The computational roots of positivity and confirmation biases in reinforcement learning

Stefano Palminteri, Maël Lebreton

Title:

The computational roots of positivity and confirmation biases in reinforcement-learning

Authors:
Stefano Palminteri(1,2,3) and Maël Lebreton(4,5,6)

Affiliations:
(1) Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et Recherche Médicale, Paris, France
(2) Département d'études cognitives, Ecole Normale Supérieure, Paris, France
(3) Université de Recherche Paris Sciences et Lettres
(4) Paris School of Economics, Paris, France
(5) LabNIC, Department of Fundamental Neurosciences, University of Geneva, Geneva, Swiss
(6) Swiss Center for Affective Science, Geneva, Swiss

Correspondence:
stefano.palminteri@ens.fr, mael.lebreton@pse.fr

Keywords (6)
learning, confirmation, gain, loss, update, decision

Abstract
Humans do not integrate new information objectively: outcomes carrying a positive affective value and evidence confirming one's own prior belief are overweighed. Until recently, theoretical and empirical accounts of the positivity and confirmation biases assumed them to be specific to "high level" belief updates. We present evidence against this account. Learning rates in reinforcement learning tasks, estimated across context and species, generally present the same characteristic asymmetry, suggesting that belief and value updating processes share key computational principles and distortions. This bias generates over-optimistic expectations about the probability of making the right choices and, consequently, generates over-optimistic reward expectations. We discuss the normative and neurobiological roots of these reinforcement learning biases and their position within the greater picture of behavioral decision-making theories.

From belief updating to reinforcement learning

Our decisions critically depend on the beliefs we have about the options available to us: their probability of occurrence, conditional on the actions that we undertake, and their value – i.e. how good they are. It is therefore not surprising that an ever-growing literature in cognitive psychology and behavioral economics focuses on how humans form and update their beliefs. While Bayesian inference principles provide a normative solution for how beliefs can be optimally updated when we receive new information, in humans, belief-updating behaviors often deviate from this normative benchmark . Among the most prominent systematic deviations, the positivity and the confirmation **biases** (see Glossary) stand out for their pervasiveness and ecological relevance [1].

The **positivity bias** characterizes the fact that decision-makers tend to update their beliefs more when new evidence conveys a positive valence [1,2]. This bias has notoriously been revealed in situations where subjects learn something about themselves and preferentially integrate information that convey a positive signal (e.g. a higher IQ or a lower risk of disease) [3–5]. The **confirmation bias** characterizes the fact that decision-makers tend to update their beliefs more when new evidence confirms their prior beliefs and past decisions compared to when it disconfirms or contradicts them [1,6]. This bias can take many forms – extending to positive test strategies and selective information sampling – and it has been robustly reported in a variety of natural or laboratory experimental setups [6,7]. Of note, in most ecological settings, positivity and confirmatory biases co-occur [8,9]. Indeed, unless a cogent experimental design carefully orthogonalizes them, we typically hold opinions and select actions that we believe have a positive subjective value (e.g., a higher payoff in economic settings). Therefore, after a better than expected outcome, such actions result in a positive and confirmatory update [3,10,11].

To date, the dominant framework used to explain the existence and persistence of asymmetric belief updating posits that they stem from a 'rational' cost-benefit trade-off. The cost of holding objectively wrong (i.e., overly optimistic) beliefs is traded against the psychological benefits of them being self-serving: believing in a world that is pleasant and reassuring per se (consumption value) [2,12–15]. While originally designed to account for the positivity bias, this logic arguably extends to the confirmation bias, when one considers that being right (as signaled by confirmatory information) is also valuable and self-serving ("the ego utility consequences of being right", cit. [3]). Importantly, both original and more recent versions of this theoretical account of asymmetric belief updating explicitly suggest that this class of learning biases is specific to high-level and ego-relevant beliefs [2,12,14,15], a position that seems supported by the fact that the positivity bias (as opposed to the confirmatory bias) does not clearly extend to belief updates that are not ego-relevant, such as in purely financial contexts [10,16–19].

In the present article, we review recent empirical and modelling studies that challenge the standard account, and suggest that the asymmetries that affect high-level belief updates are shared with more elementary forms of updates. This set of empirical findings cannot be purely explained by the dominant, self-serving bias account of asymmetric updates, and shows that some forms of positivity and confirmatory biases occur across a wide variety of species and contexts.

Testing asymmetric updating in the reinforcement-learning framework

Arguably, the reinforcement-learning (RL) framework represents the ideal elementary form of motivated belief updating. RL characterizes the behavioral processes that consist of selecting among alternative courses of action, based on inferred economic (or affective) values that are learned by interacting with the environment [20]. In addition to being computationally simple, elegant, and tractable, the most popular RL algorithms can solve (or be a core component of the solution to) higher-level cognitive tasks, such as spatial navigation, games involving strategic interactions, and even complex video games, thereby consisting an ideal basic building block for higher-level cognitive processes [21,22].

The basic experimental framework of a two-armed bandit task (often referred to as a two-alternative forced-choice task) provides all the key elements necessary to assess the pervasiveness of positivity and confirmation bias. In this simplified set-up, the decision-maker faces two neutral cues, associated with different reward distributions (**Figure 1A**). In the most popular RL formalism, the decision-maker learns, through an error-correction mechanism, to attach subjective values (Q(:)) to each option, which they use to make later choices (**Figure 1B**).

Concretely, once an option is chosen ('c'), the decision-maker receives an outcome R. The outcome is compared to its subjective value, generating a **prediction error**

$$\mathrm{PE}(c) = \mathrm{R}(c) - \mathrm{Q}(c)$$

The prediction error is then used to update the subjective value of the chosen option via an error correction mechanism involving a weighting parameter, the **learning rate**:

$$\mathrm{Q}(c) \leftarrow \mathrm{Q}(c) + \alpha * \mathrm{PE}(c)$$

Reframed in terms of belief updating, the magnitude of the prediction error quantifies how surprising the experienced outcome is, while its sign (positive or negative) specifies the valence of the information carried by the experienced outcome. In other words, positive prediction errors follow outcomes that are better than expected (i.e. they signal relative gains or good news), while negative prediction errors follow outcomes that are worse than expected (i.e. they signal relative losses or bad news). In addition, a positive prediction error following the chosen option confirms that the decision-maker was right to pick the current course of action (and the converse is true for a negative prediction error). In theory, it is possible to define two different learning rates, following these two types of prediction errors:

$$\mathrm{Q}(c) \leftarrow \mathrm{Q}(c) + \begin{cases} \alpha_+ * \mathrm{PE}(c), \text{if } \mathrm{PE}(c) > 0 \\ \alpha_- * \mathrm{PE}(c), \text{if } \mathrm{PE}(c) < 0 \end{cases}$$

As a consequence, in this simplified experimental and computational framework, an elementary counterpart of both the positivity and confirmation bias should be reflected in a learning rate asymmetry – i.e. in the fact that positive learning rates (α_+) are higher than negative ones (α_-). In the following sections, we review evidence in favor (or challenging) the hypothesis that updating biases analogous to the positivity bias and confirmation bias occur in simple reinforcement learning tasks.

Value update biases in reinforcement learning

Positivity bias in reinforcement learning

About fifteen years ago, a few studies incidentally started fitting variants of the Q-learning model to human data collected in simple reinforcement learning tasks [23–27]. Notably, they fitted Q-learning models with separate learning rates depending on prediction error valence (Q($\alpha\pm$)). Comparisons between the two learning rates generally revealed a positivity bias ($\alpha_+ > \alpha_-$), although sometimes results were mixed across groups or learning phases.

Arguably, a strong demonstration of a positivity bias requires three steps, which were usually absent in these incidental observations: first, the Q($\alpha\pm$) model should outperform the standard model with one learning rate Q(α) in a stringent **model comparison** [24]; second, although allowed to vary across individuals, the comparison of the two learning rates estimated from **model fitting** should reveal a significant asymmetry on average, such as $\alpha_+ > \alpha_-$; third, behavioural data should exhibit at least one qualitative pattern which falsifies the standard model, while being explained by the Q($\alpha\pm$) (see [28] and **Box 1** for a survey of the behavioural signatures of the positivity bias). These three levels of demonstration were unambiguously achieved in a recent study investigating asymmetric updating in a simple two-armed bandit task in humans [29]. The fact that individuals update the option values more following positive rather than negative prediction errors leads to optimistic overestimating of reward expectations and a heightened probability of selecting what the decision-maker believes is the best option. Importantly, the key aspects of such optimistic reinforcement-learning were later replicated in fully incentivized experiments, which included various types of outcome ranges, such as gain (+0.5€ / 0.0€), loss (0.0€ / -0.5€) and mixed contexts (+0.5€ / -0.5€) [29,30]. These results confirm that negative prediction errors are down-weighted relative to positive prediction errors even when they are associated with actual monetary losses. Moreover, the positivity bias cannot be neutralized, nor reverted, by either increasing the saliency of negative outcomes (0.0€ → -0.5€) or decreasing the saliency of the positive outcomes (+0.5€ → 0.0€). Finally, this also tells us that the bias depends on the valence (or sign) of the prediction error and not the outcome.

Generalizing the results

Since then, several other studies featuring different experimental designs also fitted the Q($\alpha\pm$) model, thus putting learning asymmetry to the test. In a task featuring different regimens of outcome uncertainty, learning rates are typically adaptively modulated as a function of this environmental volatility: learning rates in a volatile condition are higher than those in a stable condition [31]. In addition to this adaptive modulation, a positivity bias can be observed in human participants in both the low and high volatility conditions [32]. When the same volatility task features an 'appetitive' treatment (winning money versus nothing) and an 'aversive' treatment (getting a mild electric shock versus nothing), a positivity bias is reported in human participants in all treatments (rewarding and aversive) and conditions (stable and volatile) (**Figure 2A**) [33].

The positivity bias in learning rates has been found beyond two-armed bandit task contexts, such as in foraging situations [34], in multi-attribute reinforcement learning (e.g. instantiated by Wisconsin Card Sorting Test [35]), in strategic interactions and multi-step decisions with delayed rewards [36], and in learning transitivity relations [37].

These results suggest that the positivity bias is robust to major variations of experimental protocols, from uncertainty about the outcomes (stable vs volatile) to differences in the nature of the outcomes themselves (e.g. primary, like electric shocks, or secondary, like money) and the extension of the state-transition structure of the task beyond two-armed bandits. It is worth noting, however, that on some occasions, studies failed to find a positivity bias or even reported a negativity bias ($\alpha_+ < \alpha_-$) [38–43]. We argue that sources of such inconsistencies could sometimes be found in specific choices concerning model specification that can hinder the identification of a positivity bias (see **Box 2**). Other features of the design, such as mixing instrumental (or 'free') choices and Pavlovian (or 'forced') trials may also have blurred the result (see the section below).

From positivity to confirmatory bias

The studies surveyed so far all feature what is often referred to as partial feedback conditions, i.e., the standard situation where the subject is informed only about the outcome of the chosen option (**Figure 1A** and [44]). Critically, under this standard set-up, it is not possible to assess whether the reported positivity bias actually reflects a saliency bias ('all positive prediction errors are overweighed') or a choice-confirmation bias ('only positive prediction errors following obtained outcomes are overweighed'). To tease apart these interpretations, we conducted a series of studies leveraging complete feedback conditions, that consist of also displaying the forgone (or counterfactual) outcome, i.e. the outcome associated with the unchosen option in a two-armed bandit task [45,46]. Under the saliency bias hypothesis, one expects larger learning rates for positive prediction errors, independent of them being associated with the chosen or unchosen option. Under the confirmation bias hypothesis, one expects an interaction between the valence of the prediction error and its association with the chosen or the unchosen option (**Figure 1B**). The rationale is that a better-than-expected forgone outcome can be interpreted as a relative loss, as it indicates that the alternative course of action could have been beneficial (a disconfirmatory signal). Symmetrically, a worse-than-expected forgone outcome can be interpreted as a relative gain as it indicates that the current course of action is advantageous (a confirmatory signal).

In a recent study that explicitly and systematically exploited this rationale, we observed the interaction characterising the confirmation bias hypothesis: positive and negative learning-rates associated with the unchosen option mirrored the learning rates associated with the chosen option (**Figure 3**; left). Additional model-comparison analyses showed that the four learning-rate model could be reduced to a two learning-rate model, featuring a single parameter for all confirmatory and all disconfirmatory feedback, respectively (**Figure 1C**). The symmetrical pattern of learning rates, as well as the superiority of this implementation of choice confirmation bias against other models, has been replicated several times in RL tasks that include both partial and complete feedback information [47–49].

In a follow-up study that further investigated the choice-related aspects of the positivity bias, standard instrumental trials were interleaved with observational trials, where participants observed the computer making a choice for them and the resulting outcome [45]. Results from model-fitting and model-comparison indicated that the update bias was specific to freely chosen outcomes, further corroborating the presence of a proper choice-confirmation bias. Importantly, the fact that agency seems mandatory to observe the choice confirmation bias [45,50] is reminiscent of the ego-relevance aspect of belief-updating biases.

Finally, several studies have experimentally manipulated participants' beliefs about the option values through task instructions (e.g., by explicitly indicating option values to the participants before the beginning of the experiment) [51,52]). Behavioral results in this task are consistent with a model that assumes that the usual learning asymmetry is further exacerbated by the (instructed) prior about the option value, such that positive prediction errors following options with a positive prior are over-weighted (and the reverse is true for options with a negative prior). Therefore the available evidence is consistent with the idea that belief-confirmation bias can be induced in the context of reinforcement learning via semantic instructions, thus suggesting a permeability between cognitive representations and instrumental associations.

Positivity and confirmation biases across evolution and development

A valuable aspect of reinforcement-learning tasks in general (and n-armed bandits in particular) is that they are routinely used in non-human research, opening up the possibility of testing the comparative validity of the positivity bias results. To our knowledge, to date, few studies have tested the dual-learning model in other species. Among those few, one study featuring stable and volatile phases tested both humans and rhesus monkeys (*macaca mulatta)* with the same task [32]. Like humans, monkeys displayed a positivity bias, whose size, was, if anything, larger than that observed in humans (**Figure 2B;** see **Box 1** for possible behavioral consequences). A couple of recent studies in rodents (*rattus norvegicus*) also provide support for the positivity bias [53,54]. In addition, they suggest that the bias could be modulated by factors such as the stage of learning (the bias being larger in the exploratory phase) and the overall value of the decision problem (the bias being larger in 'poor' environments) (**Figure 2C**).

Regarding the developmental aspects of positivity and confirmation bias, a series of recent studies investigated learning behaviour in a simple two-armed bandit task in cohorts including children and young adults. While most of these studies actually report a positivity bias in all age groups [55–58] (but see [59]), they draw conflicting conclusions regarding the developmental trajectories of the bias. Further studies are therefore required to better assess the trajectory of these biases during development and ageing, as well as identify the individual traits and tendencies that promote or counteract them.

Is confirmatory updating a flaw or a desirable feature of reinforcement learning?

The presence of update bias (such as the positivity and the confirmation bias) in basic reinforcement learning across species and contexts naturally raises the question of why evolution has selected and maintained what can be perceived, prima facie, as error-introducing processes that generate apparently irrational behavioral tendencies (**Box 1**).

Statistical normativity of choice-confirmation bias

Early simulations restricted to specific task contingencies and partial feedback regimens demonstrated that a positivity bias is optimal in learning contexts with a low overall reward rate ('poor' environments) but detrimental in learning contexts with a high overall reward rate ('rich' environments) [60]. This result can be intuitively understood as a consequence of the fact that, in partial feedback situations, it is

rational to preferentially take into account the prediction errors that are rare (i.e., positive prediction errors in 'poor' environments and negative prediction errors in 'rich' environments) (**Figure 3A**). However, to date, experimental data has not provided convincing evidence in favor of an inversion of the learning bias as a function of task demands [39,45] (but see [55] for a partial adaptation). Accordingly, a positivity bias following partial feedback is maintained in tasks involving contingency reversals and volatility [33,46], even though these reduce the learner's capacity to quickly adapt their responses in these conditions (**Box 1**). However, the fact that the positivity bias appears maladaptive in some (laboratory-based) conditions does not rule out the possibility that it has been selected and maintained by evolution because it could still be adaptive in most ecologically relevant scenarios [61]. Indeed, the fact that the bias is documented in several species, suggests that its statistical advantages should apply across a broad range of ecological contexts.

A recent study systematic analysed the performance of the choice-confirmation bias in complete feedback contexts to clarify its statistical properties. Specifically, the study assessed its optimality in a larger space of learning problems, including 'rich' and 'poor', 'stable' and 'volatile' environments, as well as more demanding decision problems [62]. The authors reported that confirmatory-biased RL algorithms generally outperform their unbiased counterparts (**Figure 3B**). This counterintuitive result, replicated by other simulation studies, arises from the fact that confirmatory RL algorithms mechanistically neglect uninformative –stochastic– negative prediction errors associated with the best response. Thereby they accumulate resources (i.e. collect rewards and avoid losses) more efficiently than their unbiased counterparts [62–64]. Thus, confirmatory updating appears to facilitate and optimize learning and performance in a broad range of learning situations [61,65].

Metacognitive efficiency potentiates the positivity bias

Finally, positivity and confirmatory bias may be normative or advantageous in combination with other features of cognition. Supporting this idea, recent work proposes that learning biases are normative when coupled with efficient metacognition [66]. This is because when one can efficiently tease apart one's own correct decisions from one's mistakes, the probabilistic negative feedback (that sometimes inevitably follows correct choices) can be neglected. This creates a normative ground for positivity and confirmation biases. Note that this mechanism might not be restricted to humans, as efficient metacognition has been reported in animals, from non-human primates to rodents [67,68].

A challenge to this idea lies in the fact that learning biases and metacognitive (in)efficiencies might not be independent. Indeed, a yet unpublished study shows that in a two-armed bandit task where confidence in choice is elicited, the confirmation bias can cause overconfidence, which is a metacognitive bias [48]. While these findings challenge the idea that metacognition ensures that updating biases are normative, they might connect the asymmetric updating observed in RL to the original theoretical accounts of asymmetric belief updating, if overconfidence (i.e. the metacognitive illusion of accuracy) is considered self-serving per se, i.e. carries an ego-relevant utility [15,69].

In conclusion, although this section reviewed the evidence that learning asymmetry may be normative in some contexts – and as such may provide justification for its selection in that context – its persistence in contexts where it is unfavorable along with its lack of modulation in many circumstances reinforce the idea that learning

asymmetry constitutes a hardcoded learning bias [39,45,54,55]. A complementary perspective on the normativity of this bias could emerge from different modelling perspectives. For example, a recent unpublished study suggests that asymmetric updating can be derived from Bayesian-optimal principles [70].

Neuronal bases

Neural circuits for biased updating

An important question concerns the neurobiological bases of positivity and confirmatory bias in RL [71]. A prerequisite to answering this question is a consensus concerning the neural bases of RL, per se. The dominant hypothesis, stemming from the repeated and robust electrophysiological and pharmacological observations, postulates that reinforcement is instantiated by dopaminergic modulation of cortico-striatal synapses [72–75]. A neural model of biased (or asymmetric) updates then further requires that the neural channels for positive and negative prediction errors are dissociable. In line with this assumption, anatomically plausible neural network models of cortico-striatal circuits suggest that positive and negative reinforcements are mediated by specific sub-populations of striatal neurons, which exhibit different receptors with excitatory (D1) or inhibitory (D2) properties [76]. These models (as well as their more recent developments [77,78]), can therefore support, in principle, asymmetric updating, by implementing the processing of positive and negative reinforcements in different neurobiological pathways. Crucially, recent extensions of these models also account for the absence of biases following observational trials and its exacerbation induced by instruction priors [50,51].

A conceptually similar but structurally different neural network model put forward an alternative theory, which suggests a key computational role for meta-plasticity in the generation of update biases [79]. While the meta-plasticity framework does not necessitate the emergence of a positivity bias, this bias naturally emerges under most outcome contingencies and confirms its advantageous properties [62–64,80].

Neural signatures in human studies

Several lines of evidence suggest that the neurotransmitter dopamine and a basal-ganglia structure, the striatum, govern the relative sensitivity to positive and negative prediction errors. First, in both healthy and neurological patients, dopaminergic modulation affects the learning rate bias, such that higher dopamine is associated with a higher positivity bias [81–84]. Second, in healthy subjects, inter-individual differences in positivity bias are associated with higher striatal activation in response to rewards [29]. Inter-individual differences in the positive bias have also been associated with pupil dilation (another physiological proxy of neuromodulator activity during outcome presentation in classic two-armed bandit tasks) [85]. Finally, the choice-confirmation bias model supposes that positive and negative predictions associated, respectively, with obtained and forgone outcomes, are treated by the same learning rate as confirmatory signals. FMRI studies of two-armed bandit tasks with complete feedback (**Figure 1A**) confirm that obtained and forgone outcome signals are both encoded in the dopaminergic striatum, with opposite signs, thereby suggesting that the neurocomputational role currently attributed to this structure can be extended to accommodate the choice-confirmation bias without major structural changes [86,87].

Loss aversion versus loss neglect

Overall, the studies reviewed here suggest that in reinforcement learning, outcomes are processed in a choice-confirmatory manner. This bias takes the form of a selective neglect of losses (i.e., obtained punishments and forgone rewards) relative to gains (i.e., obtained rewards and forgone punishments) when updating outcome expectations. Superficially, this pattern seems in stark contrast with a vast literature in behavioral economics revolving around the notion of loss aversion [88]. According to loss aversion, prospective losses loom greater than corresponding gains in determining individuals' economic choices [89]. In the RL framework, this valuation asymmetry would directly translate into the negative prediction error having a larger relative influence on value expectation. Consequently, the choice-confirmation bias observed in RL does not align, at least prima facie, with dominant behavioural economics theories, potentially representing an additional instance of the experience-description gap [44,90] (**Figure 4**). However, a more in-depth consideration of the processes at stake may help reconcile these apparently contradictory findings. First, loss aversion pertains to the calculation of subjective decision values, while loss neglect, in the context of reinforcement learning, applies to the retrospective subjective assessment of experienced outcomes. It is well known that different heuristics and biases apply to expected and experience utilities [91,92]. Second, most of the findings reviewed here, although properly incentivized, use relatively small outcomes (primary or secondary). Evidence in behavioral science and economics suggests that the utility function may display specific features in the range of small amounts usually involved in reinforcement learning studies, making them unsuited to test – and to challenge – the general structure of loss aversion [93,94] (but note some recent studies claim that loss aversion also extends to small outcomes [95]). Finally, it is worth noting that prospective loss aversion and retrospective loss neglect, although superficially antithetic, provide complementary explanations for the status quo bias. While loss aversion would explain the bias by the fear of losing current assets [94,96,97], loss neglect rather posits that we disregard the feedback that suggests we made a wrong decision (**Box 1**). Retrospective loss neglect (or choice-confirmation bias), however, provides a putative, new computational explanation for the puzzling phenomenon of (pathological) gambling, which is difficult to accommodate with loss aversion (**Figure IC** in **Box 1**) [98,99].

Concluding remarks

The evidence reviewed here suggests that, contrary to what was previously thought [2,69], positivity and confirmation biases permeate reinforcement learning, leading to an over-optimistic estimation of outcome expectations. This results in characteristic behavioral consequences (**Box 1**), that may explain phenomena such as choice inertia (or status quo bias) and risky decision-making (gambling).

Empirical investigations of the choice-confirmation bias in reinforcement learning have mostly relied on inferring model parameters from choice data. Therefore, no matter how carefully this inferential process is carried out [100], it is still conceivable that a surrogate, spurious computational process is responsible for the observed patterns of behavioral and neurobiological results. While we believe the current competing interpretations are not supported by available experimental evidence (**Box 2**; [101–106]), future research should carefully combine model fitting and clever designs, to provide unambiguous evidence for the neuro-computational mechanisms of positivity and confirmatory biases [28].

Recently, a stream of studies from cognitive (neuro)science has described behavioral patterns consistent with this emerging account of positivity and confirmatory bias.

Indeed, confirmation bias was recently described in a simple perceptual task [107,108],. within the time-evolving dynamic of the decision [109]. Crucially, in this latter case, the act of choosing was critical to the expression of the bias [110]. These findings suggest that confirmation bias is not purely a reflection of a high-level reasoning bias, nor restricted to the domain of abstract, semantic beliefs.

In sum, a growing body of empirical studies in humans and animals reveal that the asymmetries that affect high-level belief updates are shared with more elementary forms of updates, notably in the form of the choice-confirmation bias observed in reinforcement-learning. Whether those update asymmetries are caused by shared neuro-computational mechanisms, or whether they have emerged independently in two separate pathways remains an open question (see also Outstanding Questions). Finally, at the conceptual level, it seems that important links between concepts of agency, metacognition and ego-relevance could help reconciliate fundamental aspects of belief and value update asymmetries.

Glossary

Belief-confirmation bias: the tendency to overweight or selectively sample information that confirms our own beliefs (‘what I believe is true’). Also referred to as prior-biased updating, belief perseverance, or conservatism, among other nomenclatures.

Bias: a feature of a cognitive process that introduces systematic deviations between state of the world and an internal representation

Choice-confirmation bias: the tendency to overweight information that confirms our own choice (‘what I did was right’).

Learning rate: a model parameter that traditionally indexes the extent to which prediction errors affect future expectations

Model comparison: collection of methods aimed at determining what is the best model in a given dataset combining model fitting and model simulations, to assess, respectively the falsifiability of the rejected models and the parsimony of the accepted one.

Model fitting: statistical method aimed at estimating the values of a model parameters that maximise the likelihood of observing the empirical data. Model fitting is not to be confounded with model comparison (see below)

Positivity bias: the tendency to overweight events with a positive affective valence. In the specific context of reinforcement learning it would consist in overweighting positive prediction errors (regardless of them being associated with chosen of forgone option).

Positivity bias is also sometimes referred to as the good-news bad-news effect or preference-biased updating.

Prediction error: the discrepancy between an expectation and the reality. In the context of reinforcement learning, prediction errors are defined as the difference between an expected and an obtained outcome and they therefore have a valence: they are positive when the outcome is better than expected, and they are negative when the outcome is worse than expected.

Figure legends

Figure 1: Typical behavioral task and computational reinforcement learning framework. (A) A typical trial of a two-armed bandit task. Both a partial and complete feedback condition are presented. Labels in black indicate the objective steps of the trial, while labels in grey indicate the corresponding hidden cognitive processes. **(B)** Box-and-arrow representation of a reinforcement learning model of a two-armed bandit task. The figure presents a complete feedback task, where both the obtained (i.e., following the **c**hosen option: R(c)) and forgone (i.e., following the **u**nchosen option: R(u)) outcomes are displayed. The figure also presents a 'full' model with a learning rate specific to each combination of prediction error valence (positive '+' or negative '-') and relation to choice ('c' vs. 'u') [45,46]. **(C)** A figure of how the learning rates of the full (i.e., a model for a different learning rate for any possible combination of outcome types and prediction error valences) model relate to the those of the confirmation bias model, which bundles together the learning rates for positive obtained and negative forgone (i.e., confirmatory - 'CON') prediction errors and the learning rates for negative obtained and positive forgone (i.e., disconfirmatory - 'DIS') prediction errors.

Figure 2: Reinforcement learning biases across tasks, species, and outcome types. (A) The panel displays learning rates from [33] plotted as a function of the nature of the outcomes used in the task (appetitive/money versus aversive/electric shocks), the volatility of option-outcome contingencies (stable versus volatility as in [31]) and the prediction error valence (positive '+' versus negative '-'). (**B and C)** The panel displays learning rates from [32] (**B**) and [54] (**C**) plotted as a function of the species (monkeys versus rats), the volatility of option-outcome contingencies and the prediction error valence (positive '+' versus negative '-'). **(D** and **E)** The panels display the choice confirmation bias. The figure displays the learning rates from [45] (Experiment 2 in the paper) of a full model (i.e. a model with a different learning rate for any possible combination of choices, outcomes and prediction error types) as a function of whether the outcome followed a free (or instrumental) or a forced (or observational) trial; whether the outcome was associated with the obtained or forgone option and, finally, the valence of the prediction error (positive '+' or negative '-'). The overall pattern is consistent with a choice-confirmation bias because positive obtained and negative forgone prediction errors are overweighed only if they follow a free choice **(D)**, but not after a forced choice (observation trial; **(E)**). Data visualization is as in [111]: horizontal lines represent the mean, the error bars represent the error of the mean, the box the 95% confidence interval. Finally, the colored area is the distribution of the individual points.

Figure 3: Optimality of the learning rate biases. The figure displays the simulation results recently reported in [62]. Performance of the model is expressed as the average reward per trial obtained from by the artificial agents and is indexed by a colored gradient so that the yellow represents the highest values. Artificial agents are simulated playing a two-armed bandit task, using an exhaustive range of model parameters (learning rates) and across different task conditions. 'Partial feedback' refers to simulations where only the feedback of the chosen outcome is disclosed to the agent, while 'complete feedback' refers to simulations where both the obtained and forgone outcomes are disclosed to the agents. 'Rich task' refers to simulations in which both options have an overall positive expected value, while 'poor task' indicates the opposite configuration. 'Stable task' refers to simulations featuring a good option (positive expected value) and a bad option (negative expected value), whose values do not change across time. On the contrary 'volatile task' refers to simulations in which the options switched from good to bad (and vice versa) three times during the learning period. Performance is plotted a function of the learning rates. Cells above the diagonal correspond to positivity bias ('partial feedback') or a confirmation bias ('complete feedback'). The cell with a black circle indicates the best possible unbiased (or symmetric) combination of learning rates (in terms of average reward per trial). Cells surrounded by black lines indicate the biased (or asymmetric) combinations of learning rates that obtain a higher reward rate compared to the best unbiased combination (see the original paper for more details; adapted with permission from [62]).

Figure 4: Loss aversion versus loss neglect. This figure exemplifies the crucial computation differences between 'loss aversion' and 'loss neglect'. The former applies to decisions between explicit (or described) options, often referred to also as 'prospects' and experimentally instantiated by lotteries. The latter applies to decisions between options (often experimentally instantiated by bandits) whose values have been learnt by trial-and-error (or experience). In the former case, the slope in the loss domain, which determines the relation between subjective and objective values, corresponds to the loss aversion parameter. In the latter case, the slopes in the positive and the negative domains determine the extent at which an option estimate (Q-value) is updated as a function of the prediction error; the slopes correspond to the learning rates for positive and negative prediction errors, respectively.

Figure I: Behavioural signatures of biased updates. The panels display the two-armed bandit task contingencies (top) and simulated choice rates (bottom) as function of the trial number with three different models (unbiased $\alpha_+ = \alpha_-$, positivity bias $\alpha_+ > \alpha_-$, and negativity bias $\alpha_+ < \alpha_-$). (**A**) Two-armed bandit task with stable contingencies and no correct response (top) and preferred choice rate (bottom). The preferred choice rate is defined as the choice rate of the option most frequently selected by the simulated subject - by definition, in more than 50% of trials [29,46] (**B**) Reversal learning task (top) and correct choice rate (bottom). (**C**) risk preference task (top) and risky choice rate (bottom). The curves are obtained simulating the corresponding models using a very broad range of parameters' values. For each task ('stable', 'reversal' and 'risk') and model ('unbiased', 'positivity' and 'negativity') we simulated 10000 agents; decisions were implemented using a softmax decision rule. The parameters were drawn from uniform distributions covering all possible values of learning rates to ensure the generality of the results. See github.com/spalminteri/valence_bias_simulations for full details.

Boxes

Box 1: Behavioral signatures of positivity-biased update

Here, we illustrate some behavioral signatures that have been associated with the positivity bias in standard reinforcement learning paradigms, focusing on two-armed bandit tasks with partial feedback (**Figure 1A**). A first signature associated with positivity bias is reported in "stable" bandits (i.e., situations where the option probabilities and values do not change), specifically in situation where there is no correct option [23]. In such situations, the positivity bias predicts the development of a preferred response rate to a much greater extent compared to the other learning rate patterns (**Figure Ia**). Another signature has been uncovered in "reversal" bandits, i.e., tasks where after some time the best option becomes the worst and vice-versa. In this situations, the positivity bias first generates a high correct response rate before reversal, then induces a reluctance to switch toward the alternative option in the second phase (post reversal) [62–64] (**Figure Ib**). Both the development of a higher preferred response rate and the reluctance to reverse can be broadly understood as manifestation of the fact that the positivity bias induces choice inertia. Here, the feedback that is supposed make us change our policy, is not taken into account [112]. A third signature of positivity bias, independent from the choice inertia phenomenon, comes from bandits designed to assess risk preferences, by contrasting a risky option (i.e. the option with variable outcome) to a safe one with similar expected value (**Figure Ic**). Crucially, in these kinds of bandits, the alternative patterns of learning rates (unbiased, $\alpha_+ = \alpha_-$; and negativity bias, $\alpha_+ < \alpha_-$) predict subjects to behave in a risk avoidant manner. Although prima facie counter-intuitive, this result can be understood by considering that outcome sampling can locally generate a negative expectation for the risky option, which may never be corrected (with partial feedback). On the other hand, the positivity bias predicts a certain degree of risk seeking behavior: a pattern that has often been observed in humans [55,113] (albeit sometimes in interaction with the valence of the decision frame) and frequently in non-human primates [90]. Finally, by inducing an over-estimation of reward expectations, both positivity and choice-confirmation biases mechanistically overestimate the subjective probability of making a correct choice. Not only is this prediction weakly confirmed by the observation of widespread patterns of over-confidence in reinforcement learning tasks, but recent results also suggest that individuals levels of overconfidence and confirmatory learning are correlated [47,48].

Box 2: Misidentifying asymmetric update

Assessing reinforcement learning update biases relies on estimating learning rates from choice data. Although the logic of such inference is intuitive, fitting and interpretating parameters remains some of the trickiest analytical steps in computational cognitive modelling [28,100,114]. Here, we discuss how the estimation of learning rates asymmetries can be affected by (or mistaken for) apparently neutral choices of model specification and the omission of alternative computational processes. For instance, although counter-intuitive, Q-values initialization markedly affect learning rate and learning bias estimates. The reason is that the first prediction error plays a very important role in shaping all subsequent responses, especially in designs involving a small number of trials and stable contingencies. For instance, pessimistic initializations (i.e., setting initial Q-values lower than the true default expectation) can counter, or reverse, genuine positivity biases, by artificially amplifying the size of the first positive prediction error. Consequently, it is not surprising that many of the papers reporting a negativity bias used pessimistic initialization [38–40] - although not all, see e.g. [59]. Since the effect of priors vanishes after few trials and in volatile environments, tasks featuring long learning phases and variable contingencies are particularly well suited to tease apart pessimistic initializations from positivity and confirmation biases [33].

It has also recently been proposed that positivity and confirmation biases may spuriously arise by fitting different learning rates to models including an explicit choice-autocorrelation term [104,112]. The choice autocorrelation term is usually modelled as a (fixed or graded) bias in the choice function toward the option that was previously chosen, and is thought to account for the development of a habitual processes [115]. Intuitively, both processes naturally lead to a similar escalation of choice-repetition, as successful learning increasingly identify the best option (see **Box 1**). Yet a crucial, conceptual difference is that the autocorrelation is independent of the outcome (i.e. of the prediction-error). A recent meta-analysis showed that in nine datasets the choice confirmation bias is still detectable despite the inclusion of a choice-autocorrelation term [103]. It can be further argued that in the context of the typically short learning task (less than 1 hour), developing a strong outcome-independent habit is unlikely. As a consequence, it is possible that studies fitting explicit choice autocorrelation, actually missed occurrences of positivity and confirmation bias [27,116–118]. Tasks contrasting a riskier and safer options can tease apart these competing accounts, because only the positivity and confirmation biases predict a preference for the riskier (high variance) options, see **Box 1** and [55,90].

Acknowledgements
SP and ML thank Germain Lefebvre, Nahuel Salem-Garcia, Valerian Chambon and Héloise Théro for stimulating discussions and for leading most of the experimental work that nurtured these ideas over the last years. SP and ML thank Zoe Koopmans for proof-reading the manuscript. SP and ML thank Alireza Soltani, Hiroyuki Ohta, Sonia Bishop, Christopher Gagne and Germain Lefebvre for providing material for the figures. SP is supported by the Institut de Recherche en Santé Publique (IRESP, grant number : 20II138-00), and the Agence National de la Recherche (CogFinAgent: ANR-21-CE23-0002-02; RELATIVE: ANR-21-CE37-0008-01; RANGE : ANR-21-CE28-0024-01). The Departement d'études cognitives is supported by the Agence National de la Recherche (ANR; FrontCog ANR-17-EURE-0017). ML is supported by an SNSF Ambizione grant (PZ00P3_174127) and an ERC Starting Grant (INFORL-948671).

Bibliography

1 Benjamin, D.J. (2019) Chapter 2 - Errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1* 2 (Bernheim, B. D. et al., eds), pp. 69–186, North-Holland
2 Sharot, T. and Garrett, N. (2016) Forming Beliefs: Why Valence Matters. *Trends Cogn. Sci.* 20, 25–33
3 Eil, D. and Rao, J.M. (2011) The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *Am. Econ. J. Microecon.* 3, 114–138
4 Kuzmanovic, B. *et al.* (2018) Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. *J. Neurosci.* 38, 7996–8010
5 Sharot, T. *et al.* (2011) How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* 14, 1475–1479
6 Klayman, J. (1995) Varieties of Confirmation Bias. In *Psychology of Learning and Motivation* 32 (Busemeyer, J. et al., eds), pp. 385–418, Academic Press
7 Nickerson, R.S. (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220
8 Eskreis-Winkler, L. and Fishbach, A. (2019) Not Learning From Failure—the Greatest Failure of All. *Psychol. Sci.* 30, 1733–1744
9 Staats, B.R. *et al.* (2018) Maintaining Beliefs in the Face of Negative News: The Moderating Role of Experience. *Manag. Sci.* 64, 804–824
10 Coutts, A. (2019) Good news and bad news are still news: experimental evidence on belief updating. *Exp. Econ.* 22, 369–395
11 Tappin, B.M. *et al.* (2017) The heart trumps the head: Desirability bias in political belief revision. *J. Exp. Psychol. Gen.* 146, 1143
12 Bénabou, R. and Tirole, J. (2016) Mindful Economics: The Production, Consumption, and Value of Beliefs. *J. Econ. Perspect.* 30, 141–164
13 Loewenstein, G. and Molnar, A. (2018) The renaissance of belief-based utility in economics. *Nat. Hum. Behav.* 2, 166–167
14 Sharot, T. *et al.* (2022) Why and when beliefs change: a multi-attribute value-based decision problem. *Perspect. Psychol. Sci.* in press,
15 Bénabou, R. and Tirole, J. (2002) Self-Confidence and Personal Motivation*. *Q. J. Econ.* 117, 871–915
16 Kuhnen, C.M. and Knutson, B. (2011) The Influence of Affect on Beliefs, Preferences, and Financial Decisions. *J. Financ. Quant. Anal.* 46, 605–626
17 Barron, K. (2021) Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains? *Exp. Econ.* 24, 31–58
18 Kuhnen, C.M. (2015) Asymmetric Learning from Financial Information. *J. Finance* 70, 2029–2062
19 Buser, T. *et al.* (2018) Responsiveness to feedback as a personal trait. *J. Risk Uncertain.* 56, 165–192
20 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement learning: An introduction*, Cambridge University Press.
21 Botvinick, M. *et al.* (2019) Reinforcement Learning, Fast and Slow. *Trends Cogn. Sci.* 23, 408–422

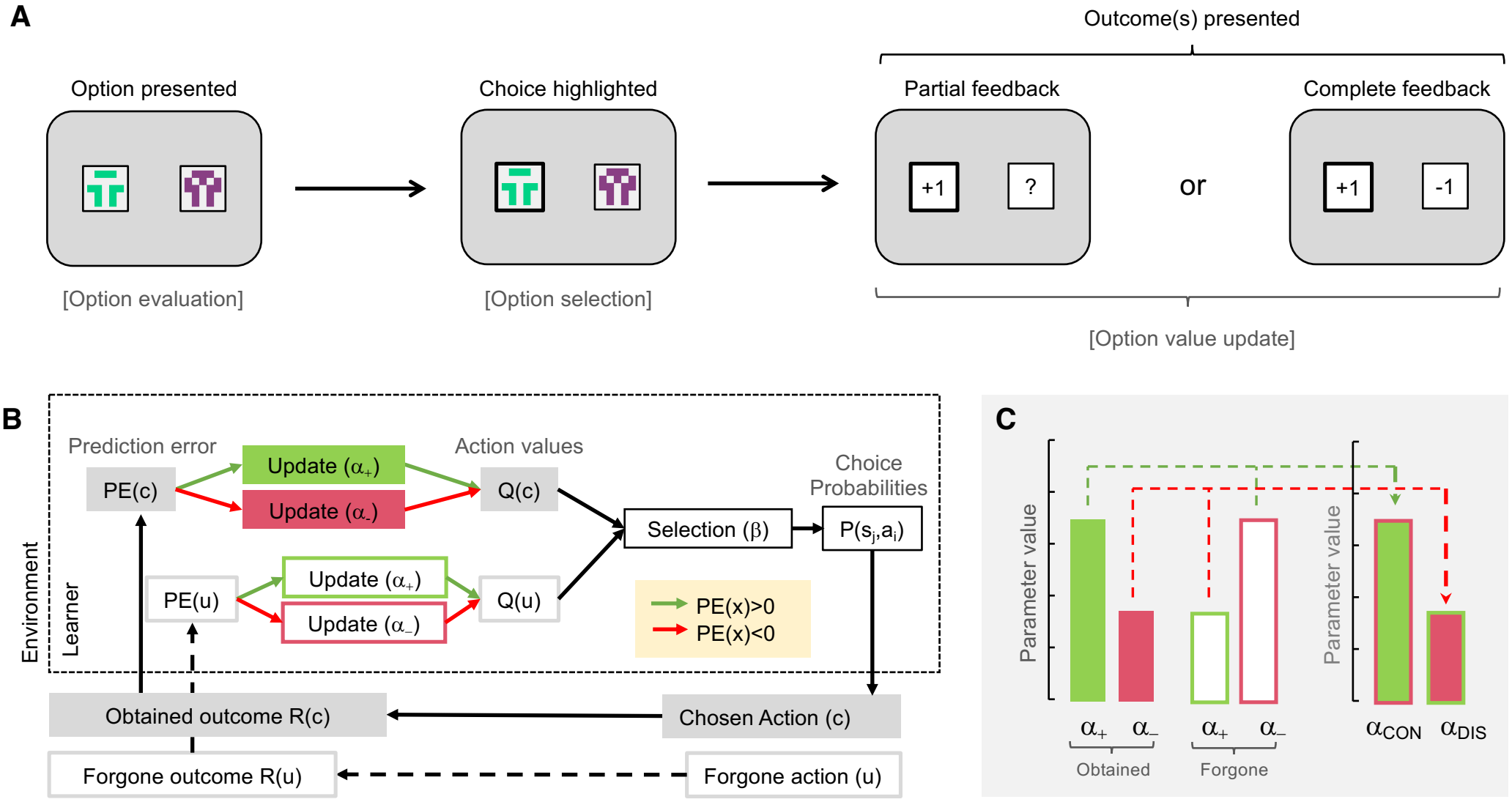
22 Hassabis, D. *et al.* (2017) Neuroscience-Inspired Artificial Intelligence. *Neuron* 95, 245–258
23 Aberg, K.C. *et al.* (2016) Linking Individual Learning Styles to Approach-Avoidance Motivational Traits and Computational Aspects of Reinforcement Learning. *PLOS ONE* 11, e0166675
24 Chase, H.W. *et al.* (2010) Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. *Psychol. Med.* 40, 433–440
25 Frank, M.J. *et al.* (2007) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl. Acad. Sci.* 104, 16311–16316
26 Kahnt, T. *et al.* (2009) Dorsal Striatal–midbrain Connectivity in Humans Predicts How Reinforcements Are Used to Guide Decisions. *J. Cogn. Neurosci.* 21, 1332–1345
27 den Ouden, H.E.M. *et al.* (2013) Dissociable Effects of Dopamine and Serotonin on Reversal Learning. *Neuron* 80, 1090–1100
28 Palminteri, S. *et al.* (2017) The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* 21, 425–433
29 Lefebvre, G. *et al.* (2017) Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* 1, 1–9
30 Ting, C.-C. *et al.* (2021) The elusive effects of incidental anxiety on reinforcement-learning. *J. Exp. Psychol. Learn. Mem. Cogn.* DOI: 10.1037/xlm0001033
31 Behrens, T.E.J. *et al.* (2007) Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221
32 Farashahi, S. *et al.* (2019) Flexible combination of reward information across primates. *Nat. Hum. Behav.* 3, 1215–1224
33 Gagne, C. *et al.* (2020) Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife* 9, e61387
34 Garrett, N. and Daw, N.D. (2020) Biased belief updating and suboptimal choice in foraging decisions. *Nat. Commun.* 11, 3417
35 Steinke, A. *et al.* (2020) Parallel model-based and model-free reinforcement learning for card sorting performance. *Sci. Rep.* 10, 15464
36 Nioche, A. *et al.* (2019) Coordination over a unique medium of exchange under information scarcity. *Palgrave Commun.* 5, 1–11
37 Ciranka, S. *et al.* (2022) Asymmetric reinforcement learning facilitates human inference of transitive relations. *Nat. Hum. Behav.* DOI: 10.1038/s41562-021-01263-w
38 Christakou, A. *et al.* (2013) Neural and Psychological Maturation of Decision-making in Adolescence and Young Adulthood. *J. Cogn. Neurosci.* 25, 1807–1823
39 Gershman, S.J. (2015) Do learning rates adapt to the distribution of rewards? *Psychon. Bull. Rev.* 22, 1320–1327
40 Niv, Y. *et al.* (2012) Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *J. Neurosci.* 32, 551–562
41 Pulcu, E. and Browning, M. (2017) Affective bias as a rational response to the statistics of rewards and punishments. *eLife* 6, e27879

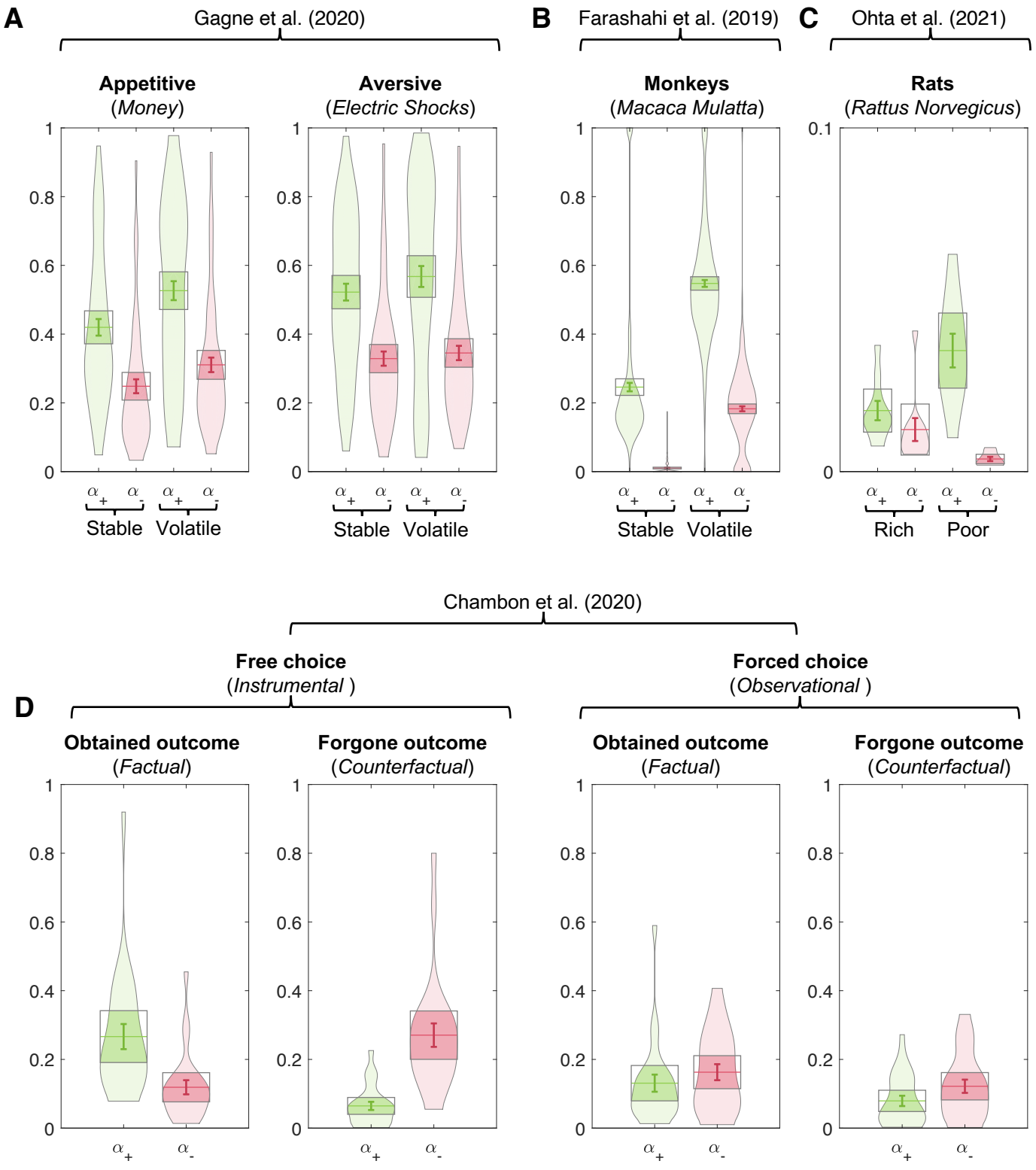
42 Wise, T. and Dolan, R.J. (2020) Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nat. Commun.* 11, 4179
43 Wise, T. *et al.* (2019) A computational account of threat-related attentional bias. *PLOS Comput. Biol.* 15, e1007341
44 Hertwig, R. and Erev, I. (2009) The description–experience gap in risky choice. *Trends Cogn. Sci.* 13, 517–523
45 Chambon, V. *et al.* (2020) Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nat. Hum. Behav.* 4, 1067–1079
46 Palminteri, S. *et al.* (2017) Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Comput. Biol.* 13, e1005684
47 Lebreton, M. *et al.* (2019) Contextual influence on confidence judgments in human reinforcement learning. *PLOS Comput. Biol.* 15, e1006973
48 Salem-Garcia, N.A. *et al.* (2021) The computational origins of confidence biases in reinforcement learning. *psyarxiv*
49 Schüller, T. *et al.* (2020) Decreased transfer of value to action in Tourette syndrome. *Cortex* 126, 39–48
50 Cockburn, J. *et al.* (2014) A Reinforcement Learning Mechanism Responsible for the Valuation of Free Choice. *Neuron* 83, 551–557
51 Doll, B.B. *et al.* (2009) Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res.* 1299, 74–94
52 Doll, B.B. *et al.* (2011) Dopaminergic Genes Predict Individual Differences in Susceptibility to Confirmation Bias. *J. Neurosci.* 31, 6188–6198
53 Harris, C. *et al.* (2020) Unique features of stimulus-based probabilistic reversal learning. *bioRxiv* DOI: 10.1101/2020.09.24.310771
54 Ohta, H. *et al.* (2021) The asymmetric learning rates of murine exploratory behavior in sparse reward environments. *Neural Netw.* 143, 218–229
55 Nussenbaum, K. *et al.* Flexibility in valenced reinforcement learning computations across development. . 16-Nov-(2021) , PsyArXiv
56 Chierchia, G. *et al.* Choice-confirmation bias in reinforcement learning changes with age during adolescence. . 06-Oct-(2021) , PsyArXiv
57 Habicht, J. *et al.* (2021) Children are full of optimism, but those rose-tinted glasses are fading—Reduced learning from negative outcomes drives hyperoptimism in children. *J. Exp. Psychol. Gen.* Advance online publication,
58 Xia, L. *et al.* (2021) Modeling changes in probabilistic reinforcement learning during adolescence. *PLOS Comput. Biol.* 17, e1008524
59 Rosenbaum, G.M. *et al.* (2022) Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. *eLife* 11, e64620
60 Cazé, R.D. and van der Meer, M.A.A. (2013) Adaptive properties of differential learning rates for positive and negative outcomes. *Biol. Cybern.* 107, 711–719
61 Gigerenzer, G. and Selten, R. (2002) *Bounded rationality: The adaptative toolbox*, Mit Press.

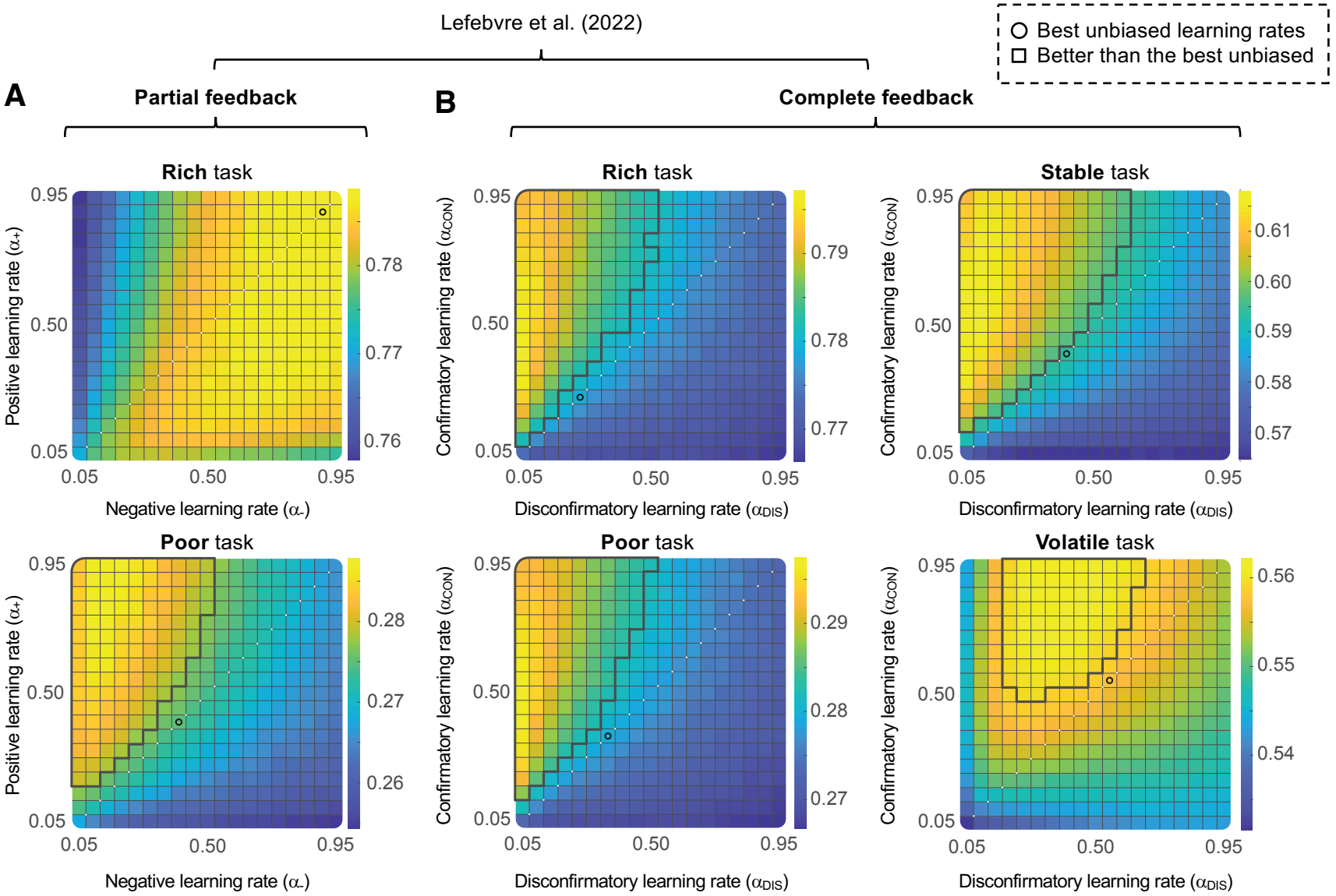
62 Lefebvre, G. *et al.* (2022) A Normative Account of Confirmation Bias During Reinforcement Learning. *Neural Comput.* 34, 307–337
63 Kandroodi, M.R. *et al.* (2021) *Optimal Reinforcement Learning with Asymmetric Updating in Volatile Environments: a Simulation Study*,
64 Tarantola, T. *et al.* (2021) Confirmation bias optimizes reward learning. *bioRxiv* DOI: 10.1101/2021.02.27.433214
65 Summerfield, C. and Tsetsos, K. (2020) Rationality and Efficiency in Human Decision-Making. *Cogn. Neurosci.*
66 Rollwage, M. and Fleming, S.M. (2021) Confirmation bias is adaptive when coupled with efficient metacognition. *Philos. Trans. R. Soc. B Biol. Sci.* 376, 20200131
67 Joo, H.R. *et al.* (2021) Rats use memory confidence to guide decisions. *Curr. Biol.* 31, 4571-4583.e4
68 Kepecs, A. and Mainen, Z.F. (2012) A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 1322–1337
69 Sharot, T. *et al.* Why and when beliefs change: A multi-attribute value-based decision problem. . 04-Nov-(2021) , PsyArXiv
70 Kobayashi, T. (2021) Optimistic Reinforcement Learning by Forward Kullback-Leibler Divergence Optimization. *ArXiv210512991 Cs* at <http://arxiv.org/abs/2105.12991>
71 Palminteri, S. and Pessiglione, M. (2017) Chapter 23 - Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. In *Decision Neuroscience* (Dreher, J.-C. and Tremblay, L., eds), pp. 291–303, Academic Press
72 Bayer, H.M. and Glimcher, P.W. (2005) Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* 47, 129–141
73 Dayan, P. (2012) Twenty-Five Lessons from Computational Neuromodulation. *Neuron* 76, 240–256
74 Di Chiara, G. (1999) Drug addiction as dopamine-dependent associative learning disorder. *Eur. J. Pharmacol.* 375, 13–30
75 Schultz, W. *et al.* (1997) A Neural Substrate of Prediction and Reward. *Science* 275, 1593–1599
76 Frank, M.J. (2006) Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Netw.* 19, 1120–1136
77 Collins, A.G.E. and Frank, M.J. (2014) Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* 121, 337–366
78 Swieten, M.M.H. van and Bogacz, R. (2020) Modeling the effects of motivation on choice and learning in the basal ganglia. *PLOS Comput. Biol.* 16, e1007465
79 Soltani, A. *et al.* (2006) Neural mechanism for stochastic behaviour during a competitive game. *Neural Netw.* 19, 1075–1090
80 Farashahi, S. *et al.* (2017) Metaplasticity as a Neural Substrate for Adaptive Learning and Choice under Uncertainty. *Neuron* 94, 401-414.e6
81 Frank, M.J. *et al.* (2004) By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science* 306, 1940–1943

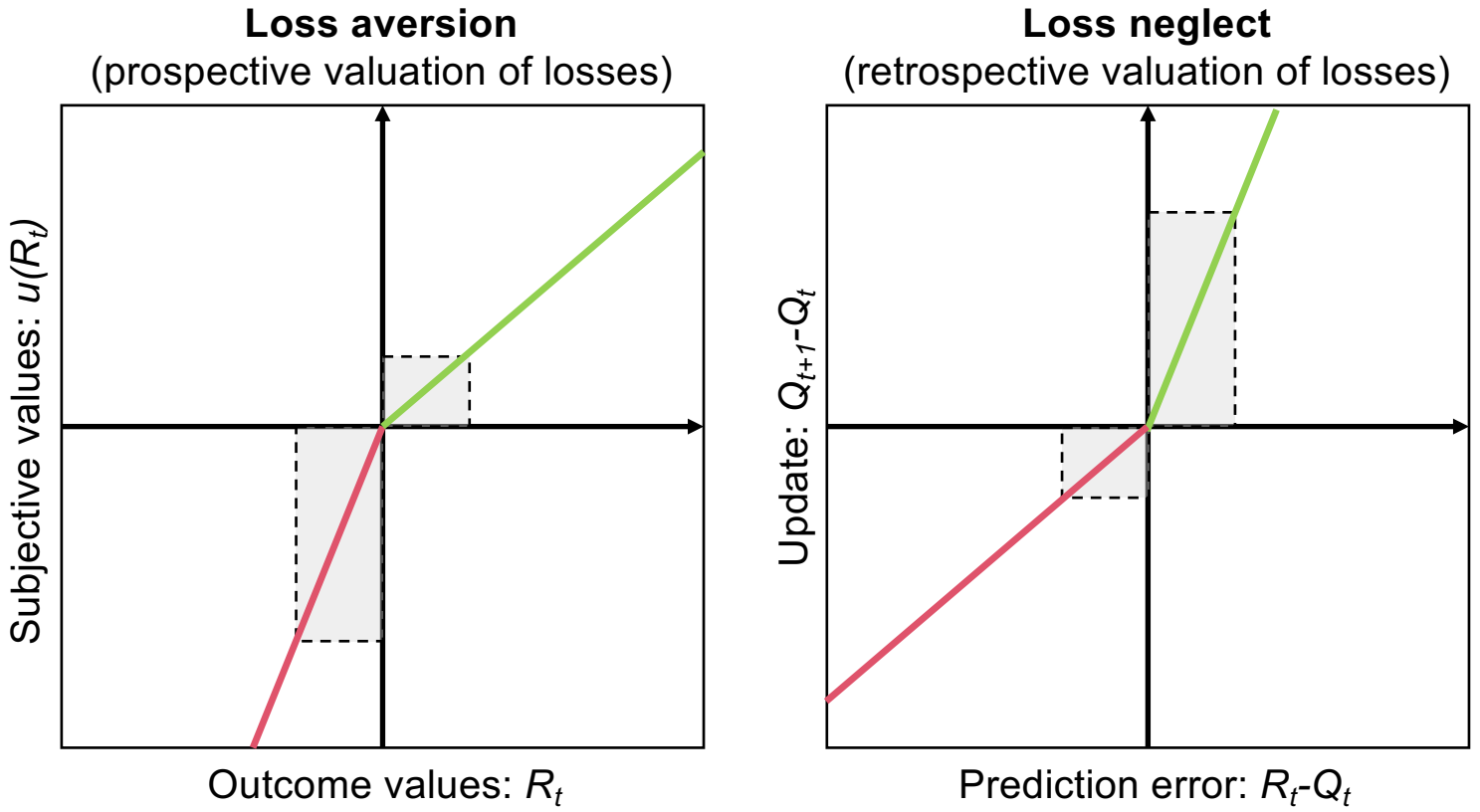
82 McCoy, B. *et al.* (2019) Dopaminergic medication reduces striatal sensitivity to negative outcomes in Parkinson's disease. *Brain* 142, 3605–3620
83 Palminteri, S. *et al.* (2009) Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. *Proc. Natl. Acad. Sci.* 106, 19179–19184
84 Pessiglione, M. *et al.* (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* 442, 1042–1045
85 Slooten, J.C.V. *et al.* (2018) How pupil responses track value-based decision-making during and after reinforcement learning. *PLOS Comput. Biol.* 14, e1006632
86 Li, J. and Daw, N.D. (2011) Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction. *J. Neurosci.* 31, 5504–5511
87 Klein, T.A. *et al.* (2017) Learning relative values in the striatum induces violations of normative decision making. *Nat. Commun.* 8, 16033
88 Ruggeri, K. *et al.* (2020) Replicating patterns of prospect theory for decision under risk. *Nat. Hum. Behav.* 4, 622–633
89 Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 263
90 Garcia, B. *et al.* (2021) The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty. *Philos. Trans. R. Soc. B Biol. Sci.* 376, 20190665
91 Kahneman, D. and Tversky, A. (2000) *Choices, Values, and Frames*, Cambridge University Press.
92 Kahneman, D. *et al.* (1997) Back to Bentham? Explorations of Experienced Utility. *Q. J. Econ.* 112, 375–406
93 Yechiam, E. (2019) Acceptable losses: the debatable origins of loss aversion. *Psychol. Res.* 83, 1327–1339
94 Anderson, C.J. (2003) The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychol. Bull.* 129, 139–167
95 Sokol-Hessner, P. and Rutledge, R.B. (2019) The Psychological and Neural Basis of Loss Aversion. *Curr. Dir. Psychol. Sci.* 28, 20–27
96 Jachimowicz, J.M. *et al.* (2019) When and why defaults influence decisions: a meta-analysis of default effects. *Behav. Public Policy* 3, 159–186
97 Kahneman, D. *et al.* (1991) Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *J. Econ. Perspect.* 5, 193–206
98 Fauth-Bühler, M. *et al.* (2017) Pathological gambling: a review of the neurobiological evidence relevant for its classification as an addictive disorder. *Addict. Biol.* 22, 885–897
99 Clark, L. *et al.* (2019) Neuroimaging of reward mechanisms in Gambling disorder: an integrative review. *Mol. Psychiatry* 24, 674–693
100 Wilson, R.C. and Collins, A.G. (2019) Ten simple rules for the computational modeling of behavioral data. *eLife* 8, e49547
101 Agrawal, V. and Shenoy, P. (2021) Tracking what matters: A decision-variable account of human behavior in bandit tasks. *Proc. Annu. Meet. Cogn. Sci. Soc.* 43,

102 Harada, T. (2020) Learning From Success or Failure? – Positivity Biases Revisited. *Front. Psychol.* 11, 1627
103 Palminteri, S. Choice-confirmation bias and gradual perseveration in human reinforcement learning. . 06-Jul-(2021) , PsyArXiv
104 Sugawara, M. and Katahira, K. (2021) Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Sci. Rep.* 11, 3574
105 Tano, P. *et al.* (2017) Variability in prior expectations explains biases in confidence reports. *bioRxiv* DOI: 10.1101/127399
106 Zhou, C.Y. *et al.* (2020) Devaluation of Unchosen Options: A Bayesian Account of the Provenance and Maintenance of Overly Optimistic Expectations. *CogSci Annu. Conf. Cogn. Sci. Soc. Cogn. Sci. Soc. US Conf.* 42, 1682–1688
107 Rajsic, J. *et al.* (2015) Confirmation bias in visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 41, 1353–1364
108 Rollwage, M. *et al.* (2020) Confidence drives a neural confirmation bias. *Nat. Commun.* 11, 2634
109 Talluri, B.C. *et al.* (2018) Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Curr. Biol.* 28, 3128-3135.e8
110 Talluri, B.C. *et al.* (2021) Choices change the temporal weighting of decision evidence. *J. Neurophysiol.* 125, 1468–1481
111 Bavard, S. *et al.* (2021) Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning. *Sci. Adv.* 7, eabe0340
112 Katahira, K. (2018) The statistical structures of reinforcement learning with asymmetric value updates. *J. Math. Psychol.* 87, 31–45
113 Madan, C.R. *et al.* (2019) Comparative inspiration: From puzzles with pigeons to novel discoveries with humans in risky choice. *Behav. Processes* 160, 10–19
114 Eckstein, M.K. *et al.* (2021) What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Curr. Opin. Behav. Sci.* 41, 128–137
115 Miller, K.J. *et al.* (2019) Habits without values. *Psychol. Rev.* 126, 292
116 Correa, C.M.C. *et al.* (2018) How the Level of Reward Awareness Changes the Computational and Electrophysiological Signatures of Reinforcement Learning. *J. Neurosci.* 38, 10338–10348
117 Gueguen, M.C.M. *et al.* (2021) Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans. *Nat. Commun.* 12, 3344
118 Voon, V. *et al.* (2015) Disorders of compulsivity: a common bias towards learning habits. *Mol. Psychiatry* 20, 345–352


A
Outcome(s) presented
Option presented
Choice highlighted
Partial feedback
Complete feedback
+1
?
or
+1
-1
[Option evaluation]
[Option selection]
[Option value update]
B
Prediction error
Action values
Choice Probabilities
PE(c)
Update (α+)
Update (α-)
Q(c)
Selection (β)
P(sj,ai)
PE(u)
Update (α+)
Update (α-)
Q(u)
PE(x)>0
PE(x)<0
Environment
Learner
Obtained outcome R(c)
Chosen Action (c)
Forgone outcome R(u)
Forgone action (u)
C
Parameter value
α+
α–
α+
α–
Obtained
Forgone
Parameter value
αCON
αDIS


A
Gagne et al. (2020)
Appetitive
(Money)
Aversive
(Electric Shocks)
B
Farashahi et al. (2019)
Monkeys
(Macaca Mulatta)
C
Ohta et al. (2021)
Rats
(Rattus Norvegicus)
Stable
Volatile
Rich
Poor
Chambon et al. (2020)
D
Free choice
(Instrumental)
Forced choice
(Observational)
Obtained outcome
(Factual)
Forgone outcome
(Counterfactual)


Lefebvre et al. (2022)
Best unbiased learning rates
Better than the best unbiased
A
Partial feedback
B
Complete feedback
Rich task
Positive learning rate (α+)
Negative learning rate (α-)
0.95
0.50
0.05
0.78
0.77
0.76
Rich task
Confirmatory learning rate (αCON)
Disconfirmatory learning rate (αDIS)
0.79
0.78
0.77
Stable task
0.61
0.60
0.59
0.58
0.57
Poor task
0.28
0.27
0.26
Poor task
0.29
0.28
0.27
Volatile task
0.56
0.55
0.54


Loss aversion
(prospective valuation of losses)
Subjective values: $u(R_t)$
Outcome values: R_t
Loss neglect
(retrospective valuation of losses)
Update: Q_{t+1}-Q_t
Prediction error: R_t-Q_t