nature human behaviour

ARTICLES

PUBLISHED: 20 MARCH 2017 | VOLUME: 1 | ARTICLE NUMBER: 0067

Behavioural and neural characterization of optimistic reinforcement learning

Germain Lefebvre[1,2], Maël Lebreton[3,4], Florent Meyniel[5], Sacha Bourgeois-Gironde[2,6] and Stefano Palminteri[1,7]*

When forming and updating beliefs about future life outcomes, people tend to consider good news and to disregard bad news. This tendency is assumed to support the optimism bias. Whether this learning bias is specific to 'high-level' abstract belief update or a particular expression of a more general 'low-level' reinforcement learning process is unknown. Here we report evidence in favour of the second hypothesis. In a simple instrumental learning task, participants incorporated better-than-expected outcomes at a higher rate than worse-than-expected ones. In addition, functional imaging indicated that inter-individual difference in the expression of optimistic update corresponds to enhanced prediction error signalling in the reward circuitry. Our results constitute a step towards the understanding of the genesis of optimism bias at the neurocomputational level.

Francis Bacon wrote[1]: "It is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives; whereas it ought properly to hold itself indifferently disposed towards both alike."

People typically overestimate the likelihood of positive events and underestimate the likelihood of negative events. This cognitive trait in (healthy) humans is known as the optimism bias and has been repeatedly evidenced in many different guises and populations[2–4] such as students projecting their salary after graduation[5], women estimating their risk of getting breast cancer[6] or heavy smokers assessing their risk of premature mortality[7]. One mechanism hypothesized to underlie this phenomenon is an asymmetry in belief updating, colloquially referred to as the 'good news/bad news effect'[8,9]. Preferentially revising one's beliefs when provided with favourable compared with unfavourable information constitutes a learning bias that could, in principle, generate and sustain an overestimation of the likelihood of desired events and a concomitant underestimation of the likelihood of undesired events (optimism bias)[10].

This good news/bad news effect has recently been demonstrated in the case where outcomes are hypothetical future prospects associated with a strong a priori desirability or undesirability (estimation of post-graduation salary or the probability of getting cancer)[5,6]. In this experimental context, belief formation triggers complex interactions between episodic, affective and executive cognitive functions[8,9,11], and belief updating relies on a learning process involving abstract probabilistic information[8,12–14]. However, it remains unclear whether this learning asymmetry also applies to immediate reinforcement events driving instrumental learning directed to affectively neutral options (with no a priori desirability or undesirability). If an asymmetric update were also found in a task involving neutral items and direct feedback, then the good news/bad news effect could be considered as a specific — cognitive — manifestation of a general reinforcement learning asymmetry. If the asymmetry were not found at the basic reinforcement learning level, this would mean that the asymmetry is specific to abstract belief updating, and this would require a theory explaining this discrepancy.

To arbitrate between these two alternative hypotheses, we fitted the instrumental behaviour of subjects performing a simple two-armed bandit task, involving neutral stimuli and actual and immediate monetary outcomes, with two learning models. The first model (a standard reinforcement learning (RL) algorithm) confounded individual learning rates for positive and negative feedback, and the second one differentiated them, potentially accounting for learning asymmetries. Over two experiments, we found that subjects' behaviour was better explained by the asymmetric model, with an overall difference in learning rates consistent with preferential learning from positive, compared with negative, prediction errors.

Previous studies also suggest that the good news/bad news effect is highly variable across subjects[12]. Behavioural differences in optimistic beliefs and optimistic update have been shown to be reflected by differences in brain activation in the prefrontal cortex[8]. However, the question remains whether and how such inter-individual behavioural differences are related to the inter-individual neural differences in the extensively documented reward circuitry[15]. Our imaging results indicate that inter-individual variability in the tendency to optimistic learning correlates with prediction-error-related signals in the reward system, including the striatum and the ventro-medial prefrontal cortex (vmPFC).

Results

Behavioural task and dependent variables. Healthy subjects performed a probabilistic instrumental learning task with monetary feedback, previously used in brain imaging, pharmacological and clinical studies[16–18] (Fig. 1a). In this task, options (abstract cues) were presented in fixed pairs (conditions). In all conditions, each cue was associated with a stationary probability of reward.

[1]Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France. [2]Laboratoire d'Économie Mathématique et de Microéconomie Appliquée (LEMMA), Université Panthéon-Assas, 75006 Paris, France. [3]Amsterdam Brain and Cognition (ABC), Nieuwe Achtergracht 129, 1018 WS Amsterdam, The Netherlands. [4]Amsterdam School of Economics (ASE), Faculty of Economics and Business (FEB), Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. [5]INSERM-CEA Cognitive Neuroimaging Unit (UNICOG), NeuroSpin Centre, 91191 Gif sur Yvette, France. [6]Institut Jean-Nicod (IJN), CNRS UMR 8129, Ecole Normale Supérieure, 75005 Paris, France. [7]Institut d'Étude de la Cognition, Departement d'Études Cognitives, École Normale Supérieure, 75005 Paris, France. *e-mail: stefano.palminteri@ens.fr

In asymmetric conditions, the two reward probabilities differed between cues (25/75%). From asymmetric conditions, we extracted the rate of 'correct' response (selection of the best option) as a measure of performance (Fig. 1b, left). In symmetric conditions, both cues had the same reward probabilities (25/25% or 75/75%), such that there was no intrinsic 'correct response'. In symmetric conditions, we extracted, for each subject and each symmetric pair, a 'preferred response' rate, defined as the choice rate of the option most frequently selected by a given subject (by definition, in more than 50% of trials). The preferred response rate, especially in the 25/25% condition, should be taken as a measure of the tendency to overestimate the value of one instrumental cue relative to the other, in the absence of actual outcome-based evidence (Fig. 1b, right). In a first experiment ($N=50$) that subjects performed while being scanned by functional magnetic resonance imaging (fMRI), the task involved reward (+€0.50) and reward omission (€0), as the best and worst outcomes, respectively. In a second purely behavioural experiment ($N=35$), the task involved reward (+€0.50) and punishment (−€0.50), as the best and worst outcomes, respectively. All the results presented in the main text concern experiment 1, except those of the section entitled "Optimistic reinforcement learning is robust across different outcome valences". Detailed behavioural and computational analyses for experiment 2 are presented in the Supplementary Information.

Computational models. We fitted the behavioural data with two reinforcement-learning models[19]. The 'reference' model was represented by a standard Rescorla–Wagner model[20], hereafter referred to as the RW model. The RW model learns option values by minimizing reward prediction errors. It uses a single learning rate (alpha: α) to learn from positive and negative prediction errors. The 'target' model was represented by a modified version of the RW model, hereafter referred to as the RW± model. In the RW± model, learning from positive and negative prediction errors is governed by different learning rates (alpha plus, α^+, and alpha minus, α^-, respectively). For $\alpha^+ > \alpha^-$, the RW± model instantiates optimistic reinforcement learning (the good news/bad news effect); for $\alpha^+ = \alpha^-$, the RW± instantiates unbiased reinforcement learning, just as in the RW model (the RW model is thus nested in the RW± model); finally, for $\alpha^+ < \alpha^-$, the RW± instantiates pessimistic reinforcement learning. In both models, the choices are taken by feeding the option values into a softmax decision rule, whose exploration/exploitation trade-off is governed by a 'temperature' parameter (β).

Model comparison and analysis of model parameters. We used Bayesian model comparison to establish which model better accounted for the behavioural data. For each model, we estimated the optimal free parameters by maximizing the likelihood of the participants' choices, given the models and sets of parameters. For each model and each subject, we calculated the Bayesian information criterion (BIC) by penalizing the maximum likelihood with the number of free parameters in the model. Random-effects BIC analysis indicated that the RW± model explained the behavioural data better than the RW model ($BIC_{RW}=99.4 \pm 4.4$, $BIC_{RW\pm}=93.6 \pm 4.7$; $t(49)=2.9$, $P=0.006$, paired t-test), even after accounting for its additional degree of freedom. A similar result was obtained when calculating the model exceedance probability using the BIC as an approximation of the model evidence[21] (Table 1). Having established that RW± was the best-fitting model, we compared the learning rates fitted for positive (good news: α^+) and negative (bad news: α^-) prediction errors. We found α^+ to be significantly higher than α^- ($\alpha^+=0.36 \pm 0.05$, $\alpha^-=0.22 \pm 0.05$, $t(49)=3.8$, $P<0.001$, paired t-test). To summarize, model comparison indicated that, in our simple instrumental learning task, the best-fitting model is the model with different learning rates for learning from positive and negative predictions errors (RW±). Crucially, learning rates

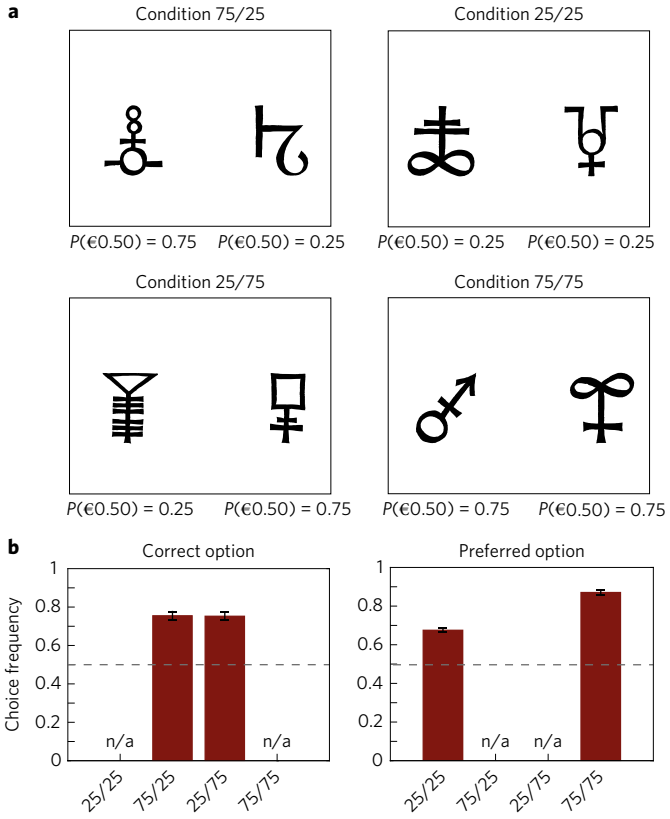

Figure 1 | Behavioural task and variables. a, The task's conditions and contingencies. Subjects selected between left and right symbols. Each symbol was associated with a stationary probability ($P=0.25$ or 0.75) of winning €0.50 and a reciprocal probability ($1-P$) of getting nothing (first experiment) or losing €0.50 (second experiment). In two conditions (rightmost column), the reward probability was the same for both symbols ('symmetric' conditions), and in two other conditions (leftmost column), the reward probability was different across symbols ('asymmetric' conditions). Note that the symbols-to-conditions assignment was randomized across subjects. **b**, Dependent variables. In the left panel, the histograms show the correct choice rate (that is, choices directed toward the most rewarding stimulus in the asymmetric conditions). In the right panel, the histograms show the preferred option choice rate (that is, the option chosen by subjects in more than 50% of the trials; this measure is relevant only in the symmetric conditions, where there is no intrinsic correct response). n/a, not applicable. Bars indicate the mean and error bars indicate the s.e.m. Data are taken from both experiments ($N=85$).

comparison indicated that instrumental values are preferentially updated following positive prediction errors, which is consistent with an optimistic bias operating when learning from immediate feedback (optimistic reinforcement learning).

Computational phenotyping. To categorize subjects, we computed for each individual the between-model BIC difference ($\Delta BIC = BIC_{RW} - BIC_{RW\pm}$) (see Methods). The ΔBIC quantifies, at the individual level, the improvement in goodness of fit on moving from the RW to the RW± model, or in other words, the fit improvement assuming different learning rates for positive and negative prediction errors. Subjects with a negative ΔBIC ($N=25$, in the first experiment) are subjects whose behaviour is better explained by the RW model and who therefore learn in an unbiased manner (hereafter referred to as RW subjects) (Fig. 2a). Subjects with a positive ΔBIC ($N=25$, in the first experiment) are subjects whose behaviour is better explained by asymmetric learning (hereafter referred to as RW± subjects).

Table 1 | Model fitting and parameters in the two experiments.

Experiment/model	LLmax	BIC	XP	MF	α	α^+	α^-	$1/\beta$
Experiment 1 ($N=50$)								
RW model	45.1 ± 2.2	99.4 ± 4.4	0.17	0.43	0.32 ± 0.05	–	–	0.16 ± 0.03
RW± model	40.0 ± 2.4	93.6 ± 4.7*	0.83	0.57	–	0.36 ± 0.05†	0.22 ± 0.05	0.13 ± 0.03
Experiment 2 ($N=35$)								
RW model	44.2 ± 2.9	97.6 ± 5.9	0.10	0.39	0.24 ± 0.05	–	–	0.53 ± 0.16
RW± model	38.1 ± 3.0	89.8 ± 6.0*	0.90	0.61	–	0.45 ± 0.06†	0.18 ± 0.05	0.30 ± 0.10

The table summarizes, for each model, its fitting performances and its average parameters: LLmax, maximal log likelihood; BIC, Bayesian information criterion (computed from LLmax); XP, exceedance probability; MF, model frequency; α, learning rate for both positive and negative prediction errors (RW model); α^+, learning rate for positive prediction errors; α^-, average learning rate for negative prediction errors (RW± model); $1/\beta$, average inverse of model temperature. Data are expressed as mean ± s.e.m. *$P<0.01$ comparing the two models. †$P<0.001$ comparing the two learning rates.

To test this hypothesis, learning rates fitted with the RW± model were entered into a two-way ANOVA with group (RW and RW±) and learning rate type (α^+ and α^-) as between- and within-subjects factors, respectively. The ANOVA showed a main effect of learning rate type ($F(1,48)=16.5$, $P<0.001$) with α^+ higher than α^-. We also found a main effect of group ($F(1,48)=10.48$, $P=0.002$) and a significant interaction between group and learning rate type ($F(1,48)=7.8$, $P=0.007$). Post-hoc tests revealed that average learning rates for positive prediction errors did not differ among the two groups: $\alpha^+_{RW}=0.45\pm0.08$ and $\alpha^+_{RW\pm}=0.27\pm0.06$ ($t(48)=1.7$, $P=0.086$,

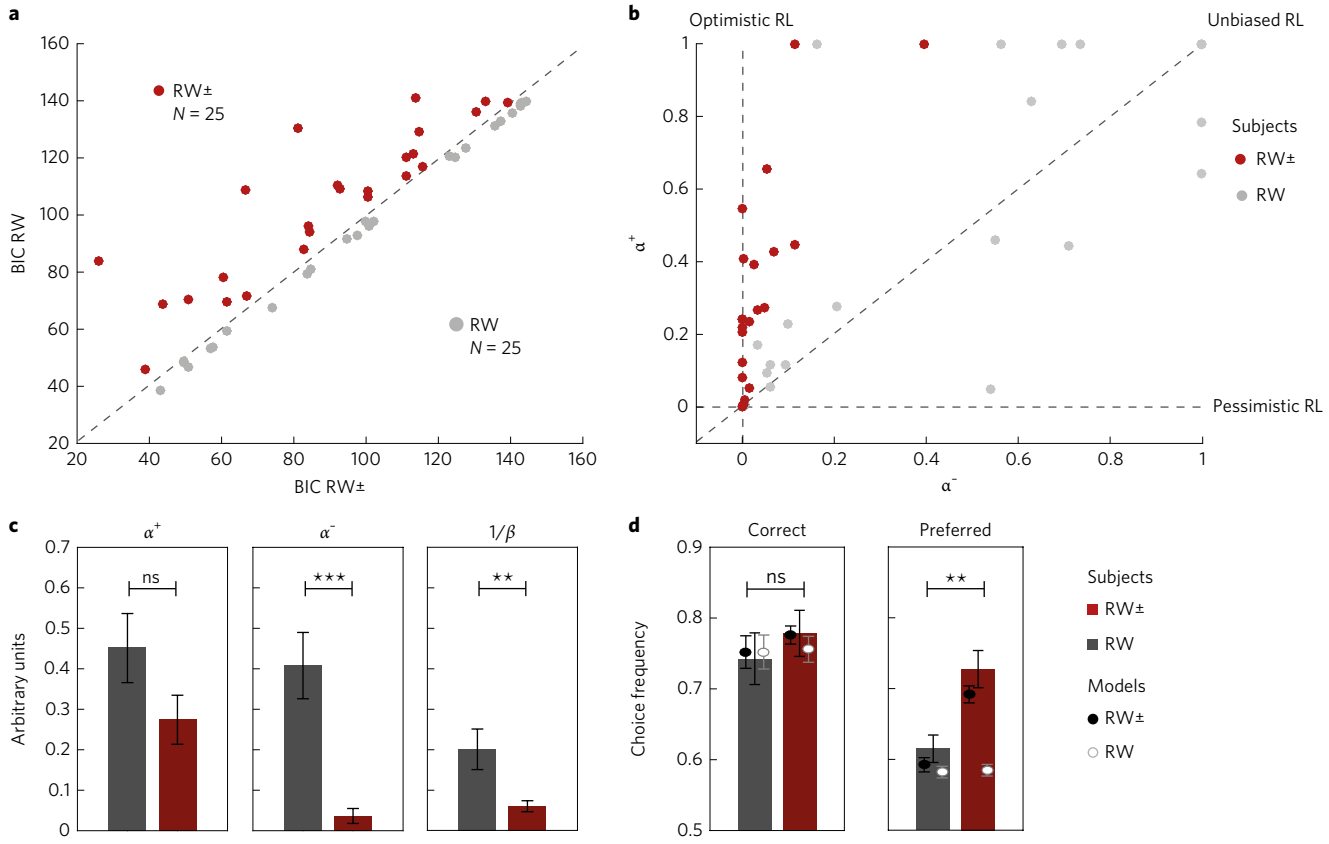

Figure 2 | Behavioural and computational identification of optimistic reinforcement learning. a, Model comparison. The graphic displays the scatter plot of the BIC calculated for the RW model as a function of the BIC calculated for the RW± model. Smaller BIC values indicate better fits. Subjects are clustered in two populations according to the BIC difference ($\Delta BIC = BIC_{RW} - BIC_{RW\pm}$) between the two models. RW± subjects (displayed in red) are characterized by a positive ΔBIC, indicating that the RW± model better explains their behaviour. RW subjects (grey) are characterized by a negative ΔBIC, indicating that the RW model better explains their behaviour. **b**, Model parameters. The graphic displays the scatter plot of the learning rate following positive prediction errors (α^+) as a function of the learning rate following negative prediction errors (α^-), obtained from the RW± model. 'Unbiased' learners are characterized by similar learning rates for both types of prediction errors. 'Optimistic' learners are characterized by a bigger learning rate for positive than for negative prediction errors. 'Pessimistic' learners are characterized by the opposite pattern. **c**, The histograms show the RW± model free parameters (the learning rates + and –, and the inverse temperature $1/\beta$) as a function of the subjects' populations. **d**, Actual and simulated choice rates. Histograms represent the observed and dots represent the model simulations of choices for both populations and both models, respectively for correct option (extracted from asymmetric condition), and for preferred option (extracted from the symmetrical condition 25/25%; see Fig. 1a). Model simulations are obtained using the individual best-fitting free parameters. Bars indicate the mean, and error bars indicate the s.e.m. ** $P<0.01$, ***$P<0.001$, two-sample, two-sided t-test. ns, not significant. Data are taken from the first experiment ($N=50$).

Table 2 | Behavioural and simulated data.

Experiment/model	Correct response	Correct response (RW model)	Correct response (RW± model)	Preferred response	Preferred response (RW model)	Preferred response (RW± model)
Condition(s)	**Asymmetric**			**Symmetric (25/25%)**		
Experiment 1 ($N=50$)						
RW group	74.25 ± 3.65	75.20 ± 2.55	75.35 ± 2.42	61.5 ± 1.94	58.14 ± 0.67	59.47 ± 0.81
RW± group	77.83 ± 3.25	75.58 ± 1.94	77.75 ± 1.44	72.75 ± 2.63*	58.84 ± 0.55	69.36 ± 0.99
Experiment 2 ($N=35$)						
RW group	73.28 ± 4.63	73.65 ± 3.58	73.72 ± 3.49	61.89 ± 2.12	57.34 ± 1.14	59.82 ± 1.35
RW± group	75.23 ± 4.73	75.29 ± 2.63	77.70 ± 1.86	73.73 ± 3.34*	58.33 ± 0.98	70.74 ± 2.08

The table summarizes for each experiment and each group of subjects, behavioural and simulated dependent variables: both real and simulated correct response rates in asymmetric conditions, and both real and simulated preferred response rates in 25/25% condition. Data are expressed as mean ± s.e.m (in percentage). *$P<0.01$, two-sample t-test.

two-sample t-test). In contrast, average learning rates for negative prediction errors were significantly different between groups, $\alpha^-_{RW}=0.41\pm0.08$ and $\alpha^-_{RW\pm}=0.04\pm0.02$ ($t(48)=4.6$, $P<0.001$, two-sample t-test). In addition, an asymmetry in learning rates was detected within the RW± group, where α^+ was higher than α^- ($t(24)=5.1$, $P<0.001$, paired t-test) but not within the RW group ($t(24)=0.9$, $P=0.399$, paired t-test). Thus, RW± subjects specifically drove the learning rate asymmetry found in the whole population. In contrast, the RW subjects displayed 'unbiased' (as opposed to 'optimistic') instrumental learning (Fig. 2b and c).

Interestingly, the exploration rate (captured by the $1/\beta$ 'temperature' parameter) was also found to be significantly different between the two groups of subjects, $1/\beta_{RW}=0.20\pm0.05$ and $1/\beta_{RW\pm}=0.06\pm0.01$ ($t(48)=2.9$, $P=0.006$, two-sample t-test). Importantly, the maximum likelihood of the reference model (RW) did not differ between the two groups of subjects, indicating similar baseline quality of fit (94.94 ± 5.00 and 103.91 ± 3.72 for RW and RW± subjects respectively, $t(48)=-1.0$, $P=0.314$, two-sample t-test). Accordingly, the difference in the exploration rate parameter cannot be explained by differences in the quality of fit (noisiness of the data). This suggests that optimistic reinforcement learning, observed in RW± subjects, is also associated with exploitative, as opposite to explorative, behaviour (Fig. 2c). Importantly, model simulation-based assessment of parameter recovery indicated that the two effects (learning rate asymmetry and lower exploration/exploitation trade-off) can be independently and correctly retrieved, ruling out the possibility that this twofold result is an artifact of the parameter optimization procedure (see Supplementary Information and Supplementary Fig. 7). To summarize, RW± subjects are characterized by two computational features: over-weighting positive prediction errors and over-exploiting previously rewarded options.

Behavioural signature distinguishing optimistic from unbiased subjects. To analyse the behavioural consequences of optimistic, as opposed to unbiased, learning and to confirm our model-based results with model-free behavioural observations, we compared the task's dependent variables for our two groups of subjects (Fig. 2d, Table 2). The correct response rate did not differ between groups ($t(48)=-0.7323$, $P=0.467$, two-sample t-tests). However, the preferred response rate in the 25/25% condition was significantly higher for the RW± group than for the RW group ($t(48)=-3.4$, $P=0.001$, two-sample t-test). Note that the same analysis performed on the 75/75% condition provided similar results ($t(48)=-2.66$, $P=0.01$, two-sample t-test).

To validate the ability of the RW± model to capture this difference, we performed simulations using both models and submitted them to the same statistical analysis as actual choices (Fig. 2d). The preferred response rates simulated using the RW± model were significantly higher in the RW± group compared with the RW group (25/25% $t(48)=-5.4496$, $P<0.001$; 75/75% $t(48)=-2.2670$, $P=0.028$; two-sample t-tests), which is in accordance with observed human behaviour. The preferred response rates simulated using the RW model were similar in the two groups (25/25% $t(48)=0.566$, $P=0.566$; 75/75% $t(48)=0.7448$, $P=0.4600$; two-sample t-test), which is in contrast with observed human behaviour. This effect was particularly interesting in a poorly rewarding environment (25/25%), where optimistic subjects tended to overestimate the value of one of the two options (Supplementary Fig. 1). Finally, the preferred response rate in the symmetric conditions significantly correlated with both the computational features distinguishing RW and RW± subjects (normalized learning rates asymmetry $(\alpha^+-\alpha^-)/(\alpha^++\alpha^-)$: $R=-0.475$, $P<0.001$; choice randomness $1/\beta$: $R=-0.630$, $P<0.001$). The preferred response rate thus provides a model-free signature of optimistic reinforcement learning that is congruent with our model simulation analysis: the preferred response rate was higher in the RW± group than the RW group, and only simulations realized with the RW± model were able to replicate this pattern of responses.

Neural signature distinguishing optimistic from unbiased subjects. To investigate the neural correlates of the computational differences between RW± and RW subjects, we analysed the brain activity both at the decision and outcome moments, using fMRI and a model-based fMRI approach[22]. We devised a general linear model in which we modelled the choice and the outcome onset as separate events, each modulated by different parametric modulators. In a given trial, the choice onset was modulated by the Q-value of the chosen option ($Q_{chosen}(t)$), and the outcome onset was modulated by the reward prediction error ($\delta(t)$). Concerning the choice onset, we found a neural network including the dorsomedial prefrontal cortex (dmPFC) and anterior insulae that negatively encoded $Q_{chosen}(t)$ ($P_{FWE}<0.05$ with a minimum of 60 continuous voxels, where FWE is family-wise error) (Fig. 3a and b; Table 3). We then tested for between-group differences within these two regions and found no significant difference (dmPFC: $t(48)=0.0985$, $P=0.9220$; insulae $t(48)=-0.0190$, $P=0.9849$; two-sample t-tests) (Fig. 3c). Concerning the outcome onset, we found a neural network including the striatum and vmPFC positively encoding $\delta(t)$ ($P_{FWE}<0.05$ with a minimum of 60 continuous voxels) (Fig. 3d and e; Table 3). We then tested for between-group differences within these two regions and found significant differences (striatum: $t(48)=-3.2769$, $P=0.0020$; vmPFC $t(48)=-2.2590$, $P=0.0285$; two-sample t-tests) (Fig. 3f). It therefore seems that the behavioural difference that we observed between RW and RW± subjects finds its counterpart in a differential outcome-related signal in the ventral striatum. Within the regions displaying a between-group difference, we looked for correlation with the two computational features distinguishing optimistic from unbiased subjects. Interestingly, we found a positive and significant correlation between the striatal and vmPFC $\delta(t)$-related activity and the normalized difference between

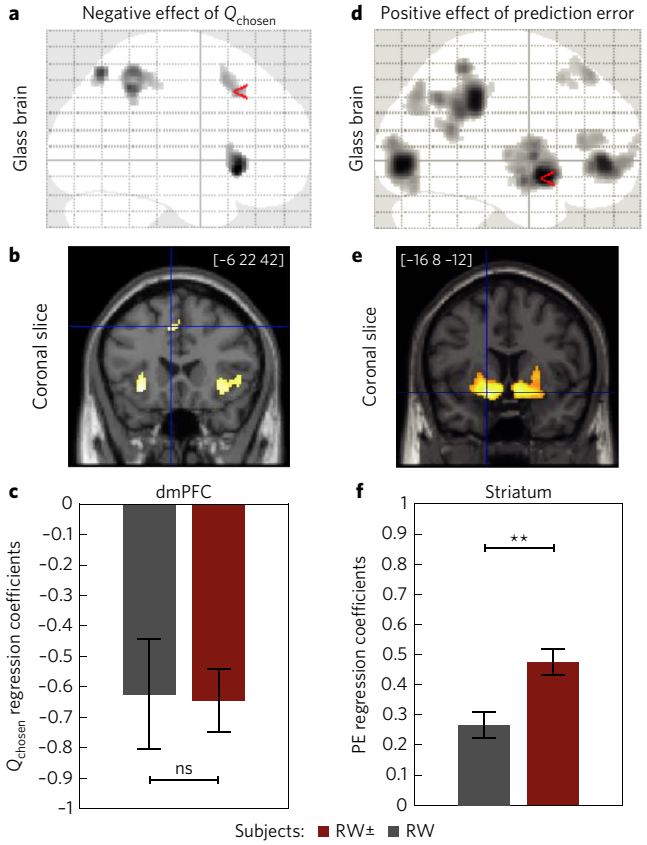

Figure 3 | Functional signatures of the optimistic reinforcement learning. **a**,**b**, Choice-related neural activity. Statistical parametric maps of blood oxygen level-dependent (BOLD) signal negatively correlating with the $Q_{chosen}(t)$ at the choice onset. Areas coloured in grey-to-black gradient on the axial glass brain and red-to-white gradient on the coronal slice show a significant effect ($P<0.001$ corrected). **c**, Inter-individual differences. Histogram shows $Q_{chosen}(t)$-related signal change in dmPFC at the time of choice onset for both populations. Bars indicate the mean, and error bars indicate the s.e.m. $^{*}P<0.05$, unpaired *t*-tests. Data are taken from the first experiment ($N=50$). **d**,**e**, Outcome-related neural activity. Statistical parametric maps of BOLD signal positively correlating with $\delta(t)$ at the outcome onset. Areas coloured in grey-to-black gradient on the axial glass brain and red-to-white gradient on the coronal slice show a significant effect ($P<0.001$ corrected). **f**, Inter-individual differences. Histogram shows $\delta(t)$-related signal change in the striatum at the time of reward onset for both populations. Bars indicate the mean, and error bars indicate the s.e.m. $^{**}P<0.01$ unpaired *t*-tests. Data are taken from the first experiment ($N=50$). [*x*, *y*, *z*] coordinates are given in the MNI (Montreal Neurological Institute) space.

learning rates (striatum: $R=0.4324$, $P=0.0017$; vmPFC: $R=0.3238$, $P=0.0218$), but no significant difference between the same activity and $1/\beta$ (striatum: $R=-0.130$, $P=0.366$; vmPFC: $R=-0.272$, $P=0.3665$), which suggests a specific link between this neural signature and the optimistic update.

Optimistic reinforcement learning is robust across different outcome valences. In the first experiment, getting nothing (€0) was the worst possible outcome. It could be argued that optimistic reinforcement learning (that is, greater learning rate for positive than negative prediction errors: $\alpha^{+}>\alpha^{-}$) is dependent on the low negative motivational salience attributed to a neutral outcome and would not resist if negative prediction errors were accompanied by actual monetary losses. To confirm the independence of our results from outcome valence, in the second experiment the worst possible outcome was represented by a monetary loss (−€0.50), instead of reward omission (€0) as in the first experiment.

First, the second experiment replicated the model comparison result of the first experiment. Group-level BIC analysis indicated that the RW± model again explains the behavioural data better than the RW model ($BIC_{RW}=97.6\pm5.9$, $BIC_{RW\pm}=89.8\pm6.0$), even after accounting for its additional degree of freedom ($t(34)=2.6414$, $P=0.0124$, paired *t*-test (Table 1 and Supplementary Fig. 4a).

To confirm that the asymmetry of learning rates is not a peculiarity of our first experiment, in which the worst possible outcome ('bad news') was represented by a reward omission, we performed a two-way ANOVA, with experiment (1 and 2) as the between-subject factor and learning rate type (α^{+} and α^{-}) as the within-subject factor. The analysis showed no significant effect of experiment ($F(1,83)=0.077$, $P=0.782$) and no significant interaction between valence and experiment ($F(1,83)=3.01$, $P=0.0864$), indicating that the two experiments were comparable, and, if anything, the effect size was bigger in the presence of punishments. Indeed, we found a significant main effect of valence ($F(1,83)=29.03$, $P<0.001$) on learning rates. Accordingly, post-hoc tests revealed that α^{-} was also significantly smaller than α^{+} in the second experiment ($t(34)=3.8639$, $P<0.001$ paired *t*-test) (Fig. 4f). These results confirm that optimistic reinforcement learning is not particular to situations involving only rewards but is still maintained in situations involving both rewards and punishments.

Discussion

We found that, in a simple instrumental learning task involving neutral visual stimuli associated to actual monetary rewards, participants preferentially updated option values following better-than-expected outcomes, compared with worse-than-expected outcomes. This learning asymmetry was replicated in two experiments and proved to be robust across different conditions. Our results support the hypothesis that the good news/bad news effect stands as a core psychological process generating and maintaining unrealistic optimism[9]. In addition, our study shows that this is not specific to probabilistic belief updating, and that the good news/bad news effect can parsimoniously be considered as an amplification of a primary instrumental learning asymmetry. In other terms, following recently proposed nomenclature[10], we found that asymmetric update applies to 'prediction errors' and not only to 'estimation errors', as reported in previous studies[10]. Recently, a debate emerged over whether the good news/bad news effect is an artifact due to the absence of positive life events and uncontrolled baseline event probabilities in the belief updating task[23,24]. Our results add to this debate by showing that the learning asymmetry persists in the presence of actual positive outcomes and controlled outcome probabilities.

The asymmetric model (RW±) included two different learning rates following positive and negative prediction errors, and we found the 'positive' learning rate to be higher than the 'negative' one[25,26]. When comparing RW± ('optimists') and RW ('unbiased') subjects, we found that the former had a significantly reduced negative learning rate. Thus, the good news/bad news effect seems to come not from overemphasizing positive prediction errors, but from underestimating negative ones. This is congruent with recent studies, in which optimism was related to reduced coding of undesirable information in the frontal cortex (right inferior frontal gyrus)[8], depression was linked to enhanced learning from bad news[13], and dopamine[27], as well as younger age[12], was related to diminished belief updating after the reception of negative information. However, since the RW± ('optimists') and RW ('unbiased') subjects also differed in the exploration rate, interpretations based on the absolute value of the learning rate, should be made with care.

The fact that the learning asymmetry was replicated when the negative prediction errors ('bad news') were associated with both

Table 3 | Activation table.

Variable	Chosen option value (negative correlation)		
Region (AAL)	**Coordinates [*x y z*]**	***t* value**	**Cluster size**
Insula (left)	−30 22 −8	7.15	131
Insula (right)	34 24 −6	6.76	147
Superior parietal gyrus	−20 −66 54	6.56	112
Angular gyrus	40 −48 44	6.52	288
Superior frontal gyrus/medial	−6 22 42	5.99	116
Variable	**Prediction error (positive correlation)**		
Region (AAL)	**Coordinates [*x y z*]**	***t* value**	**Cluster size**
Putamen	−16 8 −12	11.02	1137
Calcarine fissure	2 −84 −4	10.7	1346
Median cingulate	0 −36 36	10.17	1533
Caudate	10 6 −10	10.15	984
Anterior cingulate	−6 46 −4	9.56	911
Angular gyrus	−50 −44 56	7.68	103
Superior frontal gyrus/dorsolateral	−18 38 52	6.7	167
Angular gyrus	−40 −74 38	6.63	219
Inferior frontal gyrus/triangular part	−44 32 10	6.54	165
Cerebellum	22 −76 −18	6.41	62

FWE < 0.05, whole brain corrected and 60 minimum voxels. AAL, automatic anatomical labelling.

reward omissions (experiment 1) and monetary punishments (experiment 2) indicates that our results cannot be interpreted as a consequence of different processing of outcome values[28]. In other words, the learning asymmetry is not driven by the valence of the outcome but by the valence of the prediction error.

In principle, RW± subjects could have displayed both an optimistic and a pessimistic update, meaning that the ΔBIC is not — a priori — a measure of optimism. However, in the light of our results, this metric was a posteriori associated with the good news/bad news effect at the individual level. Categorizing subjects based on the ΔBIC, instead of the learning rate difference, has the advantage that the learning rate difference can take positive and negative values in RW subjects, but this difference merely captures noise, because it is not justified by model comparison. Our subject categorization was further supported by unsupervised Gaussian-mixtures analysis, which indicated that (1) two clusters explained the data better than one cluster and that (2) the two clusters corresponded to positive and negative ΔBIC respectively. The combination of individual model comparison with clustering techniques may represent a useful practice for computational phenotyping and for investigating cognitive differences between individuals[29].

A higher learning rate for positive compared with negative prediction errors was not the only computational metric distinguishing optimistic from unbiased subjects. In fact, we also found that optimistic subjects had a greater tendency to exploit a previously rewarded option, as opposed to unbiased subjects who were more prone to explore both options. Importantly, the higher stochasticity of unbiased subjects was associated neither with lower performance in the asymmetrical conditions, nor with a lower baseline quality of fit, as measured by the maximum likelihood. This overexploitation tendency was particularly striking in the symmetrical 25/25% condition, in which both options are poorly rewarding compared with the average task reward rate.

Whereas some previous studies suggest that optimists are more likely to explore and take risks (that is, they are entrepreneurs)[30], we found an association between optimistic learning and higher propensity to exploit. Indeed, the tendency to ignore negative feedback about chosen options was linked to considering a previously rewarded option better than it is, and hence to stick to this preference. A possible link between optimism and such 'conservatism' is not new; it can be dated back to Voltaire's work *Candide ou l'Optimisme*, where the belief of 'living in the best of all possible worlds' was consistently associated with a strong rejection and condemnation of progress and explorative behaviour. In the words of the eighteenth-century philosopher[31]:

> "Optimism," said Cacambo, "What is that?" "Alas!" replied Candide, "It is the obstinacy of maintaining that everything is best when it is worst."

Such optimism bias has been recently recognized as an important psychological factor that helps to maintain inaction over pressing social problems, such as climate change[32].

Recent studies have investigated the neural implementation of the good news/bad news effect when analysed in the context of probabilistic belief updating. At the functional level, decreased belief updating after worse-than-expected information has been associated with a reduced neural activity in the right inferior prefrontal gyrus[8]. Subsequent studies from the same group also showed that boosting dopaminergic function increases the good news/bad news effect and that this bias is correlated with striatal white matter connectivity, suggesting a possible role for the brain reward system[14,33]. In agreement with this, a more recent study showed differences in the reward system, including the striatum and the vmPFC[34]. Consistent with these results, we found that reward prediction error encoded in the brain reward network, including the striatum (mostly its ventral parts) and the vmPFC, was higher in optimistic than in unbiased subjects. Replicating previous findings, we also found a neural network, encompassing the dmPFC and the anterior insula, that negatively represented the chosen option value[35,36]. When comparing the two groups, we found no difference between optimists and pessimists in these decision-related areas[37,38]. Our results suggest that at the neural level, outcome-related activity discriminates between optimistic and unbiased subjects. Remarkably, by identifying functional differences between the two groups, our imaging data corroborate our model comparison-based classification of subject (neurocomputational phenotyping).

In our fMRI task, the outcome values (€0.50 or €0) and the prediction error signs (positive and negative) were coincident. This feature of the design undermines our capability to properly assess whether the observed neural difference is driven by a difference in outcome or prediction error encoding. Future research, involving paradigms in which outcome values and prediction error signs are orthogonalized, is needed to address this question. An important question is unanswered by our study and remains to be addressed. Although our results clearly show an asymmetry in the learning process, we cannot decide whether the learning process itself involves the representational space of values or that of probabilities. This question is related to the broader debate over whether the reinforcement or the Bayesian learning framework better captures learning and decision-making: two views that have been hard to disentangle, because of largely overlapping predictions, both at the behavioural and neural levels[39–41]. At this stage, our results cannot establish whether this optimistic bias is a valuation or a confirmation bias. In other terms, do subjects preferentially learn from positive prediction error because of its valence or because a positive prediction error 'confirms' the choice subjects just made? Future studies, decoupling valence from choice, are required to disentangle these two hypotheses.

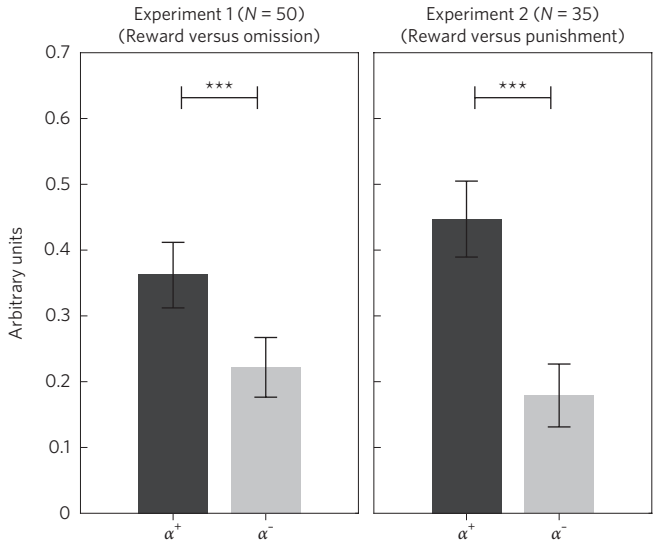

Figure 4 | Robustness of optimistic reinforcement learning. Histograms show the learning rates following positive prediction errors (α^+) and negative prediction errors (α^-), in experiment 1 ($N=50$) and experiment 2 ($N=35$). Experiment 1's worst outcome was getting nothing (€0). Experiment 2's worst outcome was losing money (−€0.50). Bars indicate the mean, and error bars indicate the s.e.m. $^{***}P<0.001$, paired t-tests.

It is worth noting that whereas some previous studies reported similar findings[42,43], another study reported the opposite pattern[44]. The difference between that study and ours might rely on the fact that the former involved Pavlovian conditioning[44]. It may therefore be argued that optimistic reinforcement learning is specific to instrumental (as opposite to Pavlovian) conditioning.

A legitimate question is why such learning bias has survived the course of evolution. An obvious answer is that being (unrealistically) optimistic is or has been, at least in certain conditions, adaptive, meaning that it confers an advantage. Consistent with this idea, in everyday life, dispositional optimism[45] has been linked to better global emotional well-being, interpersonal relationships or physical health. Optimists are less likely, for instance, to develop coronary heart disease[46], and they have a broader social network[47] and are less subject to distress when facing adversity[45]. Over-confidence in one's own abilities has been shown to be associated with higher performance in competitive games[48]. Such advantages of dispositional optimism could explain, at least in part, the pervasiveness of an optimistic bias in humans. Concerning the specific context of optimistic reinforcement learning, a recent paper[49] showed that in certain conditions (low rewarding environments), an agent learning asymmetrically in an optimistic manner (that is, with a higher learning rate for positive than for negative feedback) objectively outperforms another 'unbiased' agent in a simple probabilistic learning task. Thus, before any social, well-being or health consideration, it is normatively advantageous (in certain contingencies) to take more account of positive feedback than negative feedback. A possible explanation for an asymmetric learning system is, therefore, that the conditions identified in ref. [49] closely resemble the statistics of the natural environment that shaped the evolution of our learning system.

Finally, when reasoning about the adaptive value of optimism, a crucial point to consider is the inter-individual variability of unrealistic optimism[8,12–14]. As social animals, humans face both private and collective decision-making problems[50]. An intriguing possibility is that multiple 'sub-optimal' reinforcement learning strategies are maintained in the natural population to ensure an 'optimal' learning repertoire, flexible enough to solve, at the group level, the value learning and exploration–exploitation trade-off[51]. This hypothesis needs to be formally addressed using evolutionary simulations.

To conclude, our findings shed light on the nature of the good news/bad news effect and therefore on the mechanistic origins of unrealistic optimism. We suggest that optimistic learning is not specific to 'high-level' belief updating but a particular consequence of a more general 'low-level' instrumental learning asymmetry, which is associated to enhanced prediction error encoding in the brain reward system.

Methods

Subjects. The first dataset ($N=50$) served as a cohort of healthy control subjects in a previous clinical neuroimaging study[16]. The second dataset involved the recruitment of new subjects ($N=35$). The local ethics committees approved both experiments. All subjects gave written informed consent before inclusion in the study, and the study was carried out in accordance with the declaration of Helsinki (1964, revised 2013). In both studies, the inclusion criteria were being older than 18 years and having no history of neurologic or psychiatric disorders. In experiments 1 and 2, ratios of men to women were 27/23 and 20/15, respectively, and the age means were 27.1 ± 1.3 and 23.5 ± 0.7, respectively (expressed as mean ± s.e.m). In the first experiment, subjects believed that they would be playing for real money; the final pay-off was rounded up to a fixed amount of €80 for every participant. In the second experiment, subjects were paid the exact amount of money earned in the learning task, plus a fixed amount (average pay-off €15.70 ± 7.60).

Behavioural task and analyses. Subjects performed a probabilistic instrumental learning task described previously[17] (Fig. 1a). Briefly, the task involved choosing between two cues that were associated with stationary reward probability (25% or 75%). There were four pairs of cues, randomly constituted and assigned to the four possible combinations of probabilities (25/25%, 25/75%, 75/25% and 75/75%). Each pair of cues was presented 24 times, and each trial lasted on average 7 s. Subjects were encouraged to accumulate as much money as possible and were informed that some cues would result in a win more often than others (the instructions have been published in the appendix of the original study[17]). Subjects were given no explicit information on reward probabilities, which they had to learn through trial and error. The positive outcome reward was winning money (+€0.50); the negative outcome was getting nothing (€0) in the first experiment and losing money (−€0.50) in the second experiment. Subjects made their choice by pressing left or right (L or R) response buttons with a left- or right-hand finger. Two given cues were always presented together, thus forming a fixed pair (choice context).

Regarding pay-off, learning mattered only for pairs with unequal probabilities (75/25% and 25/75%). As dependent variable, we extracted the correct response rate in asymmetric conditions (the left response rate for the 75/25% pair and the right response rate for the 25/75% pair) (Fig. 1b). In symmetrical reward probability conditions, we calculated the 'preferred response rate'. The preferred response was defined as the most chosen option: that is, chosen by the subject in more than 50% of the trials. This quantity is therefore, by definition, greater than 50%. The analyses focused on the preferred choice rate in the low-reward condition (25/25%), for which standard models predict a greater frequency of negative prediction errors. Behavioural variables were compared within-subjects using a paired two-tailed t-test and between-subjects using a two-sample, two-tailed t-test. Interactions were assessed using ANOVA.

Computational models. We fitted the data with reinforcement learning models. The model space included a standard Rescorla–Wagner model (or Q-learning)[19,20] (hereafter referred to as RW) and a modified version of the latter accounting differentially for learning from positive and negative prediction errors (hereafter referred to as RW±)[26,43]. For each pair of cues, the model estimates the expected values of left and right options, Q_L and Q_R, on the basis of individual sequences of choices and outcomes. These Q-values essentially represent the expected reward obtained by taking a particular option in a given context. In the first experiment, which involved only reward and reward omission, Q-values were set at €0.25 before learning, corresponding to the a priori expectation of 50% chance of winning €0.50 plus a 50% chance of getting nothing. In the second experiment, which involved reward and punishment, Q-values were set at €0 before learning, corresponding to the a priori expectation of 50% chance of winning €0.50 plus 50% chance of losing €0.50. These priors on the initial Q-values are based on the fact that subjects were explicitly told in the instruction that no symbol was deterministically associated to either of the two possible outcomes, and that subjects were exposed to the average task outcome during the training session. Further control analyses, using post-training ('empirical') initial Q-values, were performed and are presented in the Supplementary Information and Supplementary Figure 6. After every trial t, the value of the chosen option (for example L) was updated according to the following rule:

$$Q_L(t+1) = Q_L(t) + \alpha\delta(t) \quad (1)$$

In equation (1), $\delta(t)$ was the prediction error, calculated as:

$$\delta(t) = R(t) - Q_L(t) \quad (2)$$

and $R(t)$ was the reward obtained as an outcome of choosing L at trial t. In other words, the prediction error $\delta(t)$ is the difference between the expected reward $Q_L(t)$ and the actual reward $R(t)$. The reward magnitude R was +0.5 for winning €0.50, 0 for getting nothing, and −0.5 for losing €0.50. The learning rate, α, is a scaling parameter that adjusts the amplitude of value changes from one trial to the next. Following this rule, option values are increased if the outcome is better than expected and decreased in the opposite case, and the amplitude of the update is similar following positive and negative prediction errors.

The modified version of the Q-learning algorithm (RW±) differs from the original one (RW) by its updating rule for Q values, as follows:

$$Q_L(t+1) = Q_L(t) + \begin{cases} \alpha^+\delta(t) & \text{if } \delta(t) > 0 \\ \alpha^-\delta(t) & \text{if } \delta(t) < 0 \end{cases} \quad (3)$$

The learning rate α^+ adjusts the amplitude of value changes from one trial to the next when prediction error is positive (when the actual reward $R(t)$ is better than the expected reward $QL(t)$), and the second learning rate α^- does the same when prediction error is negative. Thus, the RW± model allows the amplitude of the update to be different following positive ('good news') and negative ('bad news') prediction errors and permits us to account for individual differences in the way that subjects learn from positive and negative experience. If both learning rates are equivalent, $\alpha^+ = \alpha^-$, the RW± model equals the RW model. If $\alpha^+ > \alpha^-$, subjects learn more from positive than negative events. We refer to this case as optimistic reinforcement learning. If $\alpha^+ < \alpha^-$, subjects learn more from negative than positive events. We refer to this case as pessimistic reinforcement learning (Fig. 2b).

Finally, given the Q-values, the associated probability (or likelihood) of selecting each option was estimated by implementing the softmax rule for choosing L, which is as follows:

$$P_L(t) = e^{(Q_L(t)\beta)} / \left(e^{(Q_L(t)\beta)} + e^{(Q_L(t)\beta)}\right) \quad (4)$$

This is a standard stochastic decision rule that calculates the probability of selecting one of a set of options according to their associated values. The temperature, β, is another scaling parameter that adjusts the stochasticity of decision-making and by doing so controls the exploration–exploitation trade-off.

Model comparison. We optimized model parameters by minimizing the negative log-likelihood of the data given different parameters settings using Matlab's fmincon function, as previously described[52]. Additional parameter recovery analyses based on model simulations show that our parameter optimization procedure correctly retrieves the values of parameters (Supplementary Information and Supplementary Fig. 7). Negative log-likelihoods (LLmax) were used to compute at the individual level (random effects) the Bayesian information criterion for each model as follows:

$$\mathrm{BIC} = \log(n_{\mathrm{trials}})\mathrm{df} + 2\mathrm{LLmax} \quad (5)$$

Where df represent the degrees of freedom (that is, the number of free parameters) of the model. We then computed the inter-individual average BIC in order to compare the quality of fit of the two models, while accounting for their difference in complexity. The intra-individual difference in BIC ($\Delta\mathrm{BIC} = \mathrm{BIC}_{\mathrm{RW}} - \mathrm{BIC}_{\mathrm{RW}\pm}$) was also computed in order to categorize subjects into two groups (Fig. 2a): RW± subjects (those whose ΔBIC is positive) are better explained by the RW± model; RW subjects (whose ΔBIC is negative) are better explained by the RW model. We note that lower BIC indicated better fit. We also calculated the model exceedance probability and the model expected frequency based on the BIC as an approximation of the model evidence (Table 1). Individual BIC values were fed into the mbb-vb-toolbox, a procedure that estimates the expected frequencies and the exceedance probability for each model within a set of models, given the data gathered from all participants. Exceedance probability (denoted XP) is the probability that a given model fits the data better than all other models in the set.

The model parameters (α^+, α^- and $1/\beta$) were also compared between the two groups of subjects. Learning rates were compared using a mixed ANOVA with group (RW versus RW±) as a between-subject factor and learning rate type (+ or −) as a within-subject factor. The temperature was compared using a two-sample, two-tailed t-test. The normalized learning rates asymmetry $(\alpha^+ - \alpha^-) / (\alpha^+ + \alpha^-)$ was also computed as a measure of the good news/bad news effect and used to assess correlation with behavioural and neural data.

Subject classification. Subjects were classified based on the ΔBIC, which is the intra-individual difference in BIC between the RW and RW± model. While controlling for model parsimony, positive value indicates that the RW± better fits the data; negative value indicates the RW model better fit. The cut-off of $\Delta\mathrm{BIC} = 0$ is a priori meaningful because it indicates the limit beyond which there is enough (Bayesian) evidence to consider that a given subject's behaviour corresponds to a more complex model involving two learning rates. We also validated the $\Delta\mathrm{BIC} = 0$ cut-off a posteriori with unsupervised clustering. We fitted Gaussian mixed distributions to individual ΔBICs ($N = 85$, corresponding to the two experiments) using Matlab function gmdistribution.m. The analysis indicated that two clusters explain the variance significantly better than one cluster ($k = 1$, BIC = 716.4; $k = 2$, BIC = 635.6). The two clusters largely corresponded to subjects with negative ($N = 40$, min = −6.4; mean = −3.6, max = −0.9) and positive ΔBIC ($N = 45$, min = −0.5, mean = 15.7, max = 60.6). The two clusters differed in both the normalized difference in learning rates (0.14 versus 0.73; $t(83) = 7.2$, $P < 0.001$) and exploration rate (0.32 versus 0.09; $t(83) = 7.2$, $P = 0.006$).

Model simulations. We also analysed the models' generative performance by means of model simulations. For each participant, we devised a virtual subject, represented by a set of individual best-fitting parameters. Each virtual subject dataset was obtained averaging 100 simulations, to avoid any local effect of the individual history of choice and outcome. The model simulations included all task conditions. The evaluation of generative performances involved the assessment of the 'winning model's' ability to reproduce the key behavioural effect of the data, relative to the 'losing model'. Unlike Bayesian model comparison, model simulation comparison is bounded to a particular behavioural effect of interest (in our case the preferred response rate). The model simulation analysis, which is focused on the evidence against the losing model, is complementary to the Bayesian model comparison analysis, which is focused on the evidence in favour of the winning model (model falsification)[53,54].

Imaging data acquisition and analysis. Subjects of the first experiment ($N = 50$) performed the task during MRI scanning. T1-weighted structural images and T2*-weighted echo planar images (EPIs) were acquired during the first experiment and analysed with the Statistical Parametric Mapping software (SPM8; Wellcome Department of Imaging Neuroscience, London, UK). Acquisition and preprocessing parameters have been extensively described previously[16,17]. We refer to these publications for details about image acquisition and preprocessing.

Functional magnetic resonance imaging analysis. The fMRI analysis was based on a single general linear model. Each trial was modelled as having two time points, stimuli and outcome onsets. Each time point was regressed with a parameter modulator. Stimuli onset was modulated by the chosen option value, $Q_{\mathrm{chosen}}(t)$; outcome onset was modulated by the reward prediction error, $\delta(t)$. Given that different subjects did not implement the same model, the choice of the model used to generate the parametric regressors is not obvious. As the RW± and the RW models are nested and the RW± model was the group-level best-fitting model, we opted for using its parameters to generate the regressors. Note, however, that confirmatory analyses using, for each group, its best-fitting model's parameters lead to similar results. The parametric modulators were z-scored to ensure between-subject scaling of regression coefficients[55]. Linear contrasts of regression coefficients were computed at the subject level and compared against zero (one-sample t-test). Statistical parametric maps were threshold at $P < 0.05$ with a voxel-level family-wise error correction and a minimum of 60 contiguous voxels. Whole brain analysis included both groups of subjects and led to the identification of functionally characterized neural networks used to define unbiased regions of interest (ROIs). The dmPFC and the insular ROIs were defined as the intersection of the voxels significantly correlating with $Q_{\mathrm{chosen}}(t)$ and automatic anatomical labelling (AAL) masks of the medial frontal cortex (including the superior frontal gyrus, the supplementary motor area and the anterior medial cingulate) and the bilateral insula, respectively. The vmPFC and the striatal ROIs were defined as the intersection of the voxels significantly correlating with $\delta(t)$ and AAL masks of the ventral prefrontal cortex (including the anterior cingulate, the gyrus rectus and the superior frontal gyrus, orbital part and medial orbital part) and the bilateral caudate and putamen, respectively. Within ROIs, the regression coefficients were compared between-group using a two-sample, two-tailed t-test.

Data availability. The behavioural data are available here: https://dx.doi.org/10.6084/m9.figshare.4265408.v1. The fMRI results are available here: http://neurovault.org/collections/2195/.

Received 18 April 2016; accepted 10 February 2017; published 20 March 2017

References

1. Burtt, E. A. *The English Philosophers from Bacon to Mill* (Modern Library, 1939).
2. Weinstein, N. D. Unrealistic optimism about future life events. *J. Pers. Soc. Psychol.* **39,** 806–820 (1980).
3. Shepperd, J. A., Klein, W. M. P., Waters, E. A. & Weinstein, N. D. Taking stock of unrealistic optimism. *Perspect. Psychol. Sci.* **8,** 395–411 (2013).
4. Shepperd, J. A., Waters, E. A., Weinstein, N. D. & Klein, W. M. P. A primer on unrealistic optimism. *Curr. Dir. Psychol. Sci.* **24,** 232–237 (2015).
5. Shepperd, J. A., Ouellette, J. A. & Fernandez, J. K. Abandoning unrealistic optimism: performance estimates and the temporal proximity of self-relevant feedback. *J. Pers. Soc. Psychol.* **70,** 844–855 (1996).
6. Waters, E. A. *et al.* Correlates of unrealistic risk beliefs in a nationally representative sample. *J. Behav. Med.* **34,** 225–235 (2011).

7. Schoenbaum, M. Do smokers understand the mortality effects of smoking? Evidence from the health and retirement survey. *Am. J. Public Health* **87,** 755–759 (1997).
8. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* **14,** 1475–1479 (2011).
9. Eil, D. & Rao, J. M. The good news–bad news effect: asymmetric processing of objective information about yourself. *Am. Econ. J. Microecon.* **3,** 114–138 (2011).
10. Sharot, T. & Garrett, N. Forming beliefs: why valence matters. *Trends Cogn. Sci.* **20,** 25–33 (2016).
11. Sharot, T., Riccardi, A. M., Raio, C. M. & Phelps, E. A. Neural mechanisms mediating optimism bias. *Nature* **450,** 102–105 (2007).
12. Moutsiana, C. *et al.* Human development of the ability to learn from bad news. *Proc. Natl Acad. Sci. USA* **110,** 16396–16401 (2013).
13. Garrett, N. *et al.* Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Front. Hum. Neurosci.* **8,** 639 (2014).
14. Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X. & Sharot, T. Human frontal-subcortical circuit and asymmetric belief updating. *J. Neurosci.* **35,** 14077–14085 (2015).
15. Garrison, J., Erdeniz, B. & Done, J. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci. Biobehav. Rev.* **37,** 1297–1310 (2013).
16. Worbe, Y. *et al.* Reinforcement learning and Gilles de la Tourette syndrome: dissociation of clinical phenotypes and pharmacological treatments. *Arch. Gen. Psychiatry* **68,** 1257–1266 (2011).
17. 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).
18. Palminteri, S. *et al.* Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* **76,** 998–1009 (2012).
19. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
20. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning: Current Research and Theory* 64–99 (Appleton Century Crofts, 1972).
21. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* **10,** e1003441 (2014).
22. O'Doherty, J. P., Hampton, A. & Kim, H. Model-based fMRI and its application to reward learning and decision making. *Ann. N. Y. Acad. Sci.* **1104,** 35–53 (2007).
23. Shah, P., Harris, A. J. L., Bird, G., Catmur, C. & Hahn, U. A pessimistic view of optimistic belief updating. *Cogn. Psychol.* **90,** 71–127 (2016).
24. Sharot, T. & Garrett, N. The myth of a pessimistic view of optimistic belief updating — a commentary on Shah *et al.* Preprint at http://dx.doi.org/10.2139/ssrn.2811752 (2016).
25. Doll, B. B., Hutchison, K. E. & Frank, M. J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.* **31,** 6188–6198 (2011).
26. Niv, Y. *et al.* Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35,** 8145–8157 (2015).
27. Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R. & Dolan, R. J. How dopamine enhances an optimism bias in humans. *Curr. Biol.* **22,** 1477–1481 (2012).
28. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **47,** 263–292 (1979).
29. Huys, Q. J. M., Maia, T. V & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19,** 404–413 (2016).
30. Sharot, T. *The Optimism Bias: Why We're Wired to Look on the Bright Side* (Robinson, 2012).
31. Voltaire. *Candide, or Optimism* (Penguin, 2013).
32. Gifford, R. The dragons of inaction: pychological barriers that limit climate change mitigation and adaptation. *Am. Psychol.* **66,** 290–302 (2011).
33. Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R. & Dolan, R. J. How dopamine enhances an optimism bias in humans. *Curr. Biol.* **22,** 1477–1481 (2012).
34. Kuzmanovic, B., Jefferson, A. & Vogeley, K. The role of the neural reward circuitry in self-referential optimistic belief updates. *Neuroimage* **133,** 151–162 (2016).
35. Skvortsova, V., Palminteri, S. & Pessiglione, M. Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. *J. Neurosci.* **34,** 15621–15630 (2014).
36. 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).
37. Domenech, P. & Koechlin, E. Executive control and decision-making in the prefrontal cortex. *Curr. Opin. Behav. Sci.* **1,** 101–106 (2015).
38. Kolling, N., Behrens, T. E. J., Wittmann, M. K. & Rushworth, M. F. S. Multiple signals in anterior cingulate cortex. *Curr. Opin. Neurobiol.* **37,** 36–43 (2016).
39. Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **5,** 39 (2011).
40. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18,** 1159–1167 (2015).
41. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26,** 8360–8367 (2006).
42. van Den Bos, W., Cohen, M. X., Kahnt, T. & Crone, E. A. Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cereb. Cortex* **22,** 1247–1255 (2012).
43. Frank, M. J., Moustafa, A. A, Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl Acad. Sci. USA* **104,** 16311–16316 (2007).
44. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32,** 551–562 (2012).
45. Carver, C. S., Scheier, M. F. & Segerstrom, S. C. Optimism. *Clin. Psychol. Rev.* **30,** 879–889 (2010).
46. Tindle, H. A. *et al.* Optimism, cynical hostility, and incident coronary heart disease and mortality in the Women's Health Initiative. *Circulation* **120,** 656–662 (2009).
47. Macleod, A. K. & Conway, C. Well-being and the anticipation of future positive experiences: the role of income, social networks, and planning ability. *Cogn. Emot.* **19,** 357–374 (2005).
48. Johnson, D. D. P. & Fowler, J. H. The evolution of overconfidence. *Nature* **477,** 317–320 (2011).
49. Cazé, R. D. & van der Meer, M. A. A. Adaptive properties of differential learning rates for positive and negative outcomes. *Biol. Cybern.* **107,** 711–719 (2013).
50. Raafat, R. M., Chater, N. & Frith, C. Herding in humans. *Trends Cogn. Sci.* **13,** 420–428 (2009).
51. Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D. & Couzin, I. D. Exploration versus exploitation in space, mind, and society. *Trends Cogn. Sci.* **19,** 46–54 (2015).
52. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6,** 8096 (2015).
53. Popper, K. *The Logic of Scientific Discovery* (Routledge, 2005).
54. Dienes, Z. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference* (Palgrave Macmillan, 2008).
55. Lebreton, M. & Palminteri, S. Assessing inter-individual variability in brain-behavior relationship with functional neuroimaging. Preprint at *bioRxiv* http://dx.doi.org/10.1101/036772 (2016).

Acknowledgements

We thank Y. Worbe and M. Pessiglione for granting access to the first dataset, V. Wyart, B. Bahrami and B. Kuzmanovic for comments, and T. Sharot and N. Garrett for providing activation masks. S.P. was supported by a Marie Sklodowska-Curie Individual European Fellowship (PIEF-GA-2012 Grant 328822) and is currently supported by an ATIP-Avenir grant (R16069JS). G.L. was supported by a PhD fellowship of the Ministère de l'enseignement supérieur et de la recherche. M.L. was supported by an EU Marie Sklodowska-Curie Individual Fellowship (IF-2015 Grant 657904) and acknowledges the support of the Bettencourt-Schueller Foundation. The second experiment was supported by the ANR-ORA, NESSHI 2010–2015 research grant to S.B.-G. The Institut d'Étude de la Cognition is supported by the LabEx IEC (ANR-10-LABX-0087 IEC) and the IDEX PSL* (ANR-10-IDEX-0001-02 PSL*). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

G.L. performed the experiment, analysed the data and wrote the manuscript. M.L. provided analytical tools, interpreted the results and edited the manuscript. F.M. provided analytical tools and edited the manuscript. S.B.-G. interpreted the results and edited the manuscript. S.P. designed the study, performed the experiments, analysed the data and wrote the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.P.

How to cite this article: Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S. & Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1,** 0067 (2017)

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing interests.