Information about action outcomes differentially affects learning from self-determined versus imposed choices

Valerian Chambon, Héloïse Théro, Marie Vidal, Henri Vandendriessche, Patrick Haggard, Stefano Palminteri

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0919-5

Information about action outcomes differentially affects learning from self-determined versus imposed choices

Valérian Chambon[1,5]✉, Héloïse Théro[2,5]✉, Marie Vidal[1,3], Henri Vandendriessche[2], Patrick Haggard[2,4] and Stefano Palminteri[2]✉

The valence of new information influences learning rates in humans: good news tends to receive more weight than bad news. We investigated this learning bias in four experiments, by systematically manipulating the source of required action (free versus forced choices), outcome contingencies (low versus high reward) and motor requirements (go versus no-go choices). Analysis of model-estimated learning rates showed that the confirmation bias in learning rates was specific to free choices, but was independent of outcome contingencies. The bias was also unaffected by the motor requirements, thus suggesting that it operates in the representational space of decisions, rather than motoric actions. Finally, model simulations revealed that learning rates estimated from the choice-confirmation model had the effect of maximizing performance across low- and high-reward environments. We therefore suggest that choice-confirmation bias may be adaptive for efficient learning of action–outcome contingencies, above and beyond fostering person-level dispositions such as self-esteem.

Determining whether similar valence-induced biases exist in reinforcement learning and probabilistic reasoning may be crucial to help refine our understanding of adaptive and maladaptive decision-making through the lens of a unified computational approach. Standard reinforcement learning models conceive agents as impartial learners: they learn equally well from positive and negative outcomes alike[1]. However, empirical studies have recently come to challenge this view by demonstrating that human learners, rather than processing information impartially, consistently display a valence-induced bias: when faced with uncertain choice options, they tend to disregard bad news by integrating worse-than-expected outcomes (negative prediction errors) at a lower rate relative to better-than-expected ones (positive prediction errors)[2–4]. This positivity bias would echo the asymmetric processing of self-relevant information in probabilistic reasoning, whereby good news on average receives more weight than bad news[5,6].

A bias for learning preferentially from better-than-expected outcomes would reflect a preference for positive events in general. However, this prediction is at odds with recent findings. In a two-armed bandit task featuring complete feedback information, we previously found that participants would learn preferentially from better-than-expected obtained outcomes while preferentially learning from worse-than-expected forgone outcomes (that is, from the outcome associated with the option they had not chosen[7]). This learning asymmetry suggests that what has been previously characterized as a positivity bias may, in fact, be the upshot of a more general, and perhaps ubiquitous, choice-confirmation bias, whereby human agents preferentially integrate information that confirms their previous decision[8].

Building on these previous findings, we reasoned that if human reinforcement learning is indeed biased in a choice-confirmatory manner, learning from action–outcome couplings that were not voluntarily chosen by the subject (forced choice) should present no bias. To test this hypothesis, we conducted three experiments involving instrumental learning and computational model-based analyses. Participants were administrated new variants of a probabilistic learning task in which they could freely choose between two options, or were 'forced' to implement the choice made by a computer. In the first experiment, participants were only shown the obtained outcome corresponding to their choice (factual learning). In the second experiment, participants were shown both the obtained and the forgone outcome (counterfactual learning). Finally, to address a concern raised during the review process, a third experiment was included in which both free- and forced-choice trials featured a condition with a random reward schedule (50/50). The rationale for implementing this reward schedule was to test whether or not the confirmation bias was due to potential sampling differences between types of trials. Indeed, in the free-choice condition, the most rewarding symbol should be increasingly selected as the subject learns the structure of the task. Having a random reward schedule eliminates the possibility of such unbalanced sampling between free- and forced-choice conditions.

We had two key predictions. With regard to factual learning, participants should learn better from positive prediction error, but they should only do so when free to choose (free-choice trials), while showing no effect when forced to match a computer's choice (forced-choice trials). With regard to counterfactual learning from forgone outcomes, we expected the opposite pattern: in free-choice trials, negative prediction errors should be more likely to be taken

[1]Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, EHESS, CNRS, PSL University, Paris, France. [2]Laboratoire de Neurosciences Cognitives et Computationnelles, Département d'Études Cognitives, École Normale Supérieure, INSERM, PSL University, Paris, France. [3]Institute of Psychiatry and Neuroscience of Paris (IPNP), INSERM U1266, Université de Paris, Paris, France. [4]Institute of Cognitive Neuroscience, University College London, London, UK. [5]These authors contributed equally: Valérian Chambon, Héloïse Théro. ✉e-mail: valerian.chambon@gmail.com; thero.heloise@gmail.com; stefano.palminteri@gmail.com

into account than positive prediction errors, while we expected no bias in forced-choice trials. Put another way, we expected to observe a confirmation bias only when outcomes derived from self-determined choices. Finally, we predicted that including a random reward schedule in both forced- and free-choice conditions would not reduce, nor would it negate, the confirmation bias in self-determined trials.

To verify our predictions, we fitted subjects' behavioural data with several variants of reinforcement learning model, including different learning rates as a function of whether the outcome was positive or negative, obtained or forgone, and followed a free or forced choice. Learning rate analyses were coupled with model comparison analyses aimed at evaluating evidence for the current hypothesis (that is, no confirmation bias in observational learning) and ruling out alternative interpretations of the results (that is, a perseveration bias[9]).

Another central question regarding the nature of the valence-induced bias concerns its relationship with action requirements. An influential theory and related previous findings suggest that positive outcomes favour the learning of choices involving action execution, while negative outcomes favour the learning of choices involving action withdrawal[10,11]. Extending this framework to feedback processing, one should expect the positivity bias to disappear, or even reverse, following trials in which a decision is made by refraining from action. However, if the positivity bias emerges as a consequence of choice confirmation, only making a choice (versus following an instruction) should matter, irrespective of whether this choice is executed through making an action or not. Using a modified version of our design, we tested this prediction in a fourth experiment that varied the requirements of motor execution by including both go and no-go trials. Learning rates were analysed as a function of both outcome valence (negative versus positive) and the requirement for motor execution in order to implement the selected action (key press versus no key press).

Results

Participants performed instrumental learning tasks involving free- and forced-choice trials and go or no-go trials (see Methods and Supplementary Table 1). The task consisted of cumulating as many points as possible by selecting whichever of two symbols was associated with the highest probability of reward. Symbols were always presented in pairs, which comprised one more rewarding and one less rewarding option. In all experiments, each block was associated with a specific pair of symbols, meaning that the participant had to learn from scratch the reward contingencies at the beginning of each block.

In the first three experiments, free-choice trials were interleaved with forced-choice trials. In experiment 4, the computer randomly preselected a symbol, forcing the participant to match the computer's choice (Fig. 1a). Experiment 1 featured partial feedback information, since only the obtained outcome (that is, the outcome of the chosen symbol) was shown (experiment 1 in Fig. 1a; top panel). Experiment 2 featured complete feedback information, since both the obtained and forgone outcomes (that is, the outcome of the unchosen symbols) were shown (experiment 2 in Fig. 1a; bottom panel). As for experiment 1, experiment 3 only featured partial feedback. In contrast with the other experiments, both free- and forced-choice trials of experiment 3 implemented a condition with a random reward schedule (50/50) (experiment 3 in Fig. 1a; top panel). In experiment 4, action requirements were varied within trials where the choice could either be made by performing an action (key press; go trials) or by refraining from acting (no key press; no-go trials) (experiment 4 in Fig. 1b). Note that the nature of the trial (go or no go) was not manipulated by design but depended on the participant's choice (pressing a key or refraining from pressing a key).

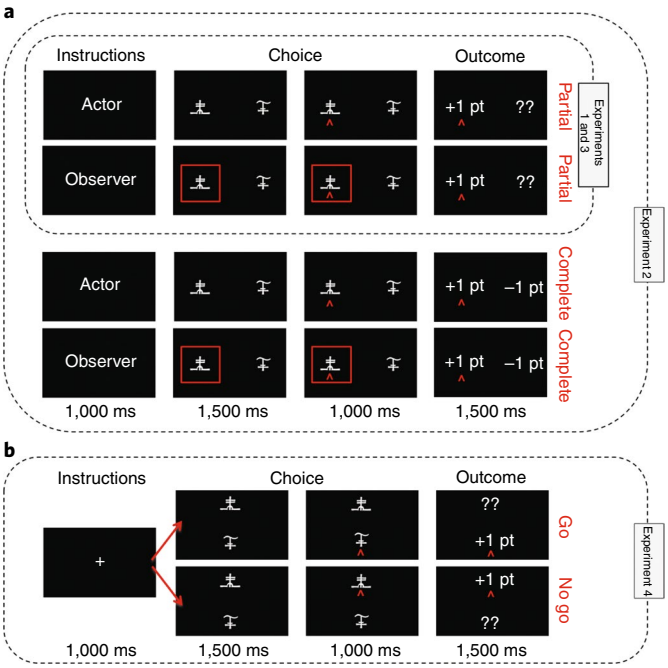

Fig. 1 | Schematic of the trial procedure and stimuli. a, Schematic of the four trial types implemented in experiments 1 and 3 (top) and experiment 2 (bottom). In free-choice trials (actor), participants could freely choose between two options, while in forced-choice trials (observer) participants had to match a preselected option, which was indicated by a red square. In partial trials, participants were only shown the outcome (+1 or −1) associated with the chosen option, while in complete trials participants were shown the outcomes associated with both chosen and unchosen options. Experiment 1 included a condition with only free-choice trials and a condition with intermixed free- and forced-choice trials. Only partial trials were used. In experiment 2, free- and forced-choice trials were intermixed within two conditions: one with partial trials and one with complete trials, where the outcomes of both chosen and unchosen options were shown. Experiment 3 featured a condition with a random reward schedule (50/50) in both free- and forced-choice trials. **b**, Schematic of the two conditions implemented in experiment 4. Action requirements were varied within trials, where the choice of an option could either be made by pressing a key (go trials) or by refraining from pressing any key (no-go trials). This experiment only featured free-choice trials and partial feedback.

Learning performance. To verify that participants understood the task correctly, we analysed correct choice rate (that is, the rate of choosing the most rewarding symbol) in free-choice trials and found it to be significantly higher than the chance level in all four experiments (one-sample, two-tailed *t*-tests against 50%: experiment 1: $t(23)=13.06$; $P<0.001$; $d=2.67$; 95% confidence interval (CI) = 1.8–3.5; experiment 2: $t(23)=24.17$; $P<0.001$; $d=4.85$; 95% CI = 3.3–6.2; experiment 3 (for the 70/30 condition only): $t(29)=10.77$; $P<0.001$; $d=2$; 95% CI = 1.3–2.6; experiment 4 (for the 70/30 condition only): $t(19)=9.38$; $P<0.001$; $d=2.09$; 95% CI = 1.29–2.87). To assess learning dynamics, we also verified that learning performance was higher in the second half of the learning block relative to the first (see Fig. 2a and Supplementary Fig. 2 for full learning trajectories). Note that in conditions mixing free- and forced-choice trials, a modest boost in performance can be observed in free-choice trials (free + forced conditions; Fig. 2a). This boost in performance indicates that participants also learned from forced-choice trials. This was expected, as in forced-choice trials options featured the same outcome contingencies as in free-choice trials.

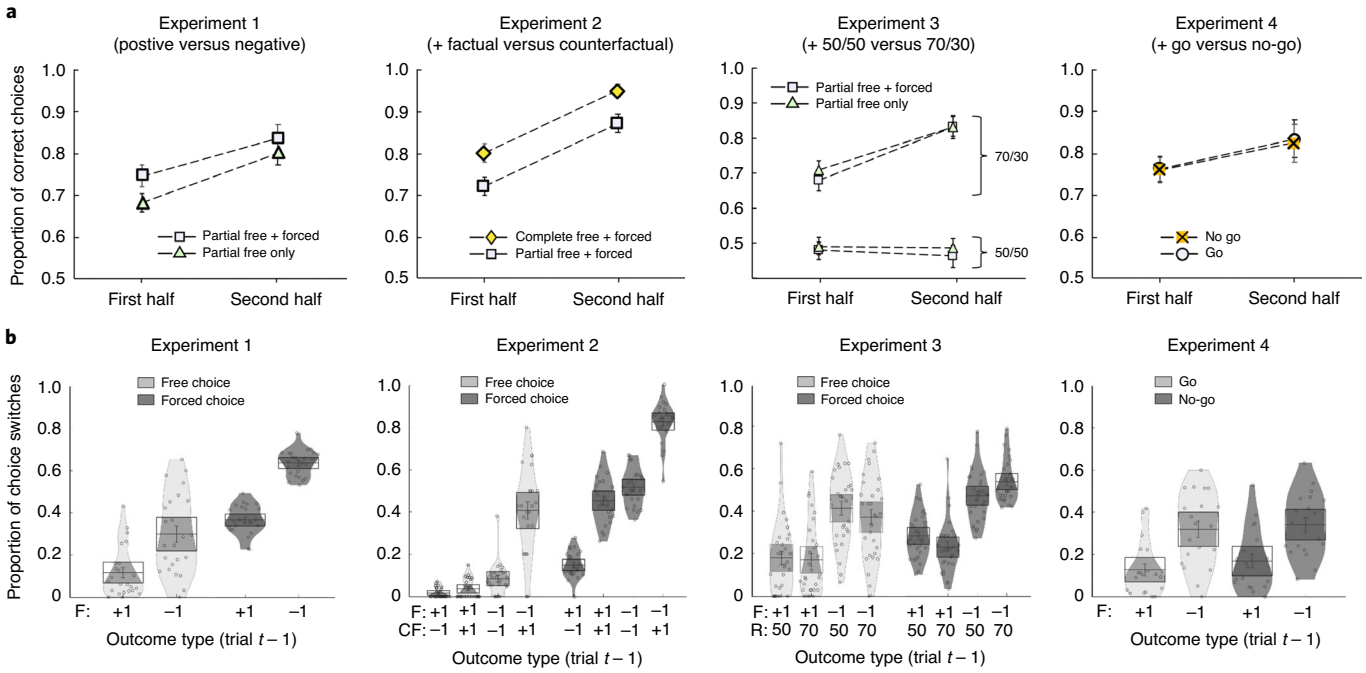

Fig. 2 | Behavioural results. a, Mean proportion of correct choices in the first and second halves of each learning block for: the two conditions of experiment 1 (free only, and intermixed free and forced) ($n=24$); the two conditions of experiment 2 (partial and complete) ($n=24$); the two reward schedules used in experiment 3 (70/30 and 50/50) ($n=30$); and go and no-go trials in experiment 4 ($n=20$). **b**, Proportion of choice switches between trial t and $t-1$ as a function of the obtained outcome (factual (F)) and the forgone (counterfactual (CF)) outcome seen on trial $t-1$, depending on whether this trial was a free- or a forced-choice trial (experiments 1–3) or a go or no-go trial (experiment 4), and depending on the reward schedule (R) being implemented (50/50 or 70/30; experiment 3). For experiments 1 and 2, the analysis was made on partial intermixed trials. For experiment 2, the analysis was made on complete intermixed trials, which contained both obtained and forgone outcomes. The free + forced data represent the correct choice rate in free-choice trials only (by definition, the correct choice rate in forced-choice trials is bounded to be 50) within blocks where both free- and forced-choice trials were mixed. For each learning rate, individual data points are displayed within an area representing their probability density function. Means ± s.e.m. are shown within a box whose height corresponds to the 95% CI.

Switching choice after a negative outcome, and repeating a choice after receiving a positive outcome, is a hallmark of feedback-based adaptive behaviour. To verify that participants took into account both free- and forced-choice outcomes, we analysed the switch rate as a function of switches depending on: (1) whether the previous obtained outcome was positive or negative; but also (2) whether the previous trial was a free- or forced-choice trial (experiments 1–3) or a go or no-go trial (experiment 4); as well as on (3) whether the forgone outcome was positive or negative (experiment 2 only); and on (4) the reward schedule being implemented (50/50 versus 70/30; experiment 3 only). Data distribution was assumed to be normal but this was not formally tested.

The repeated-measures analyses of variance (ANOVAs) revealed a main effect of the obtained outcome on switch choices in all experiments (experiment 1: $F(1,23)=131.47$; $P<0.001$; $\eta^2=0.85$; 90% CI = 0.72–0.89; experiment 2: $F(1,23)=252.34$; $P<0.001$; $\eta^2=0.91$; 90% CI = 0.84–0.94; experiment 3: $F(1,29)=161.04$; $P<0.001$; $\eta^2=0.84$; 90% CI = 0.74–0.89; experiment 4: $F(1,19)=42.74$; $P<0.001$; $\eta^2=0.69$; 90% CI = 0.44–0.79). Thus, as expected, participants switched options more often after receiving a negative, relative to a positive, outcome. This effect was observed after both a free- and a forced-choice trial alike, and in both go and no-go trials. As expected, we also found a main effect of the forgone outcome in experiment 2 ($F(1,23)=364.35$; $P<0.001$; $\eta^2=0.94$; 90% CI = 0.88–0.95), with participants switching choices significantly more when the outcome associated with the unchosen option was positive, relative to negative (see Fig. 2b). The interaction effects between the forgone outcome and the type of choice ($F(1,23)=22.52$; $P<0.001$; $\eta^2=0.49$; 90% CI = 0.22–0.64) or between the forgone outcome and the type of outcome obtained ($F(1,23)=33.27$; $P<0.001$; $\eta^2=0.59$; 90% CI = 0.33–0.71) were all statistically significant, as was the three-way interaction effect ($F(1,23)=27.15$; $P<0.001$; $\eta^2=0.54$; 90% CI = 0.27–0.67). In experiment 3, neither the main effect of the reward schedule (50/50 versus 70/30; $F(1,29)=0.49$; $P=0.48$; $\eta^2=0.01$; 90% CI = 0–0.15) nor the schedule-by-choice ($F(1,29)=2.56$; $P=0.12$; $\eta^2=0.08$; 90% CI = 0–0.25) or the schedule-by-outcome interaction ($F(1,29)=3.71$; $P=0.06$; $\eta^2=0.11$; 90% CI = 0.71–0.89) effects were statistically significant. A significant three-way interaction effect was found ($F(1,29)=20.81$; $P<0.001$; $\eta^2=0.41$; 90% CI = 0.18–0.57).

In the first three experiments, the main effect of the type of choice was significant (experiment 1: $F(1,23)=85,67$; $P<0.001$; $\eta^2=0.78$; 90% CI = 0.62–0.85; experiment 2: $F(1,23)=439.65$; $P<0.001$; $\eta^2=0.95$; 90% CI = 0.90–0.96; experiment 3: $F(1,29)=28.41$; $P<0.001$; $\eta^2=0.49$; 90% CI = 0.25–0.63), with participants switching more often after a forced-choice trial than after a free-choice trial. This effect can be accounted for by the fact that the chosen symbol was pseudo-randomly selected in forced-choice trials, while subjects preferentially chose the correct option in free-choice trials (thus forced-choice trials were more likely to involve incorrect choices). In experiments 1 and 2, the outcome-by-choice interaction effect was statistically significant (experiment 1: $F(1,23)=8,81$; $P=0.006$; $\eta^2=0.27$; 90% CI = 0.04–0.47; experiment 2: $F(1,23)=33.51$; $P<0.001$; $\eta^2=0.59$; 90% CI = 0.34–0.71). In experiment 4, neither the main effect of the execution mode (key press versus no key press: $F(1,19)=0.42$; $P=0.52$; $\eta^2=0.02$; 90%

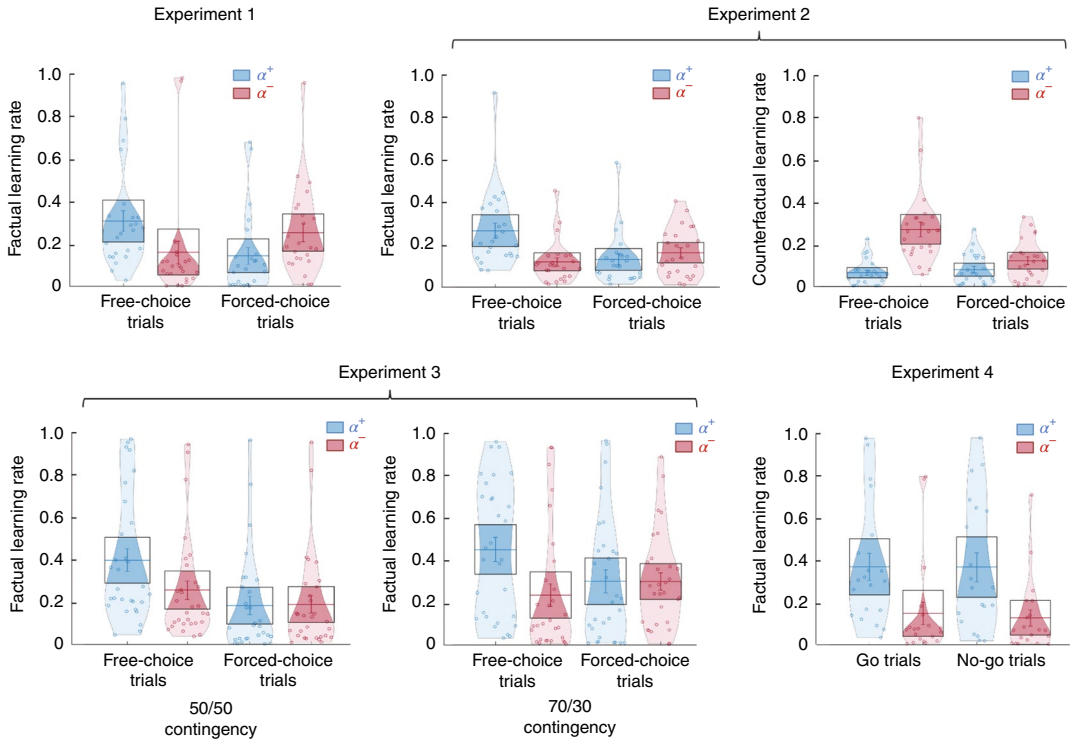

Fig. 3 | Parameter results of the full model from all four experiments. Top left: fitted factual learning rates from free- and forced-choice trials of experiment 1 ($n=24$). Top middle and top right: fitted factual and counterfactual learning rates from free- and forced-choice trials of experiment 2 ($n=24$). Bottom left and bottom middle: fitted factual learning rates from free- and forced-choice trials, and from 50/50 and 70/30 reward-schedule conditions of experiment 3 ($n=30$). Bottom right: fitted learning rates from go and no-go trials of experiment 4 ($n=20$). Note that only obtained outcomes are shown for experiments 1, 3 and 4, whereas both obtained and forgone outcomes are displayed for experiment 2, which allowed for fitting counterfactual learning rates. Positive (α^+) and negative (α^-) learning rates are represented in blue and red, respectively. For each learning rate, individual data points are displayed within an area representing their probability density function. Means ± s.e.m. are shown within a box whose height corresponds to the 95% CI.

CI = 0–0.19) nor the valence-by-execution interaction effect was significant ($F(1,19)=0.66$; $P=0.42$; $\eta^2=0.03$; 90% CI = 0–0.22).

Model parameter analyses. To test the influence of outcome valence and choice type on learning, we fitted the data with a modified Rescorla–Wagner model assuming different learning rates for positive and negative outcomes (α^+ and α^-, respectively) and for free- and forced-choice trials (experiments 1–3), or for go and no-go trials (experiment 4). In experiment 2, different learning rates were also assumed for obtained and forgone outcomes (Fig. 3, top right), whereas in experiment 3 different learning rates were fitted for the 50/50 and 70/30 reward schedules (Fig. 3, bottom left). We refer to these models as full models, because they present the highest number of parameters in the considered model space. In experiment 1, the resulting learning rates were subjected to a 2×2 repeated-measures ANOVA with outcome valence (positive versus negative) and choice type (free versus forced) as within-subject factors. In experiment 2, learning rates were subjected to a 2×2×2 repeated-measures ANOVA with outcome valence (positive versus negative), choice type (free versus forced) and outcome type (obtained versus forgone) as within-subject factors. In experiment 3, learning rates were subjected to a 2×2×2 repeated-measures ANOVA with outcome valence (positive versus negative), choice type (free versus forced) and reward schedule (50/50 versus 70/30) as within-subject factors. Finally, in experiment 4, learning rates were subjected to a 2×2 repeated-measures ANOVA with outcome valence (positive versus negative) and execution mode (key press versus no key press) as within-subject factors (Fig. 3 and Supplementary Table 2).

In experiment 1, no main effect was significant (choice type: $F(1,23)=1.59$; $P=0.21$; $\eta^2=0.06$; 90% CI = 0–0.25; outcome valence: $F(1,23)=0.25$; $P=0.61$; $\eta^2=0.01$; 90% CI = 0–0.15). In this experiment, the valence-by-choice interaction was significant ($F(1,23)=9.11$; $P=0.006$; $\eta^2=0.28$; 90% CI = 0.05–0.47).

In experiment 2, the main effects of choice ($F(1,23)=10.59$; $P=0.003$; $\eta^2=0.31$; 90% CI = 0.07–0.50) and outcome type ($F(1,23)=4.37$; $P=0.047$; $\eta^2=0.15$; 90% CI = 0–0.36) were significant, whereas the main effect of valence was not ($F(1,23)=3.6$; $P=0.07$; $\eta^2=0.13$; 90% CI = 0–0.34). We found a significant valence-by-outcome type ($F(1,23)=58.45$; $P<0.001$; $\eta^2=0.71$; 90% CI = 0.51–0.80) and a significant valence-by-choice-by-outcome type ($F(1,23)=36.58$; $P<0.001$; $\eta^2=0.61$; 90% CI = 0.36–0.72) interaction in this experiment. Neither the valence-by-choice ($F(1,23)=0.05$; $P=0.81$; $\eta^2=0.002$; 90% CI = 0–0.09) nor the choice-by-outcome type ($F(1,23)=0.45$; $P=0.50$; $\eta^2=0.02$; 90% CI = 0–0.17) interaction effects were statistically significant.

In experiment 3, the main effects of valence, choice and reward schedule were significant (valence: $F(1,29)=6.06$; $P=0.019$; $\eta^2=0.17$; 90% CI = 0.1–0.36; choice: $F(1,29)=7.25$; $P=0.011$; $\eta^2=0.20$; 90% CI = 0.02–0.38; reward schedule = $F(1,29)=9.87$; $P=0.003$; $\eta^2=0.25$; 90% CI = 0.05–0.43). In this experiment, the valence-by-choice ($F(1,29)=7.73$; $P=0.009$; $\eta^2=0.21$; 90% CI = 0.03–0.39) and schedule-by-choice interactions ($F(1,29)=6.18$; $P=0.018$; $\eta^2=0.17$; 90% CI = 0.16–0.36) were also significant. Neither the schedule-by-valence ($F(1,29)=0.47$; $P=0.49$; $\eta^2=0.01$; 90% CI = 0–0.14) nor the schedule-by-valence-by-choice interactions ($F(1,29)=0.44$; $P=0.51$; $\eta^2=0.01$; 90% CI = 0–0.14) were significant.

In experiment 4, only the main effect of outcome valence was statistically significant ($F(1,19) = 25.54$; $P < 0.001$; $\eta^2 = 0.57$; 90% CI = 0.28–0.70), whereas the main effect of the execution mode was not ($F(1,19) = 0.04$; $P = 0.83$; $\eta^2 = 0.002$; 90% CI = 0–0.01]). We found no evidence for a statistically significant valence-by-execution interaction effect in this experiment ($F(1,19) = 0.03$; $P = 0.85$; $\eta^2 = 0.001$; 90% CI = 0–0.08).

We performed post-hoc *t*-tests to further investigate the significant valence-by-choice interactions found in the first three experiments. The difference between positive and negative learning rates was statistically significant in free-choice trials of experiments 1 and 2 (obtained outcomes in experiment 1: $t(23) = 2.5$; $P = 0.02$; $d = 0.5$; 95% CI = 0.09–1.08; obtained outcomes in experiment 2: $t(23) = 4.1$; $P < 0.001$; $d = 0.84$; 95% CI = 0.44–1.58; forgone outcomes in experiment 2: $t(23) = -6.2$; $P < 0.001$; $d = 1.4$; 95% CI = 1.02–2.49), but the difference was not statistically significant in experiment 3 (50/50 schedule: $t(29) = 2.6$; $P = 0.14$; $d = 0.47$; 95% CI = 0.10–0.94; 70/30 schedule: $t(29) = 2.59$; $P = 0.14$; $d = 0.49$; 95% CI = 0.16–1.29). In forced-choice trials of all three experiments, we found no evidence for a statistically significant difference between positive and negative learning rates (obtained in experiment 1: $t(23) = -2.0$; $P = 0.055$; obtained in experiment 2: $t(23) = -1.3$; $P = 0.20$; forgone in experiment 2: $t(23) = -1.5$; $P = 0.14$; 50/50 schedule in experiment 3: $t(29) = -0.14$; $P = 0.88$; 70/30 schedule in experiment 3: $t(29) = 0.01$; $P = 0.98$; see Fig. 3).

To sum up, we replicated that participants learned preferentially from positive compared with negative prediction errors, whereas the opposite was true for forgone outcomes[7]. We found that this learning asymmetry was significant only in free-choice trials, and was undetectable when participants were forced to match the computer's decision. Implementing a random reward schedule (schedule-by-valence-by-choice interaction in experiment 3) or varying action requirements across go and no-go trials (valence-by-execution interaction in experiment 4) had no statistically discernible effect on the learning asymmetry.

Parsimony-driven parameter reduction. Although we found no valence-induced bias in forced-choice learning rates on average, one cannot rule out that participants had opposite significant biases (for example, some would learn better from positive forced-choice outcomes, while others would learn better from negative forced-choice outcomes). We therefore ran a parsimony-driven parameter reduction to assess whether fitting different learning rates in (1) forced-choice trials and (2) go and no-go trials better predicted participants' data (see Fig. 4a). The full models (with four or eight learning rates, depending on the experiment; see below) were compared with reduced versions including either a valence-induced bias only for free-choice outcomes or no bias at all. In experiment 1, the full model had four learning rates (two learning rates following positive and negative prediction errors ($\alpha_{positive}$ and $\alpha_{negative}$) and two learning rates in forced- and free-choice trials (α_{forced} and α_{free}). Similarly, in experiment 4, the model included two learning rates following positive and negative prediction errors ($\alpha_{positive}$ and $\alpha_{negative}$) and two learning rates in go and no-go trials (α_{go} and $\alpha_{no\,go}$)). In experiment 2, the full model had eight learning rates ($\alpha_{positive}$, $\alpha_{negative}$, α_{forced} and α_{free} for both factual and counterfactual learning). Similarly, in experiment 3, there were eight learning rates ($\alpha_{positive}$, $\alpha_{negative}$, α_{forced} and α_{free} for both 50/50 and 70/30 reward schedules).

We compared the models using a Bayesian model selection procedure[12] based on the Bayesian information criterion (BIC). In all experiments, intermediate models (that is, models including valence-induced bias only for free-choice outcomes) were found to better account for the data: their average posterior probabilities (PPs) were higher than the PPs of the other models in the set (experiment 1: 2α PP = 0.13 ± 0.004; 3α PP = 0.51 ± 0.009; 4α PP = 0.35 ± 0.008; experiment 2: 4α PP = 0.15 ± 0.004; 6α PP = 0.73 ± 0.007; 8α PP = 0.10 ± 0.003; experiment 3: 6α PP = 0.03 ± 0.001; 3α PP = 0.90 ± 0.002; 8α PP = 0.06 ± 0.001; experiment 4: 4α PP = 0.20 ± 0.006; 2α positive–negative PP = 0.73 ± 0.008; 2α go–no-go PP = 0.06 ± 0.002; see Fig. 4b).

Consistent with a previous study[7], we further found that experiment 2 data were better explained by an even more parsimonious model assuming similar positive factual and negative counterfactual learning rates ($\alpha_F^+ = \alpha_C^- = \alpha_{conf}$), and similar negative factual and positive counterfactual learning rates ($\alpha_F^- = \alpha_C^+ = \alpha_{disc}$). This final model thus had three different learning rates: α_{conf}, α_{disc} and α_{forced} (3α PP = 0.92 ± 0.002; 6α PP = 0.07 ± 0.002; see Fig. 5). We refer to these learning rates as α_{conf} and α_{disc} because they embody learning from confirmatory (positive obtained and negative forgone) and disconfirmatory (negative obtained and positive forgone) outcomes, respectively.

Ruling out the perseveration bias. In previous studies, a heightened choice hysteresis (that is, an increased tendency to repeat a choice above and beyond outcome-related evidence) has been identified as a behavioural hallmark of positivity and confirmatory learning biases[2,7]. However, the same behavioural phenomenon may arise in the presence of a learning-independent choice-repetition bias (often referred to as perseveration bias[13]), which is not to be confounded with motor inertia (which was avoided in our tasks by counterbalancing the spatial position of the cues). Even more concerning, positivity and confirmation biases may spuriously arise when fitting multiple learning rates on data presenting a simple choice-repetition bias[9]. To rule out this possibility, we explicitly compared models including positivity (experiment 1: α_{free}^+, α_{free}^- and α_{forced}) and confirmation learning biases (experiment 2: α_{conf}, α_{disc} and α_{forced}) with a model including a perseveration parameter. The models with different learning rates (3α) were found to better account for the data compared with the perseveration model, with a higher average PP (experiment 1: perseveration model PP = 0.39 ± 0.008; 3α PP = 0.60 ± 0.008; experiment 2: perseveration model PP = 0.10 ± 0.003; 3α PP = 0.89 ± 0.003; experiment 3: perseveration model PP = 0.41 ± 0.007; 3α PP = 0.59 ± 0.007) (Fig. 6a).

To further quantify the extent to which the observed learning biases could be ascribed to the observed choice-repetition bias, we simulated the perseveration model (using its best-fitting parameters) and fitted the models with multiple learning rates on these synthetic data (Fig. 6b). While model parameter analyses confirmed that positivity and confirmation biases may spuriously arise from data featuring a perseveration bias, the biases retrieved from the simulations were nonetheless significantly smaller compared with those observed in the participants' data (experiment 1, comparing the bias (that is, α_{free}^+ minus α_{free}^-) between participants and the perseveration model: $t(46) = 2.23$; $P = 0.03$; $d = 0.64$; 95% CI = 0.05–1.22; experiment 2: $t(46) = 5.97$; $P < 0.001$; $d = 1.77$; 95% CI = 1.12–2.47; experiment 3: $t(58) = 2.08$; $P = 0.04$; $d = 0.53$; 95% CI = 0.02–1.05).

Parameter adaptation to task contingency. In the first two experiments, we manipulated reward contingencies to include a low-reward (reward probabilities set to 0.4 and 0.1) and a high-reward condition (reward probabilities set to 0.9 and 0.6). This manipulation was included to first assess whether learning rates were adaptively modulated as a function of the amount of reward available in the task environment (low versus high), and second to test whether this modulation extended to forced-choice outcomes. Previous optimality analyses suggested that a positivity bias would be advantageous in low-reward conditions, while the opposite would be true in high-reward conditions[14]. In other terms, it would be optimal to exhibit a higher learning rate for rare outcomes (that is, rewards in low-reward conditions and punishments in high-reward conditions). In a new computational analysis, we fitted different learning rates for high- and low-reward conditions,

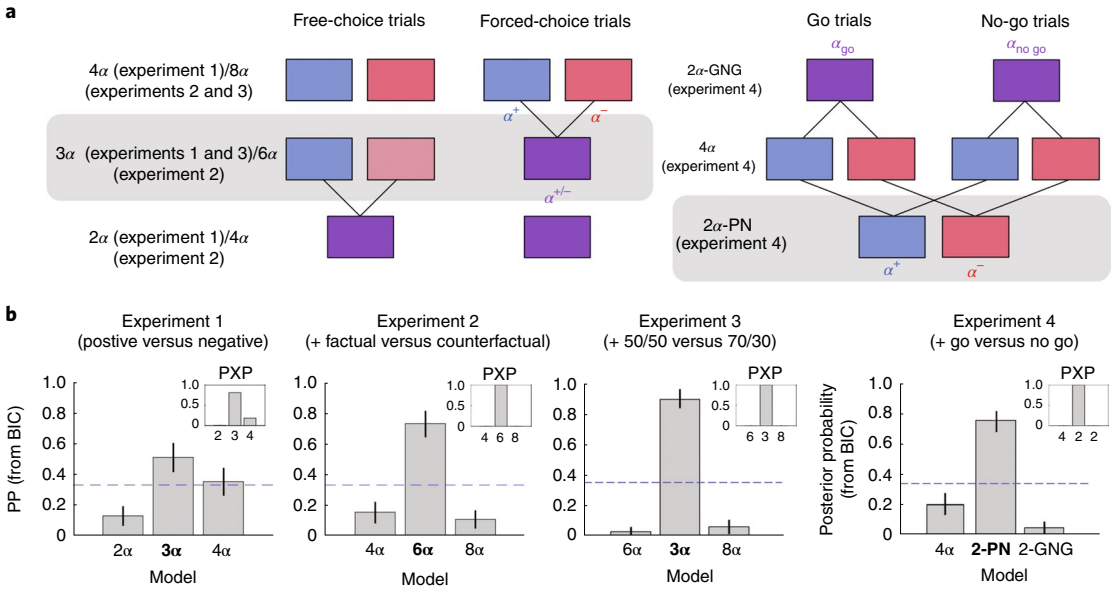

Fig. 4 | Model comparison results from all four experiments. a, Representation of the model space. Top left and top middle: in experiments 1–3, the full model (4α (experiment 1) and 8α (experiments 2 and 3)) has different learning rates for positive and negative prediction errors (blue and red squares, respectively) and for free- and forced-choice trials. The intermediate model (3α and 6α; grey background) has only a single learning rate on forced-choice trials, whereas the reduced model (2α and 4α) does not split learning rates by valence at all. The reduced model is thus nested within the intermediate model, which is itself nested within the full model. Note that in experiment 2 (n = 24), the parameter reduction operates for both factual and counterfactual learning rates, whereas in experiment 3 (n = 30) the reduction operates for both 50/50 and 70/30 reward-schedule learning rates. The grey rectangle signals the winning model in each experiment. Top right: in experiment 4 (n = 20), the full model (4α) has different learning rates for positive and negative prediction errors (blue and red squares) and for go and no-go trials. Two reduced models were tested: (1) a model with two different learning rates for positive and negative outcomes (2α-PN); and (2) a model with two different learning rates for go and no-go trials (2α-GNG). **b**, Expectations and variance of the PP for each model, based on the BIC values, with the PXP for each model shown as an insert for each experiment. The winning model is indicated in bold.

thus creating new models with six learning rates (3 (learning rate types) × 2 (low versus high)). We subjected the resulting parameter values to a 3 (learning rate types) × 2 (high- versus low-reward conditions) repeated-measures ANOVA.

As expected, the learning rate type had a significant main effect (experiment 1: $F(2,46) = 9.63$; $P < 0.001$; $\eta^2 = 0.29$; 90% CI = 0.10–0.43; experiment 2: $F(2,46) = 36.36$; $P < 0.001$; $\eta^2 = 0.61$; 90% CI = 0.44–0.69) (Fig. 7a). There was no evidence for a statistically significant main effect of the condition factor in both experiments (experiment 1: $F(1,23) = 1.29$; $P = 0.26$; $\eta^2 = 0.05$; 90% CI = 0–0.23; experiment 2: $F(1,23) = 0.35$; $P = 0.55$; $\eta^2 = 0.01$; 90% CI = 0–0.16). Regarding the condition-by-type interaction, the effect was equivocal: it was significant in experiment 1, but not significant in experiment 2 (experiment 1: $F(2,46) = 11.14$; $P < 0.001$; $\eta^2 = 0.32$; 90% CI = 0.13–0.45; experiment 2: $F(2,46) = 0.27$; $P = 0.76$; $\eta^2 = 0.01$; 90% CI = 0–0.06). Given the inconclusive nature of the above effects, we further tested learning rate adaptation in our data by turning to model comparison. The models with different learning rates for high- and low-reward conditions (H&L) were compared with models without learning rate modulation (pooled models). In both experiments, we found that the model without contingency-dependent learning rates had the highest exceedance probability (experiment 1: pooled PP = 0.78 ± 0.006; H&L PP = 0.21 ± 0.006; experiment 2: pooled PP = 0.95 ± 0.001; H&L PP = 0.04 ± 0.001) (see Fig. 7b).

Finally, using model simulations, we assessed how the observed pattern of learning rates compares to other patterns with respect to task performance (mean accuracy and variance) across low- and high-reward conditions. To do so, we simulated models with different learning rate patterns on 1,000 datasets for each participant. We set learning rates to be either choice confirmatory (CO), valence neutral (NT) or choice disconfirmatory (CD), and the learning rate patterns could be different in free-choice and forced-choice trials (see Fig. 7c and Supplementary Table 3). Replicating and extending the findings of Cazé and van der Meer[14], we found the choice-confirmatory models to outperform the other models in low-reward conditions, and the choice-disconfirmatory models to have better performances in the high-reward conditions (Fig. 7c).

When we looked at the general performance across both conditions, we found that the model corresponding to the participants' learning rate patterns (that is, the CO&NT model, whose learning rates were choice confirmatory in free choices and valence neutral in forced choices, was among the highest-performing models (Fig. 7c)). In experiment 1, the CO&NT model had a performance of 83.2%, while the CO&CD model had a slightly higher performance of 84.2%. In experiment 2, the CO&NT model had a performance of 86.6% while the CO&CD model had a slightly lower performance of 85.7%. Therefore, the learning rate patterns found in our participants can be said to be optimal, or close to optimal, in the set-up of our task and within the considered range of model parameters.

Interestingly, the performances of the CO&NT model were also quite similar across the high- and low-reward conditions. In experiment 1, the difference in performances between high- and low-reward conditions was 2% for the CO&NT model and 1.9% for the CO&CD model, while this difference was over 3% for the other models. In experiment 2, the difference in performances was the smallest for the CO&NT model (0.4%; versus 0.8% for the CO&CD model and 0.5% for the NT&CO model). Not only was the participants' best-fitting pattern very advantageous in terms of accuracy, but it also exhibited highly stable performances across low- and high-reward conditions. Performances across both conditions were not statistically discernible (paired, two-tailed t-tests comparing performances between low- and high-reward

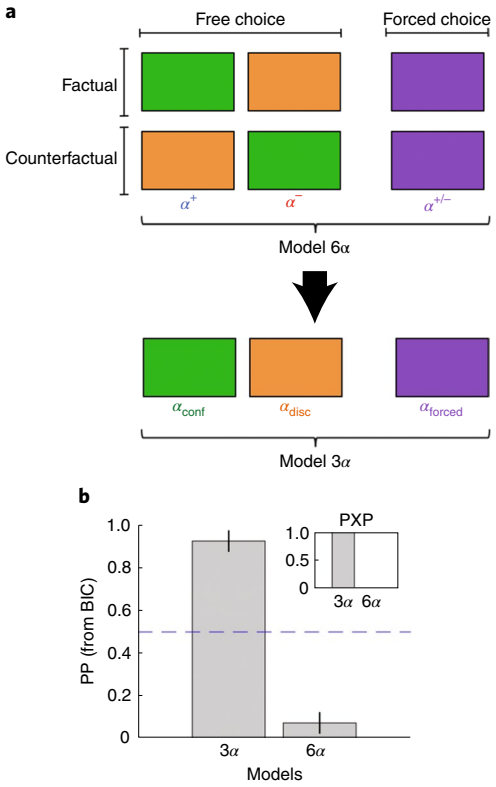

Fig. 5 | Model comparison from experiment 2. a, Representation of the model space. The model with six learning rates (top) corresponds to the winning model presented to the left in Fig. 4a. The black arrow indicates that this winning model can be reduced to a model with three learning rates, which embody learning in forced-choice trials (α_{forced}) and learning from confirmatory (α_{conf}) and disconfirmatory (α_{disc}) outcomes in free-choice trials. **b**, Expectation and variance of the PP for the initial winning model (6α) and the reduced model (3α), based on the BIC values. The insert shows the PXP for each model.

environments: experiment 1: $t(23) = 0.25$; $P = 0.80$; experiment 2: $t(23) = -0.027$; $P = 0.98$).

Discussion

While standard accounts of belief (or value) prescribe that an agent should learn equally well from positive and negative information[1,15], previous studies have consistently shown that people exhibit valence-induced biases[2,6]. These biases are traditionally exemplified in the so-called good news/bad news effect, whereby people tend to overweight good news relative to bad news when updating self-relevant beliefs[5]. An important moderator of these self-related biases would be the extent to which an individual believes they are able to control (that is, choose) the dimension concerned. Thus, it has been shown that individuals tend to rate themselves as above average on positive controllable (but not uncontrollable) traits[16]. Likewise, people self-attribute positive outcomes when the perceived controllability of the environment is high[17], and enact different behaviours[18] or process behavioural consequences differently[19] when these consequences are under their direct control, relative to uncontrollable self-relevant outcomes (for example, an asthma attack). In the present study, we sought to investigate further the link between outcome processing and control (that is, voluntary choice) in four instrumental learning experiments comparing trials with and without voluntary choices (experiments 1 and 2), featuring factual and counterfactual learning (experiment 2), and presenting different reward contingencies (experiment 3) and distinct action requirements (experiment 4).

In the first experiment, learning performance was compared between trials in which the subject could either freely choose which option to select, or was forced to match a computer's choice. As predicted, we found that participants learned better and faster from positive (relative to negative) prediction errors (that is, from better-than-expected outcomes). Crucially, this learning asymmetry (positive > negative prediction errors) was present when participants were free to choose between options, but absent (if not reversed; negative > positive) when participants were forced to match an external choice. In other terms, we observed a positivity bias (that is, a learning bias in favour of positively valued outcomes) only when learning was driven by self-determined choices.

In the second experiment, we combined free- and forced-choice trials with learning from factual (chosen action) or counterfactual (unchosen action) outcomes. Replicating previous results[7], we observed that prediction error valence biased factual and counterfactual learning in opposite directions: when learning from obtained outcomes (chosen action), positive prediction errors were preferentially taken into account compared with negative prediction errors. In contrast, when learning from forgone outcomes (unchosen action), participants integrated equally well positive and negative prediction errors. In other words, only positive outcomes that supported the participant's current choice (positive outcomes associated with the chosen option; that is, factual outcomes) were preferentially taken into account, whereas positive outcomes that contradicted this choice (notably, positive outcomes associated with the unchosen option; that is, counterfactual outcomes) were discounted. Experiment 3 further confirmed that this learning asymmetry was not driven by potential differences in outcome sampling between free- and forced-choice trials, as the learning bias was also present in the 50/50 condition. Taken together, these findings suggest that the well-documented positivity bias may be a special case of a more general choice-confirmation bias, whereby individuals tend to overweight positive information when it confirms their previous choice. In contrast, when no choice is involved, positive and negative information are weighted equally (Fig. 8).

Importantly, if learning asymmetry reflects a choice-confirmation bias, it should arise from the very first stage of the action processing chain (that is, at the decision stage rather than at the action stage). Thus, a pure choice-supportive bias should be oblivious to how the choice is implemented (for example, either through performing an action or refraining from acting). In contrast, a cognitive dissonance account would state that making an action is, in and of itself, sufficient to induce a learning asymmetry. Indeed, in the absence of any previous intention or reason to act, the mere fact of producing an action would strongly commit the agent with regard to the outcome. This commitment would then be retrospectively justified by shaping the individual's preferences in such a way that they align post hoc with subsequent action outcomes (post-action dissonance reduction[20]). Critically, most of the protocols confound choice and action execution, and hence are not well suited to disentangling the influence of choice and action execution on valence-dependent learning asymmetries. In the fourth experiment, we directly addressed this issue by varying action requirements across go and no-go trials. Learning rates were analysed as a function of both outcome valence (negative versus positive) and execution mode (go versus no go). We replicated learning asymmetries found in free-choice trials of the first three experiments, with positive prediction errors being taken into account more than negative prediction errors. We found no credible evidence of a difference between trials where the response was made by performing an action (key press) or by refraining from acting (no key press). Both dimensionality reduction and model comparison procedures supported these results. Thus, the choice-confirmation bias is truly related to choices, rather than to the physical motor events that implement those choices.

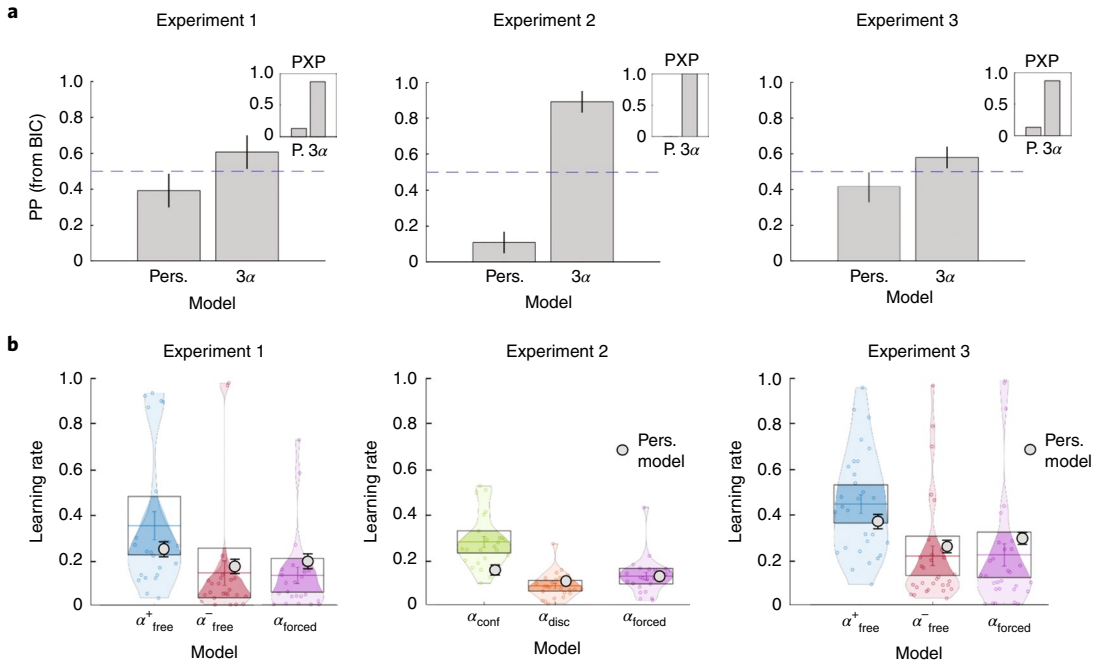

Fig. 6 | Comparison of the winning model (3α) with a model with a simple perseveration parameter. a, Expectation and variance of the PP for the perseveration model (Pers. in the main graphs; P. in the insets) and the winning model with three learning rates, based on the BIC values. The insert charts show the PXP for each model. **b**, The winning model was fitted on data from experiment 1 (left; $n = 24$), experiment 2 (middle; $n = 24$) and experiment 3 (right; $n = 30$). The violin plots represent the learning rates fitted on the participants' data, while the grey dots represent the learning rates fitted to artificial datasets, created by simulating the preservation model ten times on each participant's data. Positivity and confirmation biases may spuriously arise from data featuring a perseveration bias, but the biases retrieved from the simulations (grey dots) were significantly smaller compared with those observed in the participants' data (bars) (experiment 1, comparing the bias (that is, α^{+}_{free} minus α^{-}_{free}) between participants and the perseveration model: $t(46) = 2.23$; $P = 0.03$; $d = 0.64$; 95% CI = 0.05–1.22; experiment 2: $t(46) = 5.97$, $P < 0.001$, $d = 1.77$, 95% CI = 1.12–2.47; experiment 3: $t(58) = 2.08$, $P = 0.04$, $d = 0.53$, 95% CI = 0.02–1.05). Error bars indicate s.e.m.

In the present study, the asymmetric valuation of gains and losses is accounted for by fitting different learning rates for positive and negative outcomes. Another way of modelling this asymmetric treatment of gains and losses would be to assume that the subjective value of a given loss is, in absolute terms, greater than a comparable gain (a phenomenon that is commonly referred to as loss aversion). To directly test this hypothesis, we defined new models endowed with such a subjective loss parameter, and we adapted the structure of these models to each of our four experiments. The results fully replicated those obtained with asymmetric learning rates for positive and negative outcomes (Supplementary Methods and Results and Supplementary Fig. 3). However, let us note that this subjective loss formalism has the limitation that it cannot be straightforwardly translated to tasks where the lowest possible outcome is not a loss but zero. This is problematic as learning bias has also been found in such tasks[2].

Previous studies have suggested that learning rate asymmetries naturally implicate the development of choice hysteresis, where subjects tend to repeat previously rewarded options, despite current negative outcomes[2]. However, the very same choice behaviour may, in principle, derive from choice inertia (that is, the tendency to repeat a previously enacted choice[9,21]). To settle this issue, we directly compared these two accounts and found that choice hysteresis was overall better explained by a choice-confirmation bias in terms of both model comparison and parameter retrieval. Thus, in our task, we suggest that choice perseveration is better explained as biases in learning and updating. However, this does not exclude the possibility that learning-independent choice perseveration plays a role in decision-making; for instance, when the same pairs of cues are presented at the same spatial position across a higher number of trials.

Interestingly, theoretical simulations have suggested that preferentially learning from positive or negative prediction errors would be suboptimal in most circumstances, being only advantageous under specific and restrictive conditions (that is, in environments with extremely distributed resources). Thus, Cazé and van der Meer demonstrated that, in the long run, different learning rate asymmetries can be advantageous for certain reward contingencies, which they referred to as low- and high-reward conditions (that is, the reward probabilities associated with the two available options are both low, or both high, respectively)[14]. Consistent with previous reports, we found no detectable sign of learning rate adaptation as a function of the amount of reward available[7,22]. The absence of reward-dependent learning rates, if confirmed, is actually at odds with the above-mentioned optimality analysis, positing that it is more advantageous to have a lower positive than negative learning rate in high-reward conditions.

At first sight, it may be surprising that a biased model best accounts for participants' data in a task where the rewards are dependent on participants' performance. As a matter of fact, confirmation bias has been implicated in various suboptimal decisions, from wrongful convictions in judicial process[23] to academic underachievement[24] and misinterpretation of scientific evidence[25]. While biased learning may be suboptimal locally or under specific conditions (for example, being overly pessimistic about the consequences of other people's decisions), it could be on average well suited to adapting to periodically changing environments. In real-world situations, both the amount of resources and the causes that bring about these resources (such as when one is free to choose versus forced to take actions under influence or coercion) may vary from time to time. Overweighting positive consequences resulting from voluntary

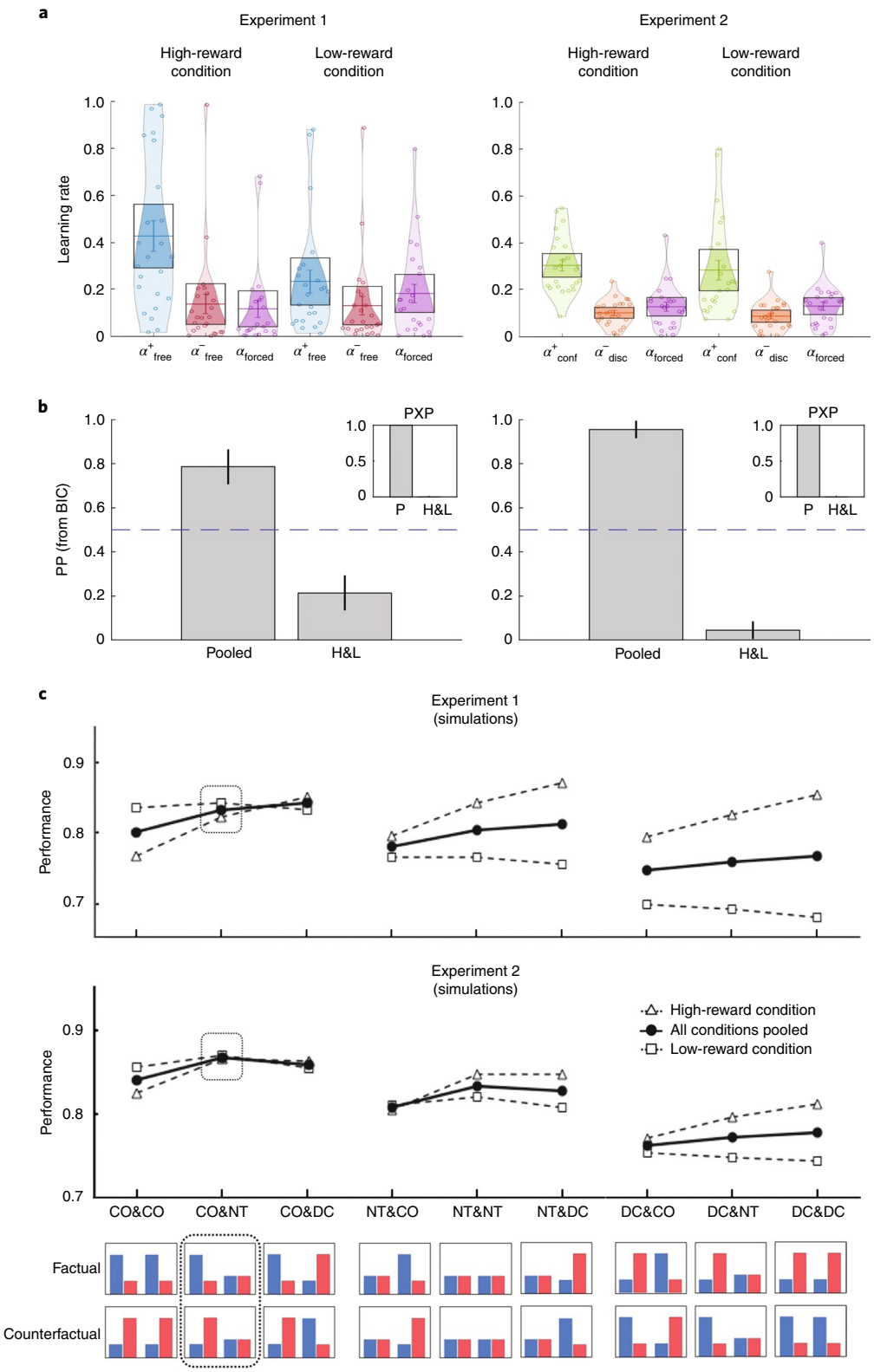

Fig. 7 | Learning rate analysis and model comparison for H&L models. a, Best-fitting learning rates of the H&L model. **b**, Posterior and exceedance probabilities of the pooled (P) and H&L models. In contrast with the H&L model, learning rates of the pooled model are not modulated by the amount of reward available. **c**, Parameter optimality was tested by simulating models with different learning rate patterns in experiments 1 ($n = 24$) and 2 ($n = 24$). The models are labelled according to their learning rate patterns, as shown in the bottom panel. For example, CO&NT designates a model with choice-confirmatory learning rates in free-choice trials and valence-neutral learning rates in forced-choice trials. The diamonds and squares correspond to the performance in high- and low-reward conditions, respectively. The black circles correspond to the performance averaged across the two conditions. Error bars are plotted, although most are so small they are not visible. Bottom: simulated learning rates. Blue and red bars represent positive and negative learning rates, respectively, in free- (left pairs) and forced-choice (right pairs) trials. The combination of learning rates that correspond to the combination observed in actual participants is highlighted by a dashed outline.

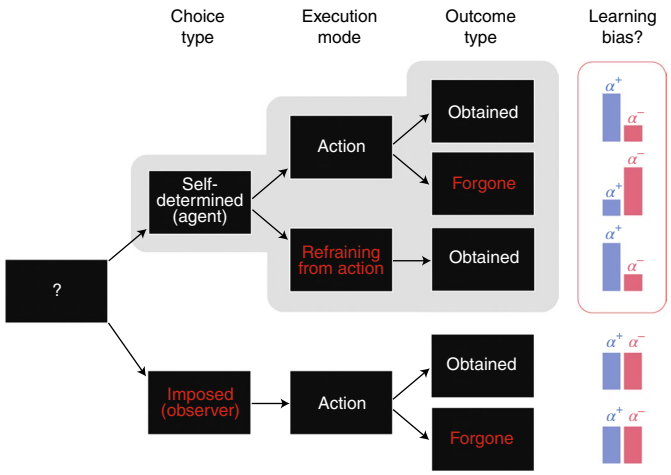

Fig. 8 | Valence-dependent learning biases as a function of the choice type, execution mode and outcome type. Top: self-determined choice stream. A valence-dependent learning asymmetry only arises when the individual is an agent (that is, controls the choice made). Positive (obtained) outcomes are better integrated than negative (obtained) outcomes (blue and red bars, respectively). This pattern reverses when learning from forgone outcomes, demonstrating that learning asymmetry reflects a choice-confirmation bias rather than a true positivity bias. Note that this bias emerges early in the action processing chain (that is, at the decision (or choice) stage rather than at the action stage). Bottom: imposed choice stream. No learning asymmetry is observed when the individual is not an agent (that is, does not control the choice made but only matches the computer's decision). In this situation, participants learn from positive and negative outcomes alike.

choices, while keeping impartiality when acting under influence, might be the most robust pattern to deal with the intrinsic volatility of disposable resources (low versus high) as well as the variety of their causal sources (internal versus external). Accordingly, we found that our best-fitting model (choice confirmatory in free choices; valence neutral in forced choices) was not only very advantageous in terms of accuracy, but also exhibited the most stable performances across low- and high-reward conditions, relative to other models with alternative patterns of learning rates (see Fig. 7c).

Previous discussions of confirmation bias often focused on person-level constructs, such as self-esteem, self-confidence and post-decisional dissonance. However, we additionally suggest that a choice-confirmation bias could be adaptive in the context of the natural environment in which the learning system evolved[26]. Previous accounts have highlighted the numerous benefits and facilitative effects of self-determined (versus forced) choices on learning performance[27]. Besides enhancing memory encoding and retention[28] and boosting selectivity to choice-consistent evidence[29,30], making self-determined choices improves learning of perceptual categories[31] and would allow for better generalization of previous knowledge to novel objects and situations[32]. Choice allows people to control the stream of evidence they experience, and hence to focus effort on information that aligns with their current needs or interests, resulting in better and better-targeted learning[33]. Choice is a powerful instrument with which to manipulate the environment so as to satisfy an individual's needs[34]. A choice-confirmation bias would lead to preferentially reinforced actions that are most likely to meet these needs (that is, freely chosen actions). In contrast, outcomes obtained from forced actions should be treated impartially as they do not necessarily align with the individual's needs, interests or values, and hence should not be assigned any special value in self-determined decisions[35].

Our results bear intriguing resemblance to recent findings on self-attribution in causal inference. In a reinforcement learning task manipulating the probability of hidden-agent interventions, Dorfman and colleagues[36] showed that when a participant believes that a benevolent agent has intervened, they learn more from negative than positive outcomes (that is, they infer that the negative outcome is a consequence of their own choice rather than due to the benevolent agent). Conversely, when they believe that an adversarial agent has intervened, the participant is more likely to learn from positive than negative outcomes. These findings highlight the relationship between valence-induced learning biases and control beliefs, and support the notion that people interpret feedback/changes in the environment differently according to perceived controllability. Controllability is a possible auxiliary hypothesis for interpreting changes in the environment[37]. Thus, if controllability is high, negative outcomes are presumably not a consequence of one's enacted choice, and so are underweighted (optimistic belief). Likewise, in our task, we found that people overweighted positive outcomes when selection of an option was under their direct control (free choice), yet were more impartial when they simply implemented an instructed choice. Note that both Dorfman's findings and ours are consistent with the notion that optimistic bias does not exclusively reflect preference for positive events in general, and hence is not only a consequence of increased salience of positive outcomes. Rather, it would reflect biased (control) beliefs about one's own causal power and the controllability of the environment[38]. Different beliefs about controllability might account for commonly observed differences between internal and external attribution profiles[39], as well as between optimistic and pessimistic explanatory styles[40].

Conclusions

In four studies mixing free- and forced-choice trials, featuring both factual and counterfactual learning and implementing distinct reward contingencies and action requirements, we showed that participants' behaviour was best accounted for by a learning model featuring a choice-confirmation bias (that is, a model amplifying positive prediction errors in free-choice trials while being valence neutral on forced-choice trials). We suggest that such a bias could be adaptive in the context of the natural environment in which the learning system evolved. Voluntary choices allow individuals to focus effort on information that aligns with their current needs. A choice-confirmation bias would thus lead to preferentially reinforced freely chosen actions, which are most likely to meet these needs. In contrast, outcomes obtained from unchosen (forced) actions should be treated impartially as they do not necessarily align with the individual's needs, interests or values, and hence should not be assigned any special value in self-determined decisions. Interestingly, choice can be seen as an opportunity to exert control over the environment. Our results support the notion that people interpret action feedback differently depending on how controllable their environment is, in line with previous findings about self-serving bias in causal inference.

Methods

Participants. The study included four experiments. Experiments 1, 2 and 4 involved 24 participants each (experiment 1: 13 males (mean age = 25.1 ± 0.8 years); experiment 2: nine males (mean age = 23.9 ± 0.5 years); experiment 4: ten males (mean age = 24.8 ± 0.7 years); Supplementary Table 1). Experiment 3 involved 30 participants (15 males; mean age = 24.2 ± 0.9). The sample size was chosen based on previous studies[7,41]. In experiment 4, four participants were excluded from further analysis because they pressed a computer key during the no-go trials more than 35% of the time. The local ethics committee approved the study (CPP C07-28). All participants gave written informed consent before inclusion in the study, which was carried out in accordance with the declaration of Helsinki (1964; revised 2013). The inclusion criteria were being older than 18 years, reporting no history of neurological or psychiatric disorders and having normal or corrected-to-normal vision. Participants were paid €10, €15 or €20, depending on the number of points they had accumulated during the experiment.

General procedure. Participants performed a probabilistic instrumental learning task modified from previous studies[2,7]. The task required choosing between two symbols that were associated with stationary outcome probabilities. The possible outcomes were either winning or losing a point. Participants were encouraged to accumulate as many points as possible and were informed that one symbol would result in winning more often than the other. They were given no explicit information about the exact reward probabilities, which they had to learn through trial and error.

Participants were also informed that some trials (indicated by the word observer displayed in the centre of the screen) were purely observational: the observed outcome would not be added to the total of points obtained so far, but it would allow them to gain knowledge about what would have happened should they have chosen the selected symbol (Fig. 1a). Crucially, in forced-choice trials, the two symbols were pseudo-randomly preselected, thus ensuring equal sampling from both low- and high-reward options. In experiment 2, participants were also informed that in some blocks they would see the outcome associated with the unchosen symbol, although they would only accumulate the points associated with the chosen symbol (Fig. 1a). As mentioned in the introduction, experiment 3 was included to address a concern raised during the review process. This third experiment only featured factual learning and both free- and forced-choice trials included a condition with a random reward schedule (50/50). In experiment 4, finally, the observer manipulation was not included, but subjects were instructed to express their decision by either making a key press (go trials) or refraining from making a key press (no-go trials) (Fig. 1b).

Data collection and analysis were not performed blind to the conditions of the experiments.

Conditions. Experiments 1–3 included four types of trials. In free-choice trials, participants could freely select between two possible symbols, while in forced-choice trials participants had to match a preselected option. In partial trials, participants were only shown the outcome (+1 or −1) associated with the chosen option, while in complete trials participants were shown the outcome of both the chosen and unchosen symbols (Fig. 1). Experiments 1 and 3 included two conditions: a condition with only partial free-choice trials (40 trials per block) and a condition with intermixed partial free- and forced-choice trials (40 + 40 = 80 trials per block). In this intermixed condition, the free- and forced-choice trials were pseudo-randomly presented within the block, and the same pair of symbols was used in both types of trial.

In addition to the condition with intermixed partial free- and forced-choice trials, experiment 2 also consisted of a condition with intermixed complete free- and forced-choice trials. For the sake of duration, the number of trials was halved in experiment 2 (20 free-choice trials + 20 forced-choice trials = 40 trials per block) (Supplementary Table 1).

Experiment 4 consisted of two free-choice conditions: a go and a no-go condition in which half of the participants had to press a computer key to select the top symbol, and to refrain from pressing any key to select the bottom symbol (the converse was true for the other half: press = bottom; refrain = top).

In experiment 1, participants underwent 12 blocks of either 40 (free) or 80 (intermixed free + forced) trials each. Six blocks were high-reward blocks and six blocks were low-reward blocks. In high-reward blocks, one of the two symbols was associated with a 0.9 probability of winning (+1 point)—and hence with a 0.1 probability of loss (−1 point). The other symbol was associated with a 0.6 probability of winning. In low-reward blocks, one symbol was associated with a 0.4 probability of winning and the other was associated with a 0.1 probability of winning. In experiment 2, each condition consisted of eight blocks of 40 (intermixed free + forced) trials each. Half of them were high-reward blocks. The low- and high-reward blocks were associated with the same reward contingencies as in experiment 1. In experiment 3, participants underwent 12 blocks of either 20 (free) or 40 (intermixed free + forced) trials each. Six blocks were random blocks and six blocks were instrumental blocks. In random blocks (50/50), each symbol was associated with a 0.5 probability of winning and losing (that is, there was no correct response in these blocks). Since the issue concerning the adaptive modulation of learning rates was addressed in experiments 1 and 2, we switched back to the contingencies used in a previous study[7] to define the instrumental blocks. Thus, in half of the instrumental blocks, one symbol was associated with a 0.7 probability of winning (+1 point)—and hence with a 0.3 probability of loss (−1 point)—and the other symbol was associated with a 0.3 probability of winning. In experiment 4, participants underwent six blocks of 100 (intermixed go + no-go) trials each. Reward contingencies were the same as those used in experiment 3.

In all experiments, each block was associated with a specific pair of symbols, meaning that the participant had to learn from scratch the reward contingencies at the beginning of each block. The first block was preceded by a short training session (60 trials for experiment 1; 40 trials for experiments 2 and 3; 40 trials for experiment 3). To ensure participants would not be biased towards expecting more frequent positive or negative outcomes in the subsequent experiment, the reward probabilities were set to 0.5 for all symbols during the training block.

Trial structure. In the first three experiments, trials began with a fixation cross, except when free- and forced-choice trials were intermixed, in which case the words actor or observer appeared for 1,000 ms before each trial, depending on the type of choice involved (free- or forced-choice, respectively) (see Fig. 1a). A pair of symbols was then presented in the left and right part of the screen (pseudo-randomly assigned on each trial). Participants made their choice by pressing the right or left arrow key with their right hand.

In forced-choice trials, the preselected cue was surrounded by a square. Participants had to press the corresponding arrow in order to move to the subsequent trial (nothing happened if they tried to press the other arrow). The cues were preselected to ensure equal sampling: one symbol was preselected in half of the trials and the other symbol was preselected in the remaining trials. The obtained outcome was then presented in the same part of the screen as the chosen symbol. In complete trials, the foregone outcome was shown in the same part as the unchosen symbol. In intermixed free- and forced-choice trials, to ensure that participants paid attention to the outcomes presented, they were asked to press the up arrow key when winning a point and the down arrow key when losing a point.

In experiment 4, trials began with a fixation cross for 1,000 ms (Fig. 1b). A pair of symbols was then presented at the top or bottom of the screen (pseudo-randomly assigned on each trial). Participants had 1,500 ms to press the instructed key: the up arrow key for half of the participants and the bottom arrow key for the other half (go trials). If no key was pressed after that delay, the other symbol was automatically selected (no-go trials). In both go and no-go trials, a feedback associated with the selected symbol was then displayed for 1,500 ms.

Computational modelling. We fitted the data with modified versions of a *Q*-learning model, including different learning rates following positive and negative prediction errors, and different learning rates in free- and forced-choice trials, for 50/50 and 70/30 reward schedules, or in go and no-go trials (see below). For each pair of symbols, the model estimated the expected value (also called the *Q* value) of the two options. The *Q* values were set to 0 at the beginning of each block, corresponding to the a priori expectation of an equal probability of winning or losing one point. After each trial, t, the value of the chosen option in a given state (s), was updated based on the prediction error, which measured the discrepancy between the actual outcome value and the expected outcome for the chosen symbol (that is, the chosen (c) Q value) as follows:

$$\delta_t(\mathrm{c}) = R_t(\mathrm{c}) - Q_t(c,s)$$

where R_t(c) represents the obtained (factual) outcome on trial t.

The prediction error was then used to update the chosen *Q* value:

$$Q_{t+1}(c,s) = Q_t(c,s) + \alpha \times \delta_t(\mathrm{c})$$

where α represents the learning rate parameter.

In the complete condition experienced in experiment 2, participants could learn from both the obtained and the forgone outcomes. Thus, in these trials, the unchosen *Q* value was also updated with the forgone (or unchosen, u) outcome using the same rule:

$$\delta_t(\mathrm{u}) = R_t(\mathrm{u}) - Q_t(u,s)$$
$$Q_{t+1}(u,s) = Q_t(u,s) + \alpha \times \delta_t(\mathrm{u})$$

As mentioned above, different learning rates (α^+ and α^-) were fitted to reflect different updating processes after a positive or negative outcome[2,7]. R could take the values $R \in \{-1, +1\}$, depending on whether or not the subject won a point. The *Q* values were initialized as $Q_{t=1} = 0$, which corresponds to unbiased prior expectations, and to the average outcome experienced during the training phase. Because we were interested in the specific effect of choice type on learning, different pairs of asymmetrical learning rates in free- and forced-choice trials (experiments 1–3), and for factual and counterfactual outcomes (experiment 2 only), were also fitted. The full model thus had four learning rates in experiment 1 and eight learning rates in experiment 2. In experiment 3, different pairs of learning rates were also fitted in random (50/50 reward schedule) and instrumental (70/30 reward schedule) blocks, in addition to positive and negative outcomes and free- and forced-choice trials. The full model therefore included eight learning rates. In experiment 4, different pairs of learning rates were fitted for go and no-go trials, and the full model included four learning rates (as in experiment 1).

In the reinforcement learning framework, the stimulus with the highest *Q* value was more likely to be selected. The probability of selecting the stimulus with the highest value was estimated with a softmax function, as follows:

$$P_t(a,s) = \frac{e^{\beta \times Q_t(a,s)}}{e^{\beta \times Q_t(a,s)} + e^{\beta \times Q_t(b,s)}}$$

where β is the exploitation intensity parameter, which represents the strength of the *Q* values on choice selection, and a and b are the two options available in a given state, s. We fitted a unique parameter β for all trials and outcome types, to avoid biasing the learning rate comparison procedure. We also designed simpler versions of the full models in order to assess, for each experiment, what was the maximum number of parameters authorized, when penalizing for their complexity (parsimony-driven dimensionality reduction). The model space ranged from full models assuming different learning rates for all possible outcomes (obtained and

forgone), choice (free and forced), reward contingencies (50/50 and 70/30) and action types (go and no go), to a fully reduced model assuming no bias at all.

Perseveration model. Following Katahira's critique[9], suggesting that learning rate asymmetries may be artifactually driven by a repetition (or perseveration) bias, we compared our models with a model including a stickiness parameter. In the latter, the action selection rule was modified as follows:

$$P_t(a,s) = \frac{e^{\beta \times (Q_t(a,s)+\rho \times C_t(a,s))}}{e^{\beta \times (Q_t(a,s)+\rho \times C_t(a,s))} + e^{\beta \times (Q_t(b,s)+\rho \times C_t(b,s))}}$$

where the parameter ρ represents the participant's tendency to perseverate, and $C_t(x, s)$ indicates which stimulus was chosen on the previous trial:

$$C_t(x,s) = \begin{cases} 1 & \text{if } x \text{ was chosen on the previous trial} \\ 0 & \text{otherwise} \end{cases}$$

When the participants were forced to match the preselected option, we considered that they would not tend to perseverate in that choice. In the subsequent free trials, we thus set $C_t(x, s)$ to zero for both the preselected and other stimuli.

Parameter estimation. We fitted the model parameters based on participants' choices on each free-choice trial, for each participant. We used a maximum posterior approach (MAP)[42] to avoid degenerate parameter estimates. The best parameters were those maximizing the logarithm of the PP (LPP):

$$\ln[P(\theta|\text{choice}_{1:N})] \propto \ln[P(\theta)] + \sum_{t=1}^{N} \ln[P(\text{choice}_t|\theta)]$$

where θ represents our parameter set, N is the total number of trials in the experiment and $P(\text{choice}_t|\theta)$ is the probability that the model would choose the same stimulus as the participant on trial t. To maximize the LPP with respect to θ, we used MATLAB's fmincon function with the following ranges: $0 < \beta <$ infinite; and $0 < \alpha_i < 1$.

The parameter prior probabilities were based on Daw et al.[43], and we used a gamma distribution with the hyperparameters 1.2 and 5 for the β parameter, and a beta distribution with the hyperparameters 1.1 and 1.1 for the learning rate (α) parameters. To avoid biasing the learning rate comparison procedure, the same priors were used for all learning rates.

Parameter recovery. We performed a parameter recovery analysis to ensure that the values of the learning rates reflected true differences in learning, and were not an artefact of the parameter-fitting procedure[44]. The aim was to check the capacity of recovering the correct parameters using simulated datasets. To do so, we first simulated performance on the two behavioural experiments using virtual participants. For each of these virtual participants, a learning rate value was randomly drawn from a uniform distribution between 0 and 1. We then averaged the correlation coefficients (R values) and P values from 100 correlations performed between the parameters manipulated and the parameters recovered from the fitting procedure applied to the simulated dataset[45]. This analysis was conducted on all of the learning rate parameters of the full models (see Supplementary Methods and Results and Supplementary Fig. 1).

Model comparison. The LPP was used to compute the BIC[46] for each model and each participant, as follows:

$$\text{BIC} = k \times \ln[N] - 2 \times \ln[P(\theta_{\text{MAP}}|\text{choice}_{1:N})]$$

where k is the number of parameters and $\ln[P(\theta_{\text{MAP}}|\text{choice}_{1:N})]$ is the LPP of the MAP parameters given the participant's choice data.

BIC of the different models were compared to verify that the extra learning rate parameters were justified by the data. As an approximation of the model evidence, individual BICs were fed into the MBB-VB toolbox[12]—a procedure that estimates how likely it is that a specific model generates the data of a randomly chosen subject (the PP of a model) as well as the protected exceedance probability (PXP) of one model being more likely than any other models in the set.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the GitHub repository (https://github.com/spalminteri/agency).

Code availability

Custom code scripts have been made available on the GitHub repository (https://github.com/spalminteri/agency). Additional modified scripts can be accessed upon request.

Received: 10 September 2019; Accepted: 26 June 2020;
Published online: 3 August 2020

References

1. Barto, A. G. & Sutton, R. S. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
2. 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).
3. Aberg, K. C., Doell, K. C. & Schwartz, S. Linking individual learning styles to approach-avoidance motivational traits and computational aspects of reinforcement learning. *PLoS ONE* **11**, e0166675 (2016).
4. 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).
5. Sharot, T. & Garrett, N. Forming beliefs: why valence matters. *Trends Cogn. Sci.* **20**, 25–33 (2016).
6. Kuzmanovic, B. & Rigoux. L. Optimistic belief updating deviates from Bayesian learning. *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2810063 (2016).
7. Palminteri, S., Lefebvre, G., Kilford, E. J. & Blakemore, S. J. Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. *PLoS Comput. Biol.* **138**, e1005684 (2017).
8. Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
9. Katahira, K. The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* **87**, 31–45 (2018).
10. Boureau, Y. L. & Dayan, P. Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology* **36**, 74–97 (2011).
11. Guitart-Masip, M. et al. Go and no-go learning in reward and punishment: interactions between affect and effect. *NeuroImage* **62**, 154–166 (2012).
12. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* **10**, e1003441 (2014).
13. Correa, C. M. et al. How the level of reward awareness changes the computational and electrophysiological signatures of reinforcement learning. *J. Neurosci.* **38**, 10338–10348 (2018).
14. Cazé, R. D. & van der Meer, M. A. Adaptive properties of differential learning rates for positive and negative outcomes. *Biol. Cybern.* **107**, 711–719 (2013).
15. Benjamin, D. J. *Errors in Probabilistic Reasoning and Judgment Biases* No. w25200 (National Bureau of Economic Research, 2018).
16. Alicke, M. D. & Govorun, O. in *The Self in Social Judgement* (eds Alicke, M. et al.) 83–106 (Psychology Press, 2005).
17. Harris, A. J. & Osman, M. The illusion of control: a Bayesian perspective. *Synthese* **189**, 29–38 (2012).
18. Ajzen, I. Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *J. Appl. Soc. Psychol.* **32**, 665–683 (2002).
19. Kool, W., Getz, S. J. & Botvinick, M. M. Neural representation of reward probability: evidence from the illusion of control. *J. Cogn. Neurosci.* **25**, 852–861 (2013).
20. Izuma, K. et al. Neural correlates of cognitive dissonance and choice-induced preference change. *Proc. Natl Acad. Sci. USA* **107**, 22014–22019 (2010).
21. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
22. Gershman, S. J. Do learning rates adapt to the distribution of rewards? *Psychon. Bull. Rev.* **22**, 1320–1327 (2015).
23. Findley, K. A. & Scott, M. S. Multiple dimensions of tunnel vision in criminal cases. *Wis. L. Rev.* **2006**, 291–397 (2006).
24. Rosenthal, R. & Jacobson, L. *Pygmalion in the Classroom* (Irvington, 1992).
25. Loehle, C. Hypothesis testing in ecology: psychological aspects and the importance of theory maturation. *Q. Rev. Biol.* **62**, 397–409 (1987).
26. Fawcett, T. W. et al. The evolution of decision rules in complex environments. *Trends Cogn. Sci.* **18**, 153–161 (2014).
27. Murayama, K. et al. How self-determined choice facilitates performance: a key role of the ventromedial prefrontal cortex. *Cereb. Cortex* **25**, 1241–1251 (2013).
28. Voss, J. L., Gonsalves, B. D., Federmeier, K. D., Tranel, D. & Cohen, N. J. Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nat. Neurosci.* **14**, 115–120 (2011).
29. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation bias through selective overweighting of choice-consistent evidence. *Curr. Biol.* **28**, 3128–3135 (2018).
30. Chambon, V. et al. Neural coding of prior expectations in hierarchical intention inference. *Sci. Rep.* **7**, 1278 (2017).
31. Markant, D. & Gureckis, T. Category learning through active sampling. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (eds Ohlsson, S. & Catrambone, R.) 248–253 (Cognitive Science Society, 2010).

32. Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007).
33. Gureckis, T. M. & Markant, D. B. Self-directed learning: a cognitive and computational perspective. *Perspect. Psychol. Sci.* **7**, 464–481 (2012).
34. Leotti, L. A. & Delgado, M. R. The inherent reward of choice. *Psychol. Sci.* **22**, 1310–1318 (2011).
35. Cockburn, J., Collins, A. G. & Frank, M. J. A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron* **83**, 551–557 (2014).
36. Dorfman, H. M., Bhui, R., Hughes, B. L. & Gershman, S. J. Causal inference about good and bad outcomes. *Psychol. Sci.* **30**, 516–525 (2019).
37. Gershman, S. J. How to never be wrong. *Psychon. Bull. Rev.* **26**, 13–28 (2019).
38. Chambon, V., Thero, H., Findling, C. & Koechlin, E. Believing in one's power: a counterfactual heuristic for goal-directed control. Preprint at *bioRxiv* https://doi.org/10.1101/498675 (2018).
39. Rotter, J. B. *Social Learning and Clinical Psychology* (Prentice-Hall, 1954).
40. Abramson, L. Y., Seligman, M. E. & Teasdale, J. D. Learned helplessness in humans: critique and reformulation. *J. Abnorm. Psychol.* **87**, 49–74 (1978).
41. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
42. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
43. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
44. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
45. Meyniel, F. et al. A specific role for serotonin in overcoming effort cost. *eLife* **5**, e17282 (2016).
46. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).

Acknowledgements

V.C. was supported by the Agence Nationale de la Recherche (ANR) grants ANR-17-EURE-0017 (Frontiers in Cognition), ANR-10-IDEX-0001-02 PSL (program 'Investissements d'Avenir') and ANR-16-CE37-0012-01 (ANR JCJ) and ANR-19-CE37-0014-01 (ANR PRC). H.T. was supported by a PSL/ENS studentship. M.V. was supported by FIRE ('Programme Bettencourt') and by a Région Île-de-France studentship. P.H. was supported by the Chaire Blaise Pascal of the Région Île-de-France. S.P. was supported by an ATIP-Avenir grant (R16069JS), the Programme Emergence(s) de la Ville de Paris, the Fyssen Foundation and the Fondation Schlumberger pour l'Education et la Recherche (FSER). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

V.C., S.P. and P.H. developed the study concept. Testing and data collection were performed by H.T. and M.V. H.V. helped to write the Psychtoolbox script for data collection. Data analysis was performed by V.C., H.T., M.V. and S.P. V.C. and H.T. drafted the manuscript. S.P. and P.H. provided critical revisions. All authors approved the final version of the manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0919-5.

Correspondence and requests for materials should be addressed to V.C., H.T. or S.P.

Peer review information Primary Handling Editor: Marike Schiffer.

nature research

Corresponding author(s): Valerian Chambon, Héloise Théro, Stefano Palminteri

Last updated by author(s): 27 April 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Matlab R2016 (Psychtoolbox) was used to collect the data
Data analysis	Matlab R2016 was used to analyze the data

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data that support the findings of this study are available from Github repository https://github.com/spalminteri/agency.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative data were collected in the 4 experiments
Research sample	The study included 4 experiments. Experiments 1, 2 and 4 involved 24 participants (Experiment 1: 13 males, mean age = 25.1 ± 0.8; Experiment 2: 9 males, mean age = 23.9 ± 0.5; Experiment 4: 10 males, mean age = 24.8 ± 0.7), whereas experiment 3 involved 30 participants (15 males, mean age = 24.2 ± 0.9). All participants were right-handed, French-native speakers, and had no history of neurological or psychiatric disorders. We defined the number of subjects in accordance with previous studies similarly contrasting factual and counterfactual learning in two-armed bandit tasks involving neutral stimuli and monetary outcomes: Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). PLoS computational biology, 13(8), e1005684 (Experiment 1: 20 participants; Experiment 2: 20 participants); Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Nature Communications, 6(1) (28 participants).
Sampling strategy	The Experiment 1 included two conditions: a condition with only "partial" free-choice trials (40 trials per block), and a condition with intermixed "partial" free- and forced-choice trials (40 + 40 = 80 trials per block). In this "intermixed" condition, the free- and forced-choice trials were pseudo-randomly presented within the block, and the same pair of symbols was used in both types of trial. In addition to the condition with intermixed partial free- and forced-choice trials, the Experiment 2 also consisted in a condition with intermixed "complete" free- and forced-choice trials. Experiment 3 included two conditions: a condition with only "partial" free-choice trials, and a condition with intermixed "partial" free- and forced-choice trials. In this experiment, participants underwent 12 blocks of either 20 (free) or 40 (intermixed free+forced) trials each. Six blocks were "random" blocks and six blocks were "instrumental" blocks. In "random" blocks (50/50), each symbol was associated with .5 probability of winning and losing (i.e. there was no correct response in these blocks). Experiment 4 consisted of two "free-choice" conditions: a "go" and a "no-go" condition in which half of the participants had to press a computer key to select the top symbol, and to refrain from pressing any key to select the bottom symbol. No saturation criteria were used.
Data collection	Behavioural data were collected using a 17-inch Apple MacBookPro laptop. Only the participant and the researcher were in the room during the experiment. The researcher was blind to experimental conditions, but not to the study hypothesis.
Timing	Experiment 1: 02/10/2017 - 20/10/2017 Experiment 2: 01/12/2017 - 22/12/2017 Experiment 3: 05/01/2018 - 19/01.2018 Experiment 4: 05/02/2018 - 02/03/2018
Data exclusions	In the experiment 4, four participants were excluded from further analysis based on a pre-established exclusion criterion, i.e. pressing a computer key during the No-Go trials in more than a third of the trials.
Non-participation	No participant dropped out/declined participation
Randomization	Participants were not allocated into experimental groups

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	Participants were recruited using a webmail list (RISC)

Ethics oversight	The local ethics committee approved the study (Comité de Protection des Personnes, CPP C07-28)

Note that full information on the approval of the study protocol must also be provided in the manuscript.