nature human behaviour

SUPPLEMENTARY INFORMATION

VOLUME: 1 | ARTICLE NUMBER: 0067

In the format provided by the authors and unedited.

Behavioral and neural characterization of optimistic reinforcement learning

Germain Lefebvre, Maël Lebreton, Florent Meyniel, Sacha Bourgeois-Gironde & Stefano Palminteri

Supplementary Notes and Figures

Preferred response rate as a behavioral measure of optimistic behavior

As previously defined in the main text, the preferred response rate is the rate of the choices directed toward the most frequently chosen option by subjects in symmetric reward probability conditions (i.e. 25/25% and 75/75%). The preferred choice rate is therefore by definition greater than 0.5. In these conditions there is no contingency-based reason to prefer one option to the other. This is particularly true in low-rewarding environment (25/25% condition), where neither option is satisfying in terms of outcome, compared to the average task outcome. We showed previously that the preferred response rate allows to behaviorally differentiating optimistic from unbiased subjects.

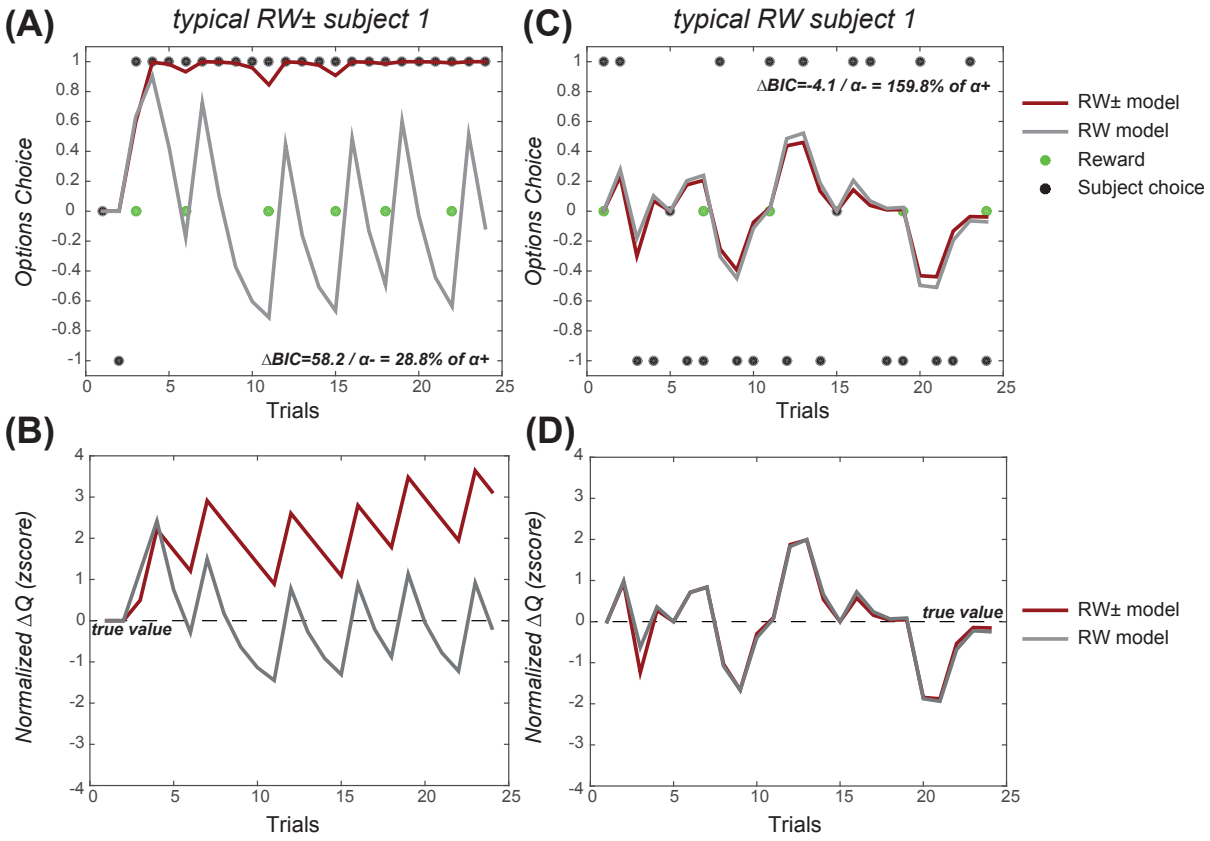

Figure S1: typical "optimistic" and "unbiased" subjects in the 25/25% condition.
(A) and **(B)** RW± (optimistic) typical subject. (**A**) Plot represents behavioral choices (represented by black dots) of a typical RW± subject (i.e. whose behavior is best fitted by the RW± model) in the 25/25% condition, together with RW and RW± models predictions (represented respectively by gray and red lines). (**B**) Plot represents Q-values (of the two options) differential evolution in each model for a typical RW± subject. **(C)** and **(D)** RW (unbiased) typical subject. **(C)** Plot represents behavioral choices (represented by black dots) of a typical RW subject (whose behavior is best fitted by the RW model) in the 25/25% condition, together with RW and RW± models predictions (represented respectively by gray and red lines). (**D**) Plot represents the evolution of Q-values (of the two options) differential in each model for a typical RW subject.

RW± subjects were characterized at the computational level by two features that are good news/bad news effect (i.e. learning rate asymmetry) and lower exploration rate. Both these computational features concur to generate this behavioral pattern. **Fig. S1A** shows the preferred choice rate in the 25/25% condition of a typical RW± subject, whose behavior is much better explained by the RW± model (ΔBIC=58.2).

We clearly see that his choices are stabilized toward one option (preferred response rate=0.94), after one single reward event. RW± model fit captures this preference, by giving less weight to negative feedback than to positive one (α^- is approximately four times smaller than α^+) and by allowing a very little exploration rate ($1/\beta$=0.001). This learning rate asymmetry creates and accentuates over the trials the "preferred minus non-preferred" ΔQ (whose true value is zero), which is further reinforced by not exploring the other option (**Fig.S1B**). At the opposite (**Fig.S1C**), a typical unbiased subject (RW) does not show clear preference toward one of the options (preferred response rate around 0.58). Accordingly the RW model better explains his behavior (ΔBIC=-4.1) and his learning rate asymmetry is moderate (α^- is approximately 1.6 time higher than α^+) and the exploration rate higher ($1/\beta$=0.235). This symmetry between positive and negative learning rates and the tendency to extensively explore the two options do not give advantage to any of the Q-Values, the differential of which gravitates around zero over learning (i.e its true value) (**Fig.S1B**). These examples nicely illustrate how the preferred response rate is affected by learning rate asymmetry and exploration rate thus allowing discriminating RW± and RW subjects.

In order to further illustrate how the preferred response rate relates to both computational signatures of optimistic reinforcement learning, we run model simulations distinguishing the effect of the latter two features, a positive learning rates asymmetry and a low exploration rate. We ran four simulations by experiment using parameters from either typical RW subjects or typical RW± subjects. The simulations were generated using the parameter sets of both experiments. In the simulations based on Experiment 1 (N=100 virtual subjects) we used either symmetric learning rates (RW: $\alpha^+ = \alpha^- = 0.41$) or optimistically asymmetric (RW± : $\alpha^+ = 0.27$ and $\alpha^- = 0.04$) and the exploration rate was either low (RW±: $1/\beta = 0.06$) or high (RW: $1/\beta = 0.21$). In the simulations based on Experiment 2 (N=105 virtual subjects) we used either symmetric learning rates (RW: $\alpha^+ = \alpha^- = 0.27$) or optimistically asymmetric (RW±: $\alpha^+ = 0.47$ and $\alpha^- = 0.10$) and the exploration rate was either low ($1/\beta = 0.15$) or high ($1/\beta = 0.73$). Results from those simulations presented in the **Fig. S2**, showed that neither of the two computational features alone (learning rate asymmetry and lower exploration rate) are sufficient to reach the preferred response rate of RW± subjects. On the contrary, simulations ran with both computational features permit to the preferred response rate to be very close to the empirical results of RW± subjects. In other terms, the learning rate asymmetry and the exploration rate have as super-additive effect on the preferred response rate, which is an essential characteristic of the RW± (optimistic) phenotype.

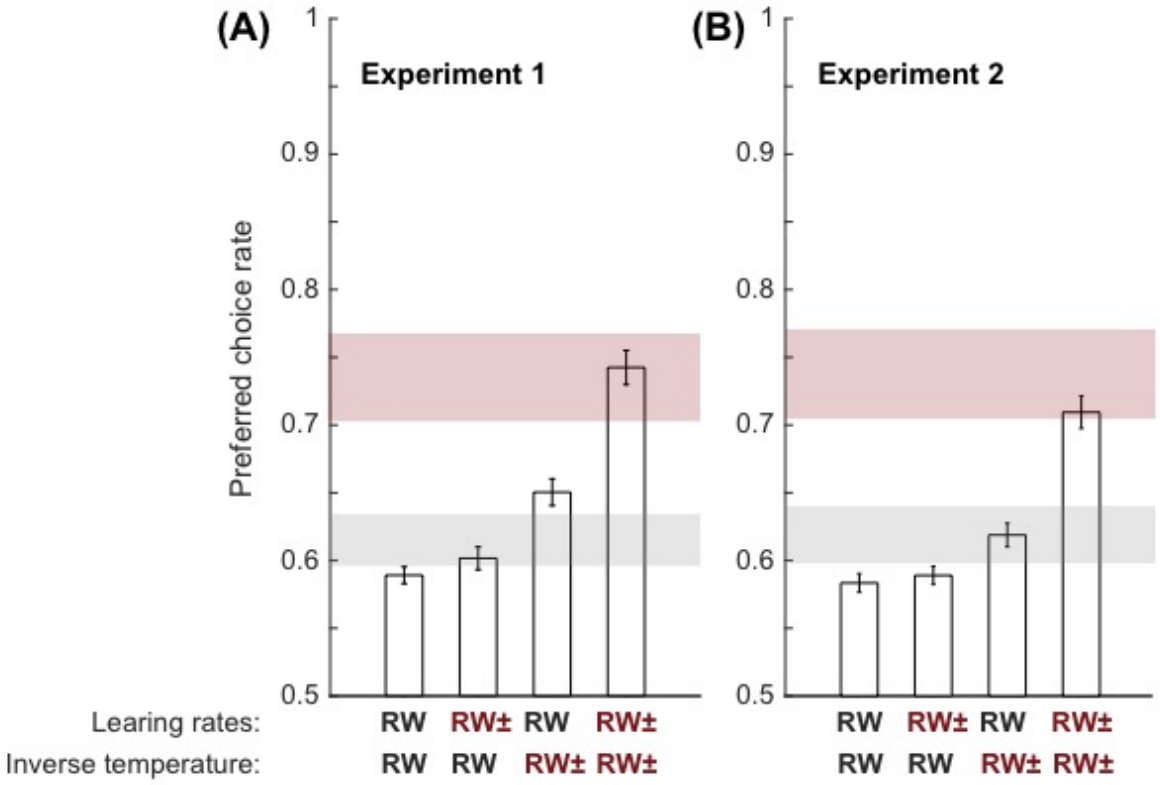

Figure S2: contribution of the learning rate asymmetry and the choice inverse temperature to the preferred choice rate. (A) and **(B)** Bars represent the simulated preferred response rate in four conditions varying according to the model parameters used to simulate the data. RW learning rates are symmetrical and correspond to the average learning rates of RW subjects while RW± learning rates are asymmetrical and correspond to the average learning rates of RW± subjects. The RW inverse temperature is high and matches the average inverse temperature of RW subjects while RW± inverse temperature is low and matches the average inverse temperature of RW± subjects. Finally, the horizontal colored areas represent the average empirical preferred response rate plus or minus its standard deviation to the mean, in RW subjects (grey area) and RW± subjects (red area).

Optimistic reinforcement learning is robust across different learning phases

Previous studies have shown that learning rates adapt with learning. More precisely the learning rate may be reduced when the confidence about choice's outcome is high and, conversely, augmented in situation of high uncertainty[1,2]. It could be argued that in our task the optimistic learning rate asymmetry ($\alpha^+ > \alpha^-$) was specifically driven by the late trials, when the reward contingences have been learnt and the subject has no longer need to monitor prediction errors as objectively as in the early trials. In order to assess that the learning rate asymmetry was not specific of the late learning, we analyzed and compared RW± model parameters separately optimized in the first and second halves of the task in both experiments (**Fig. S3A**). We performed a two-way ANOVA with part of the task (first and second halves) and learning rate type (α^+ and α^-) as within-subject factors. It shows no significant effect of task period ($F(1,84)=0.011$, $P=0.917$) and no significant valence x period interaction ($F(1,84)=0.011$, $P=0.917$) indicating that the learning rates asymmetry is not specific to the late trials. We found indeed a significant main effect of valence ($F(1,84)= 46.42$, $P<0.001$) on learning rates. Accordingly, post-hoc test revealed that α^- was significantly smaller than α^+ also in the first half ($t(84)=5.7214$, $p<0.001$ paired t-test) and in the second half ($t(84)= 5.3764$, $p<0.001$ paired t-test) of the task.

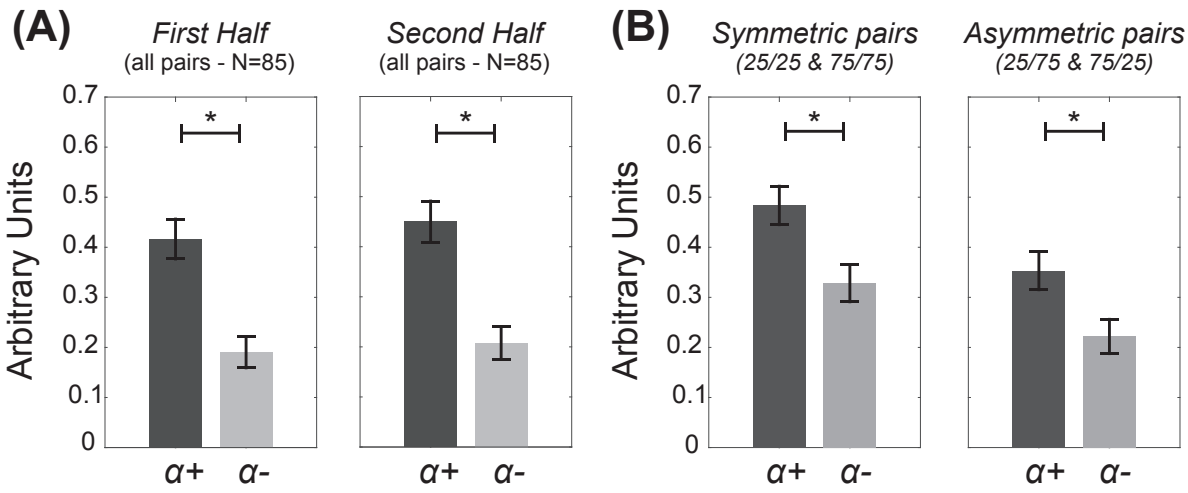

Figure S3: Robustness of optimistic reinforcement learning.
(A) Control first half / second half. Histograms show the learning rates following positive prediction errors (α^+) and negative prediction errors (α^-), obtained from parameters optimization separately performed in the first and second halves of the experiments (N=85). **(B)** Control symmetric / asymmetric conditions. Histograms show the learning rates following positive prediction errors (α^+) and negative prediction errors (α^-), obtained from parameters optimization involving only the "symmetric or the "asymmetric" conditions (N=85).

Optimistic reinforcement learning is robust across different outcome contingencies

It has also been proposed that learning rates may adapt as a function of task contingencies[3]. In our task the macroscopic (aggregate) model-free signature of optimistic behavior was found in the symmetrical conditions: higher preferred choice rate in the RW± subjects (**Fig. 2** and **S4**). In the main text we reported the results concerning the 25/25% condition, but this was also true for the 75/75% condition, where the preferred choice rate in the RW± subjects was higher compared to the RW subjects ($t(83)=3.6686$, $p<0.001$, two-sample t-test) and compared to what was predicted by the RW model ($t(42)=16.0292$, $p<0.001$, paired t-test). It might be argued that the learning rate asymmetry we observed was driven by an adaptation of the learning rates specific to the symmetrical conditions, in which there is no true correct response.

In order to verify that the asymmetry of the learning rate was not only expressed in the symmetric conditions (when options are equally rewarding), we optimized learning rates in symmetric and asymmetric conditions independently (**Fig. S3B**). A two-way ANOVA devised with condition type (symmetric and asymmetric) and learning rates valence as within subjects factors, showed a main effect of valence ($F(1,84)=21.14$, $P<0.001$) that is consistent with α^+ being higher compared to α^-. It also showed a lower effect of condition type ($F(1,84)= 9.493$, $P=0.003$) both learning rates being lower in asymmetric conditions, but importantly no significant interaction between valence and condition type ($F(1,84)=0.124$, $P=0.726$). Post hoc tests confirm this learning rates asymmetry ($\alpha^+ > \alpha^-$ in both condition types ($t(84)=3.1106$, $p=0.003$ in asymmetric conditions and $t(84)= 3.139$, $p=0.002$ in symmetric conditions, paired t-tests).

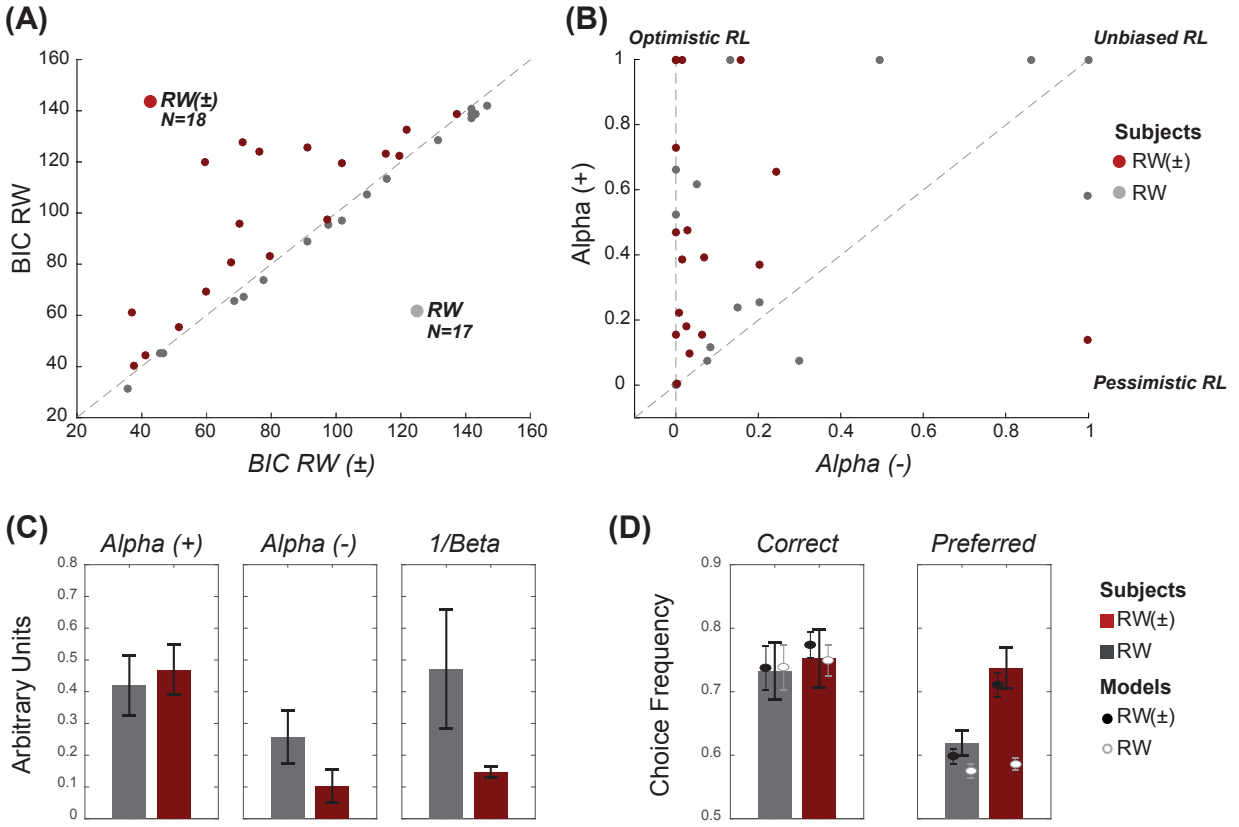

Figure S4: Replication of the computational and behavioral results in another group of subjects and using actual punishments

In order to assess the robustness of optimistic reinforcement learning in presence of actual punishments, we run an additional experiment (Experiment 2; N=35). The probabilistic contingencies, as well as the number of trials, were similar in both experiments. However, whereas Experiment 1's worst outcome was getting nothing (0€), Experiment 2's worst outcome was losing money (-0.50€). The two experiments led to strikingly similar behavioral and computational results (**Fig. 2**). **(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. 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 behavior. RW subjects (displayed in grey) are characterized by a negative ΔBIC, indicating that the RW model better explains their behavior. **(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" reinforcement learning (RL) is characterized by similar learning rates for both types of prediction errors. "Optimistic" RL is characterized by a bigger learning rate only for positive compared to negative prediction errors. "Pessimistic" RL is characterized by the opposite pattern. **(C)** The histograms represent the RW± model free parameters (the learning rates and the inverse temperature 1/beta) as function of the subjects' populations. **(D)** Actual and simulated choice rates. Histograms represent the observed and dots represent the model simulated of choices for both populations and both models, respectively for correct option (extracted from asymmetric condition), and from preferred option (extracted from the symmetric condition 25/25%, see **Fig. 1A**).

Indeed, an in depth analysis of correct response rate distribution (**Fig. S5**) showed that both behavioral and simulated response distributions are significantly different between groups although being similar within groups. The analysis focused on the distributions of correct response rates in both asymmetric conditions (25/75% and 75/25%) among real and simulated populations both dichotomized in RW± and RW subjects.

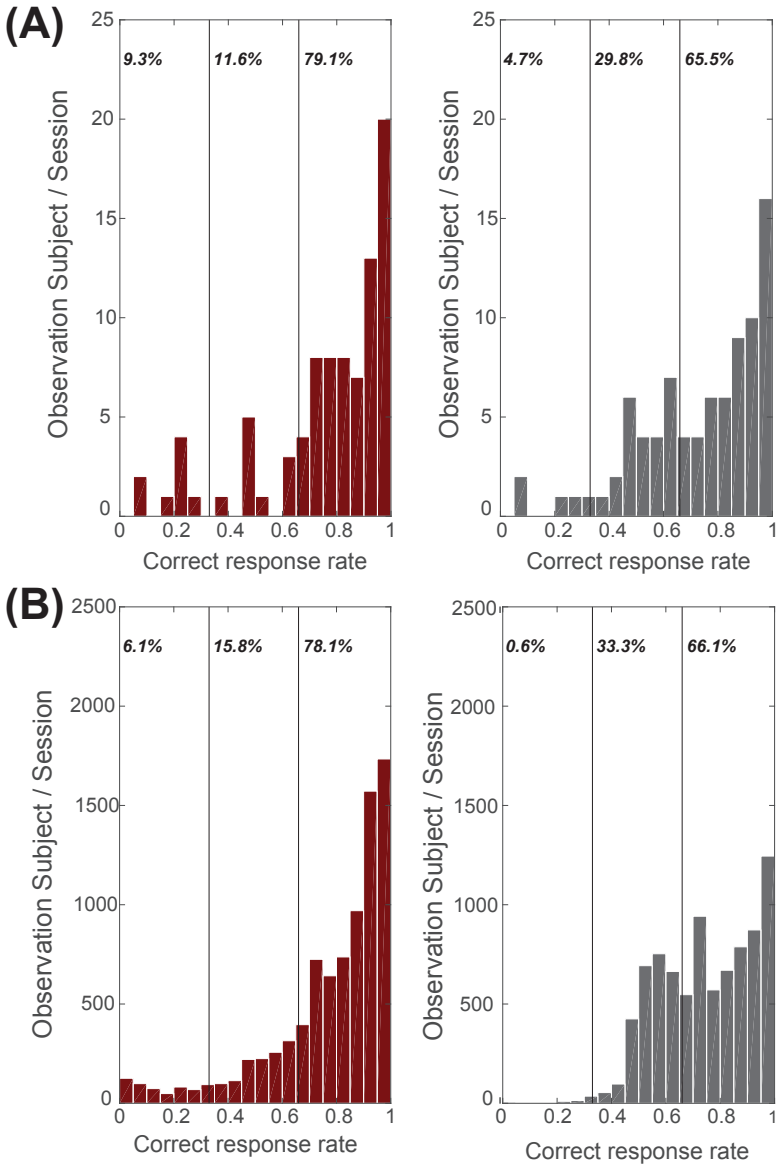

Figure S5: actual and modeled distributions of correct choice frequency
(**A)** Histograms represent distributions of correct choice rate in asymmetric conditions (25/75%, 75/25%) in RW and RW± subjects. Data are taken from both experiments (N=85). (**B)** Histograms represent distributions of correct choice rate in asymmetric conditions (25/75%, 75/25%) in RW and RW± virtual subjects. Each virtual subject (correspond to an individual set of free parameters obtained fitting the actual data with the RW± model) played the task one hundred times (N=8500). RW± real and virtual subjects are characterized by a higher frequency of "extreme" (i.e. greater than 0.66 or lower than 0.33) correct response rate, whereas RW real and virtual subjects, are characterized by a higher frequency of "intermediate" correct response rate.

To compare distributions, we split the correct response rate into three equal categories and calculated the percentage of observations belonging to each category. A first comparison between RW and RW± real populations (**Fig. S5A**) showed that their correct response rate distributions were significantly different (χ^2=10.69, p<0.005, chi-squared test). The RW± signature here being that distributions are marked by a greater presence of extreme responses due to the sensitivity (greater α^+ and lower exploration) of RW± subjects to both reward received from the correct option (79.1% and 65.5% respectively), but also to reward accidentally received from the incorrect option (9.3% and 4.7% for RW± and RW respectively). So

the insensitivity of RW± subjects to negative feedback and their relatively low tendency to explore available options make them prone to choose and stick with the worst available option. To test whether or not this feature of RW± subjects was captured by the optimistic reinforcement model, we realized the same analysis in a population of virtual RW± et RW subjects (**Fig. S5B**). Firstly, we found that simulated distributions of correct response rate are equivalent to behavioral ones (χ^2=3.49, p=0.17, for RW group and χ^2=1.09, p=0.58, for RW± group, chi-squared tests). Secondly and similarly to real distributions, simulated distributions were found to be significantly different between groups (χ^2=12.66, p<0.005, chi-squared test) with a similar representation of extreme correct response rates in RW± group compared to RW group. Being "optimistic" in this task is then often advantageous for a majority of optimistic subjects displaying a very high rate of correct option. However, when optimists receive a probabilistic reward from the worst option, they can be trapped by their insensitivity to negative feedback and by not being prone to explore alternatives.

Optimistic reinforcement learning is robust across different Q-values initializations

Learning rates asymmetry was obtained through parameters optimization using original and derivative Q-Learning models. As indicated in the **Methods** section, subjects were induced to have "neutral" priors about each stimulus value via the instructions and the training session. Accordingly, Q values were set at 0.25€ before learning, corresponding in the first experiment (that involved only reward +0.5€ and reward omission 0.0€) to the a priori expectation of 50% chance of winning 0.5€ plus a 50% chance of getting nothing. In the second experiment (which involved reward +0.5€ and punishment -0.5€) Q values were set at 0.0€ before learning, corresponding to the a priori expectation of 50% chance of winning 0.5€ plus 50% chance of losing 0.5€. In order to verify the robustness of our result in respect of the Q-value initialization, we performed another parameter optimization using the same models but initializing Q-values using individual "empirical" priors. We defined the "empirical" priors as the average outcome observed during the training session averaged across all the stimuli: we found 0.23±0.004€ (in Experiment 1) and -0.02 ± 0.01€ (in Experiment 2). These values are very close to the theoretical values used in the analysis (0.25€ and 0.00€), except for a small under-estimation that is due to the fact that the worst stimuli are less extensively sampled. Parameters optimized using initial empirical values confirmed, once again, a learning rate asymmetry, consistent with the good news/bad news effect (α^+ = 0.32±0.06 and α^- = 0.17±0.06, t(29)=3.12, p=0.0041 for Experiment 1 and α^+ = 0.46±0.06 and α^- = 0.19±0.05, t(33)=3.73, p<0.001 for Experiment 2) (**Fig S6**). Thus, empirically determined priors were similar to theoretical ones and using the firsts in our parameter optimization procedure have no impact on the results.

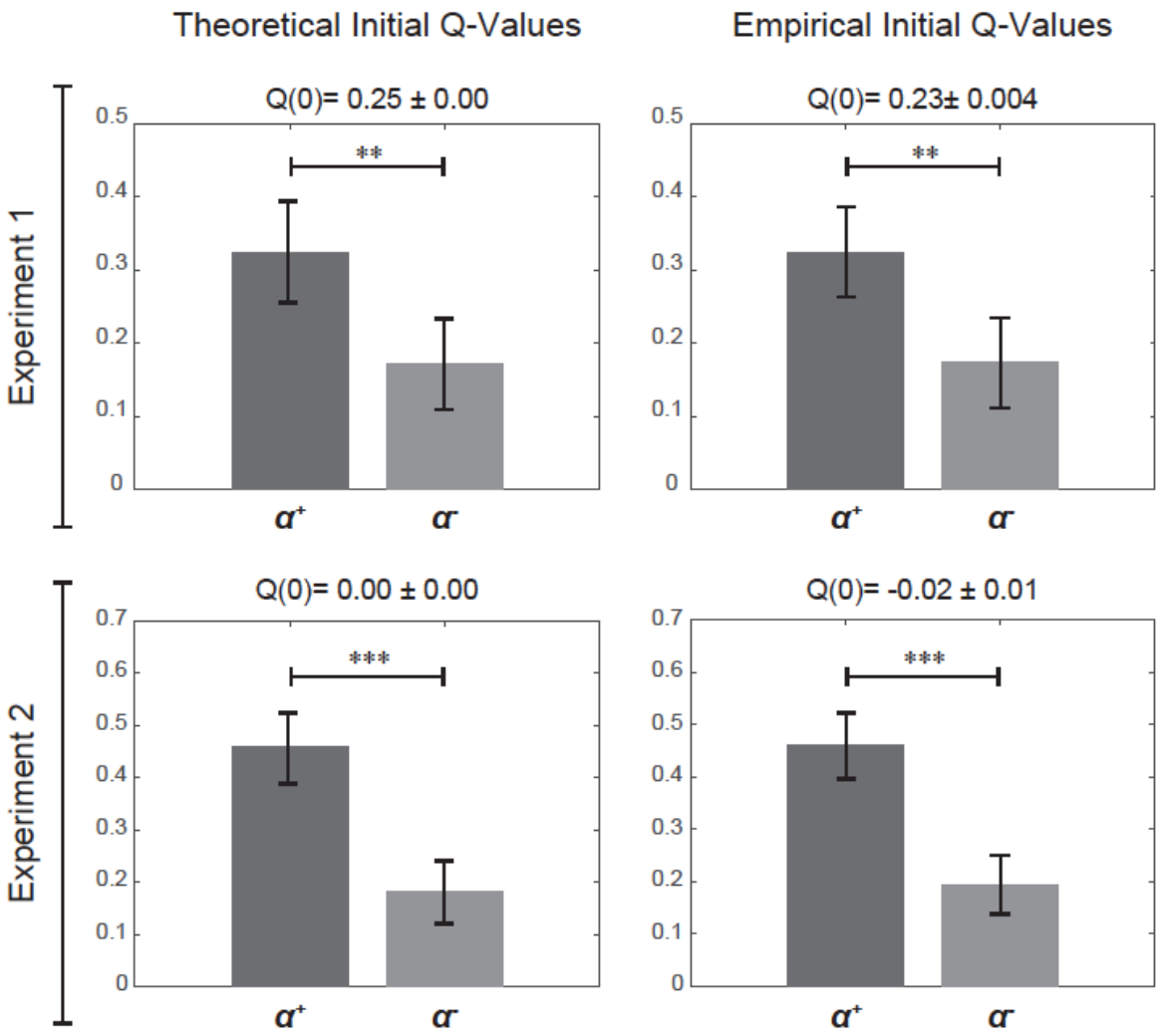

Figure S6: Bars represent for both experiments, learning rates retrieved assuming the initial Q-values equal to the mean between the best and the worst outcome *(leftmost panels)* or assuming the initial Q-values equal to the average outcome per symbol actually experienced during the training session *(rightmost panels)*. ***p<0.001 and **p<0.01, one-sample, two-tailed t-test.

Supplementary Discussion

Supplementary analyses confirm the robustness of our results and the stable nature of optimistic reinforcement learning. Firstly, we fully replicated our behavioral and computational results in a second experiment including reward and actual monetary punishments (**Fig. S3**). We found that the learning asymmetry was robust to different settings and analyses (in all learning phases and in all contingency type; **Fig. S4**). Finally, the results were robust using initial Q-values empirically derived from the training session. This robustness of the optimistic reinforcement learning to a variety of situations corroborates our conclusions, placing the good news/bad news effect on the top of a low reinforcement learning bias.

Supplementary Methods

In order to verify that the parameters optimization procedure did not introduce systematic biases in the parameters' value and to verify that both learning rate asymmetry and the exploitative behavior can be independently detected by our task and model, we run additional model simulations. We simulated four different types of subjects (N=1000 virtual subjects per computational phenotype): RW subjects (symmetric learning rates and higher exploration rate), RW± subjects (asymmetric learning rates and lower exploration rate), RW-exploitative subjects (symmetric learning rates and lower exploration rate), and RW±-explorative subjects (asymmetric learning rates and higher exploration rate). So basically our models simulations included the two computational phenotypes observed in our data plus two additional "hybrid" phenotypes. The results of the parameters optimizations indicated that the computational characteristics of each group were retrieved correctly (**Fig. S7**). Thus, the learning asymmetry and the tendency to exploit have to be considered as two independent features associated with optimistic behavior and not as an artifact of the model optimization procedure.

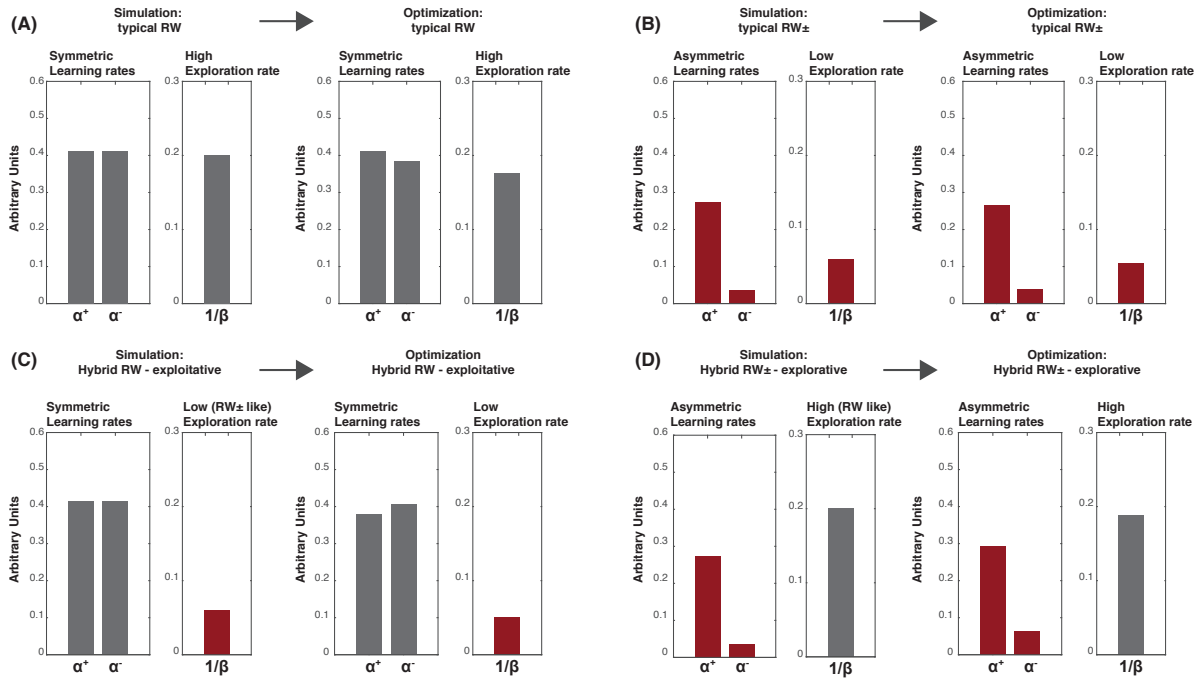

Figure S7: validation of the model optimization procedure
In each panel, the histograms represent the median parameters value used in the simulations (in the leftmost side of each panel: "Simulation") and the parameter retrieved using the same method used for the behavioral data (in the rightmost side of each panel: "Optimization"). (**A**) Typical RW subjects (symmetric learning rates and higher temperature). (**B**) Typical RW± subjects (asymmetric learning rates and lower temperature). (**C**) Hybrid RW-exploitative subjects (symmetric learning rates and lower temperature). (**D**) Hybrid RW±-explorative subjects (asymmetric learning rates and higher temperature).

Supplementary References

1. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10,** 1214–1221 (2007).

2. Vinckier, F. *et al.* Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Mol. Psychiatry* 1–10 (2015). doi:10.1038/mp.2015.73

3. 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).