELATIVE 4	$(R_C+Q(s,u))/2$	$(R_C+R_U)/2$	295±22	311±22	319±22	304±21	0.82±0.05	1.00
RELATIVE 5	R_C	R_C	297±22	313±22	321±22	308±21	0.11±0.04	0.00
RELATIVE 6	$max(R_C,Q(s,u))$	$max(R_C,R_U)$	306±20	322±20	330±20	315±20	0.01±0.01	0.00

Supplementary Note 1: behavior

I) Reaction times

Reaction time analysis provides evidence of relative value encoding. We also analyzed the reaction times, with the same statistical model used for correct choice rate, and we observed a significant effect of outcome valence (F=78.0, P<0.001), a marginally statistical effect of feedback information (F=3.2, P=0.09) and interaction between the two (F=3.8, P=0.06) (Supplementary Figure 1). Post-hoc test revealed that subjects were slower in the punishment avoidance contexts compared to the reward ones (partial and complete contexts: T>4.0, P<0.001), whereas the effect of feedback information reached statistical significance only in the punishment context (T=2.5, P<0.05), but not in the reward one (T=0.3, P>0.5). Conditioning (Pavlovian-to-instrumental transfer or PIT) as well as decision field theories established a link between chosen option value and reaction times. More precisely they predict that the subjects would take more time if choices are likely to result in negative outcomes[1–3]. We indeed observed this effect (main effect of valence), since subjects were slower when choices potentially led to negative outcomes. However, the trend toward a significant valence x information interaction (driven by faster responses in the punishment/complete context compared to the punishment/partial context) suggests that the reaction times' pattern could not be fully explained by considering option value on an absolute scale. . Importantly, the absence of difference in reaction times in the two reward contexts further indicates that observed pattern could not be explained assuming reaction times a simple function of to the correct response rate that is much higher in the reward/complete contexts compared to the reward/partial. Thus, learning and post-learning results suggest that the observed interaction may derive from relative value encoding: punishment-induced reaction times slowing in the punishment/complete context is smaller compared to the punishment/partial context, as if the option value was less negative as a result of the value contextualization process.

II) Post-learning test detailed analysis

Value inversion in the post-learning test is robust across all possible binary comparisons and confirms relative value encoding. In the main text we reported post learning choice rate in an aggregate manner, i.e. reporting the probability of choosing an option, taking into account all possible comparisons. The advantage of using this aggregate measure resides in that it is directly proportional to the underlying option value, to which it can be therefore easily compared (see Figure 2B and Figure 3C and 3D). Here we report the results of all possible comparisons involving the intermediate value options (i.e. G_{25}, the incorrect option in the reward contexts and L_{25}, the correct option in the punishment contexts) (Supplementary Figure 2). The reason to focus on these options is that the ABSOLUTE and RELATIVE models crucially diverge with respect to their post-learning choice rate prediction about G_{25} and L_{25}. We analyzed the post learning choice with a three-way ANOVA analysis including option (two levels: G_{25} or L_{25}), feedback information (two levels: partial versus complete) and absolute expected value (EV; Probability(outcome) * Magnitude(outcome)) difference between the two options (three levels: low, mid and high) as factors. Crucially, the ABSOLUTE model predicts a main effect of cue (F=716.3, P<0.001), reflecting higher choice rate for the G_{25}, compared to the L_{25} option. The ABSOLUTE model also predicts no significant option x information interaction (F=0.4, P>0.5), indicating that increased choice rate for the G_{25} compared to the L_{25} was similar in both feedback information conditions, and significant option x EV difference interaction (F=217.2, P<0.001), reflecting a non-linear increase of post-learning choice rate as a function of EV difference. Importantly, the RELATIVE model predicts a completely different pattern, with no main effect of option (i.e. similar choice rate for the G_{25} and the L_{75} options; F=1.9, P>0.1), a significant option x information interaction (i.e. an option-specific effect of feedback information on post-learning choice rate, with higher choice rate for the L_{25} in the complete feedback information; F=51.7, P<0.001), and no significant option x EV difference interaction (i.e. a linear increase of post-learning choice rate as a function of EV difference; F=0.2, P>0.8). Actual post-learning choices systematically fulfilled the predictions of the RELATIVE model, by displaying no significant effect of option (F=3.0, P>0.09), a significant option x information interaction (F=5.1, P<0.05) and no significant option x EV difference (F=0.0, P>0.9). Accordingly, direct systematic comparisons between the actual and the model predicted data, confirmed that only the ABSOLUTE model suffers from significantly diverging from subjects' post-learning behavior (see supplementary Figure 2 and supplementary Table 1).

III) Post-experiment debriefing

Post-scanning structured interview fails to reveal acquired explicit knowledge of task factors. A post-scanning structured interview was administrated to a subgroup of subjects (17/28; 60.7%). The interview was aimed to assess subjects' explicit knowledge of the learning task's features and contingencies. More precisely the structured interview assessed: i) whether or not the subjects were aware about the cues being presented in fixed pairs (choice contexts); ii) how many choice contexts they believed were simultaneously present in a learning session; iii) if they believed or not that rewards and punishments were being separated across choice contexts; iv) if they believed or not that partial and complete feedbacks were being separated across choice contexts. Subjects, on average, correctly retrieved that during learning the cues were presented in fixed choice contexts during learning (correct responses: 88.2%; P<0.001). When asked about how many pairs of cues were presented in a session, subjects, on average, answered 4.6±0.2%, slightly overestimating the correct number (i.e. 4; T=2.2, P<0.05). The task's factors (outcome valence and feedback information) were not significantly reported as discrete, mutually exclusive, features of the choice contexts. Indeed, subjects did not correctly report rewards and punishments as choice context-specific (correct responses: 35.3%; P>0.05). Similarly, subjects did not correctly report partial and complete feedbacks as choice context-specific (correct responses: 47.1%; P>0.2). Thus, as far as explicit knowledge of the task structure can be inferred by the post-scanning structured interview, whereas the existence of discrete choice contexts (states) and their number seemed explicitly grasped by the subjects, the separation between reward and punishment, as well as between partial and complete feedback, conditions, remained implicit. These two features are taken into account by our computational models that i) assume the perception of discrete states (s) but ii) treat option and context values as continuous ("model-free") variables instead of categorical ("rule or model-based") ones.

Supplementary Note 2: computational modeling

Model comparison- based justification of the algorithmic specification of the RELATIVE model reported in the main text. Four different variants of the RELATIVE model were considered, in order to select amongst different possible algorithmic implementations, such as different ways to update the state value (frequentist versus delta rule) and the heuristic employed to obviate the absence of counterfactual outcome (R_U) in the partial feedback contexts, when calculating the context-level outcome R_V. The first computational question is the learning rule used for context value update. In fact, whereas there is now strong and cumulative evidence that option values are learnt via delta rules[4] it could be that context value updates follow different learning rules. We included RELATIVE models 1 and 2 implementing frequentist inference:

$V_{t+1}(s) = ((t\text{-}1)/t)*V_t(s) + (1/t)*R_{V,t}$,

where t is the number of trials and R_V is the context-level outcome at trial t: a global measure that encompasses both the chosen and unchosen options. Frequentist inference is appropriate for environments with no volatility and instantiates a progressive reduction of the learning rate, since new experiences have less weight as the number of trials increases. RELATIVE models 1 and 2 with frequentist update of context value could be advantaged by the fact that they do not require additional free parameters, compared to the ABSOLUTE model. However, for the same reason, they cannot account for interindividual variability. We also included RELATIVE models 3 and 4 implementing the delta rule, which, for analogy with the frequentist update, can be written as:

$V_{t+1}(s) = (1\text{-}\,\alpha_3)*V_t(s) + \alpha_3*R_{V,t}$,

Where α_3 is the context value learning rate. Delta rule is appropriate for environments with unknown volatility. RELATIVE models 3 and 4 with delta rule update of context value could be disadvantaged by the fact that they require an additional free parameter (α_3), compared to the ABSOLUTE model. However, for the same reason, they can account for interindividual variability. The second computational question concerned the definition of R_V. In fact, whereas average outcome trial can be straightforwardly calculated in the complete feedback contexts as the average of the factual and the counterfactual outcomes as follows:

$R_{V,t} = (R_{C,t} + R_{U,t}) / 2$,

the question arises in the partial feedback contexts, where R_U is not explicitly provided. One possibility, (implemented in RELATIVE models 1 and 3) is to replace R_U with R_M (the central – median - task reward: 0.0€), in the partial feedback contexts,:

$R_{V,t} = (R_{C,t} + R_{M,t}) / 2$,

which we define as a "context-aspecific heuristic", in which, simplifying, $R_V = R_{C,t} / 2$, . However, given that R_V is meant to be a context-level measure, in order to incorporate unchosen option information in R_V also in the partial feedback contexts, a possibility, implemented in RELATIVE models 2 and 4, is to consider $Q_t(s,u)$ a proxy of $R_{U,t}$ and calculate $R_{V,t}$ as follows:

$R_{V,t} = (R_{C,t} + Q_t(s,u)) / 2$,

which we define as a "context-specific heuristic".

To sum up, this model space included 5 models. The ABSOLUTE model (Q-learning) and four RELATIVE models which differed in 1) context value update rule ("frequentist" versus "delta rule") and 2) the way R_V was calculated in the partial feedback contexts ("context-aspecific" or "context-specific" heuristic). We submitted these new models to the same parameters optimization procedure and model comparison analyses presented in the main text and involving the Bayesian information criterion (BIC), Akaike information criterio (AIC) and the Laplace approximation of the model evidence-based calculation of the model posterior probability and exceedance probability[5,6]. Complexity-penalizing model comparison criteria concordantly indicated that the RELATIVE model 4 better accounted for the data (see Supplementary Table 3). Note that priors-independent model comparison criteria (LLmax, AIC and BIC) were smaller (indicating better fit) in all RELATIVE models compared to the ABSOLUTE model, indicating that the finding that relative value learning better accounts for the data was robust across algorithmic variations of the context value update rule. Thus, subsequent analyses in the main text and in supplementary materials have been focused on the comparison RELATIVE model 4 only, to whom we referred simply as "RELATIVE", to stress the main feature of the model instead of its less relevant algorithmic specifications.

Position of the RELATIVE models within the family of reinforcement learning algorithms: similarity and differences with previous formulations. The RELATIVE family of models in general, and the RELATIVE 4 model in particular (the best fitting model), computationally embody the ideas behind the two-factor theory that, in simple terms, states that the instrumental action-induced punishment avoidance (cessation of fear in the original formulation) should acquire a positive reinforcement value, in order to sustain instrumental responding, in absence of further negative reinforcement (i.e. successful avoidance)[7]. The RELATIVE models capture this basic intuition of the two-factor theory assuming that, in the punishment conditions, neutral outcomes are computed relative to the negative context values (or state values as they are more frequently called in the reinforcement learning literature). The idea of computationally capturing elements of the two-factor theory by assuming some form of relative value learning has been also proposed in previous computational studies[8,9]. These studies were based on actor-critic or advantage learning models[10,11], and the models proved useful to account for classical avoidance learning results, such as the conditioned avoidance response (CAR) induced via discriminated avoidance procedure. The computational model tested here is inspired by these formulations, with whom it shares the notion of a separate track of action values (i.e. option values Q(s,a)) or policy value (i.e. 'policy' P(s,a)) and state values (V(s)), as well as the calculation of policy values relative to the state value. As a matter of fact the algorithmic implementation of our model only marginally differs from those of these previous models. Thus, in order to justify the introduction of the new model, we run supplementary model comparison analyses. In a first model comparison analysis we compared the RELATIVE 4 model with the actor-critic model, since the latter been explicitly proposed as an effective solution for punishment avoidance learning. Another algorithmic specificity of the RELATIVE 4 model is that V(s) is calculated in an random-policy manner (i.e. it depends on R_C and R_U in the complete feedback contexts and on R_C and Q(s,u) in the partial feedback contexts), as opposite to previous model in which it is calculated on-policy. Thus in the second model comparison analyses presented below, we addressed these issues by (I) comparing the RELATIVE 4 models with two variants of the actor-critic model, and (II) with two variants of the RELATIVE 4 model calculating V(s) based on the current or best policy instead of doing this in an random-policy manner.

I) Comparison with two variants of the actor-critic model

We compared the RELATIVE 4 model with two variants of the actor-critic model. At each trial t the model calculated a chosen policy prediction error defined as:

$\delta_{C,t} = R_{C,t} - V_t(s),$

where V(s) is the value of the current choice context s and R_C is the outcome of the chosen policy (factual outcome). This prediction error is then used to update the chosen policy value (P(s,c)) using a delta-rule:

$P_{t+1}(s,c) = P_t(s,c) + \alpha_1 \delta_{C,t},$

where α1 is the learning rate for the chosen option. We extended the actor-critic model in order to integrate counterfactual learning, as we have done for the other models. Thus, in the complete feedback contexts, the model also calculates an unchosen policy prediction error:

$\delta_{U,t} = R_{U,t} - V_t(s),$

where R_U is the outcome of the unchosen policy (counterfactual outcome). This prediction error is then used to update the unchosen policy value (P(s,u)) using a delta-rule:

$P_{t+1}(s,u) = P_t(s,u) + \alpha_2 \delta_{U,t},$

where α2 is the learning rate for the unchosen option. The two variants of the actor critic model differ in the way the context value V(s) is then updated. In the first, more "classical", variant (ACTOR-CRITIC 1) the chosen policy prediction error is also used to update the context value in all choice contexts:

$V_{t+1}(s) = V_t(s) + \alpha_3 \delta_{C,t},$

where α3 is the learning rate for the context value. In a second variant (ACTOR-CRITIC 2) the context value update also takes into account the unchosen policy prediction error:

$V_{t+1}(s) = V_t(s) + \alpha_3 \delta_{C,t} + \alpha_3 \delta_{U,t}.$

We submitted these new models to the same parameters optimization procedure and model comparison analyses presented in the main text and involving the Bayesian information criterion (BIC), Akaike information criterio (AIC) and the Laplace approximation of the model evidence-based calculation of the model posterior probability and exceedance probability[5,6]. The model space included the ABSOLUTE model as a reference point and the RELATIVE 4 (the best fitting model of the main model comparison). Including the choice temperature, the ACTOR-CRITIC models 1 and 2 have four free parameters, as the RELATIVE 4 model has. The results (see Supplementary Table 3 and Supplementary Figure 3) indicated that the RELATIVE 4 model provides a better account of the data, compared to both the actor-critic models.

II) Comparison with different ways to calculate the context value calculation

We also devised two additional variants of the RELATIVE models. These variants assume the context value being calculated based on the current (RELATIVE 5) or the best (RELATIVE 6) policy. More specifically, these models essentially differ from the RELATIVE 4 in the way they calculate the $R_{V,t}$: the context-level outcome at trial t, which is used to update the context, value V(s). In the RELATIVE 4 model R_V was calculated based on the R_C and Q(s,u), in the partial feedback contexts, and based on R_C and R_U, in the complete feedback contexts (i.e. "random-policy" since independent from the subjects' choice). This choice was motivated by conceiving V(s) as a reference point as much neutral as possible in respect to the current obtained outcomes, supposing that the subjects do take all feedback into account (thus being random-policy) to estimate the context value (see "Conclusions on supplementary computational analyses"). However this choice is not frequent in the current panorama of reinforcement learning algorithms. In the RELATIVE 5 model for all choice contexts the context level outcome is defined as:

$R_{V,t} = R_{C,t}$.

The context value V(s) is therefore calculated considering the ongoing policy ("on-policy"). This is the most frequent way to calculate the context value in the reinforcement learning literature. Note that the RELATIVE 5 is analogous to the advantage learning algorithm extended to also, once included the counterfactual learning module[10]. Another tempting possibility, particularly relevant in presence of complete feedback information, is to calculate the context value based on the best policy. The RELATIVE 6 model implements this possibility, in fact in the partial information choice contexts the context level outcome is defined at:

$R_{V,t} = max(R_C,Q(s,u))$,

whereas in the complete information choice contexts it is defined as:

$R_{V,t} = max(R_C,R_U)$.

We submitted these new models to the same parameters optimization procedure and model comparison analyses presented in the main text. The model space included the ABSOLUTE model, as a reference point, and the RELATIVE 4, 5 and 6 models. The RELATIVE 5 and 6 models have four free parameters, as the RELATIVE 4 model has. The results (see Supplementary Table 4 and Supplementary Figure 4) indicated that the RELATIVE 4 model provides a better account of the data, compared to RELATIVE models 5 and 6, thus supporting the random-policy calculation of the context value in this task. Another interesting metric to evaluate in each model the gain of implementing relative value learning is to look at the values of the context learning rate α_3. In fact, when $\alpha_3 = 0$ the RELATIVE models reduce to the ABSOLUTE model. In the RELATIVE model 4 only 4 subjects (14%) were fitted with $\alpha_3=0$. This percentage slightly increased in the RELATIVE model 5 (N=7; 25%) and dramatically increased in the RELATIVE model 6 (N=17; 61%), further confirming the relatively poor fitting performances of their context value update scheme (this analysis was performed on the parameters retrieved with likelihood maximization).

III) Conclusions on supplementary computational analyses

Whereas previous computational studies suggested that the actor-critic architecture could provide a good explanation for conditioned avoidance response[8,9], we found that in our task the RELATIVE 4 outperformed the actor-critic models. One important difference compared to the actor critic model is that the RELATIVE 4 model can be reduced to Q-learning assuming the contextual learning rate ($\alpha_3=0$), whereas the actor critic cannot. This lack of flexibility may at least partly explain the overall poor group-level performances. We also note the important differences between the discriminate avoidance procedure and our paradigm. In the former the contingencies are deterministic, avoidance learning is studied in isolation and the "avoidance learning paradox" consists in the long lasting insensitivity to extinction of the conditioned responses, despite the absence of further reinforcement. In our paradigm, the contingencies are probabilistic (thus with overlapping outcomes from the correct and incorrect choices), avoidance learning is not studied in isolation, but in opposition to reward seeking behavior and the "avoidance learning paradox" consists in similar performance in the reward punishment domain, despite the fact that the performance-induced sampling bias would predict enhanced performances in the reward domain. These important differences should be also taken into account, when interpreting the relatively poor performances of the actor-critic models in our task. On the other side the good potential of the actor-critic model to explain the post-test results (Supplementary Figure 3E and 3F), further illustrates the conceptual proximity between this influential algorithm and the RELATIVE model 4.

We also found that "random-policy" calculation of the context value in the RELATIVE model 4 provided a better explanation of instrumental choices compared to other forms of context value calculations (on-policy or best-policy) Interestingly, the RELATIVE models 5 and 6 also failed to capture the value inversion of the intermediate value cues in the complete feedback conditions (see Supplementary Figure 4E-F). As a matter of fact, in most of the classical instrumental conditioning (and machine learning) paradigms the agents are presented to only one type of choice context (either reward or punishment, as in the discriminated avoidance task, and there is almost no example of complete feedback information[1])[11,12]. Thus, in presence of only one type of choice context, the model predictions obtained using on-policy or random-policy context value (V(s)), can hardly diverge. In such mono-dimensional tasks, on- and random-policy context values would display a similar trend across trials and the eventual differences in their magnitude can easily be neutralized by rescaling parameters, such as coefficients or learning rates (see Supplementary Figure 4A-C). We believe that we were precisely able to rule out on-policy (and best-policy) context value, thanks to the presence of multiple, different choice contexts in our design. In particular, both the simultaneous contrasts between reward and punishment and between partial and complete feedback information contributed to highlight this feature of the best fitting model. In fact, only the random-policy context values i) were symmetrical in respect to the valence, thus permitting similar performances in the reward and punishment domains, and ii) were enhanced in magnitude in the complete feedback contexts, thus permitting the value inversion of the intermediate value cues in the complete feedback conditions (see Supplementary Figure 3A-F). Furthermore, the importance of being on-policy has been mainly stressed in problems with a risk of substantial/lethal punishments such as the cliff simulation where random-policy algorithms such as Q-learning cannot avoid sometimes falling in the cliff due to occasional exploratory decisions[11,12]. In our case, it is reasonable to consider that human subjects do not fear being harmed when interacting with the screen. In fact, we believe that this algorithmic difference between the standard view of the context value V(s) (on-policy) and ours (random-policy) betrays a more profound difference concerning the psychological intuitions behind these quantities. Whereas in most reinforcement learning models V(s) is conceived as a "Pavlovian" anticipation of the reward (or punishment) to come, aimed to elicit automatic motor effects[2,3], in our model it represents a more abstract signal, subserving value contextualization for efficient encoding purposes[13–15]. In the light of these interpretation it is easy to understand why in the framework of a "motor preparation", the context value needs to be calculated in an on-policy manner (preparation to an outcome), whereas in the framework of a "efficient coding", the context value has to be calculated in a random-policy manner. In principle both quantities (on- and off- policy context values) could exist in the brain and express their effects in different behavioral measures. Further work, probably implicating a deeper analysis of reaction times (a good candidate for Pavlovian effects) could shed light on this topic. Finally, we are not without acknowledging that an random-policy calculation of context value could rapidly become computationally challenging in learning situations implicating more than two options. Further studies are needed to uncover the learning heuristics implemented in such cases.

[1] This is less true in behavioral economics literature, where counterfactual (or "fictive") learning takes a more important place.

Supplementary Note 3: written task instructions

The subject read the learning test instructions before the training session, outside the scanner. The experimenter read the post-learning instructions to the subject, while he/she was in the scanner, after the last (fourth) functional acquisition, before starting the T1 anatomic acquisition.

Learning test instructions

The experiment is divided in four sessions, of about 12 minutes each. There will be two training sessions (a longer one outside and a shorter one inside the scanner) before the starting of the fMRI experiment.

You are asked to choose in each round one of two abstract symbols. The symbols will appear on the screen to the left or the right of a fixation cross. To choose one of the two symbols you should press the right or left button. After few seconds a cursor will appear under the chosen symbol confirming your choice. If you do not press any button, the cursor will appear at the center of the screen, and your result will be disadvantageous.

As an outcome of your choice you may:

-gain 50 cents (+0.5€)
-get nothing (0€)
-loose 50 cents (-0.5€)

The outcome of your choice will appear on the top of the chosen symbol, and will be always indicated by the position of the cursor. The two symbols are not equivalent (identical). One of the two symbols is on average more advantageous or less disadvantageous, in the sense that it makes you winning more often or loosing less often than the other. The goal of the experiment is to gain as much as you can.

In some trials the information about the outcome of the unchosen option will be also provided. Note that your earnings will correspond only to the chosen option. At the end of each session the experimenter will communicate your earnings for that session. Your final earnings will correspond to the sum of the earnings of the four sessions.

Post-learning test instructions

The test will last 5 minutes with no training.

The goal of the next test is to indicate the symbol with the higher value from the last (fourth) session. At any trial, you are asked to choose between two symbols pressing the corresponding button. Your choice will be immediately recorded and will be confirmed with the presence of a cursor that will appear under the chosen stimulus.

It will not always be the case that the shown symbols would have been presented together in the previous session. Please try to give an answer even if you are not completely sure.

Supplementary references

1. Busemeyer, J. R. & Townsend, J. T. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Maling in an Uncertain Environment. *Psychol Rev* **100,** 432–459 (1993).

2. Niv, Y., Joel, D. & Dayan, P. A normative perspective on motivation. *Trends Cogn Sci* **10,** 375–81 (2006).

3. Guitart-Masip, M., Duzel, E., Dolan, R. & Dayan, P. Action versus valence in decision making. *Trends Cogn Sci* **18,** 194–202 (2014).

4. Skvortsova, V., Palminteri, S. & Pessiglione, M. Learning To Minimize Efforts versus Maximizing Rewards: Computational Principles and Neural Correlates. *J Neurosci* 1–11 doi:10.1523/JNEUROSCI.1350-14.2014

5. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput Biol* **10,** e1003441 (2014).

6. Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F. & Procyk, E. Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. *Cereb Cortex* bhu114– (2014). doi:10.1093/cercor/bhu114

7. Mowrer, O. H. *Learning theory and behavior.* (John Wiley & Sons Inc, 1960). doi:10.1037/10802-000

8. Moutoussis, M., Bentall, R. P., Williams, J. & Dayan, P. A temporal difference account of avoidance learning. *Network* **19,** 137–60 (2008).

9. Maia, T. V. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav* **38,** 50–67 (2010).

10. Baird, L. C. Reinforcement learning in continuous time: advantage updating. in *Proc 1994 IEEE Int Conf Neural Networks* **4,** 2448–2453 (IEEE, 1994).

11. Sutton, R. S. R. S. & Barto, A. G. A. G. *Reinforcement Learning: An Introduction. IEEE Trans Neural Networks* **9,** (MIT Press, 1998).

12. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems (Computational Neuroscience).* (MIT Press, 2005). at <http://www.gatsby.ucl.ac.uk/~dayan/book/>

13. Louie, K. & Glimcher, P. W. Efficient coding and the neural representation of value. *Ann N Y Acad Sci* **1251,** 13–32 (2012).

14. Padoa-schioppa, C. & Rustichini, A. Rational Attention and Adaptive Coding : *Am Econ Rev Pap Proc* **104,** 507–513 (2014).

15. Rangel, A. & Clithero, J. a. Value normalization in decision making: theory and evidence. *Curr Opin Neurobiol* **22,** 970–81 (2012).