Context-dependent outcome encoding in human reinforcement learning

Stefano Palminteri, Maël Lebreton

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Context-dependent outcome encoding in human reinforcement learning

Stefano Palminteri[1,2,3] and Maël Lebreton[4,5,6]

A wealth of evidence in perceptual and economic decision-making research suggests that the subjective assessment of one option is influenced by the context. A series of studies provides evidence that the same coding principles apply to situations where decisions are shaped by past outcomes, that is, in reinforcement-learning situations. In bandit tasks, human behavior is explained by models assuming that individuals do not learn the objective value of an outcome, but rather its subjective, context-dependent representation. We argue that, while such outcome context-dependence may be informationally or ecologically optimal, it concomitantly undermines the capacity to generalize value-based knowledge to new contexts — sometimes creating apparent decision paradoxes.

Addresses
[1] Laboratoire de Neurosciences Cognitives et Computationnelles, INSERM, Paris, France
[2] Département d'Etudes Cognitives, ENS, PSL Research University, Paris, France
[3] Institute for Cognitive Neuroscience, HSE, Moscow, Russian Federation
[4] Swiss Center for Affective Science, Geneva, Switzerland
[5] Laboratory for Behavioural Neurology and Imaging of Cognition, Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland
[6] Paris School of Economics, Paris, France

Corresponding author: Palminteri, Stefano (stefano.palminteri@ens.fr)

Current Opinion in Behavioral Sciences 2021, **41**:144–151

This review comes from a themed issue on **Value-based decision-making**

Edited by **Laura Bradfield** and **Bernard Balleine**

For a complete overview see the Issue and the Editorial

Available online 20th July 2021

https://doi.org/10.1016/j.cobeha.2021.06.006

Introduction

The view that perceptions, sensations and evaluations depend on their *context* was already a central tenant of the late 19th century's Gestalt psychology theory [1] and of early Utility theory [2]. A century later, the pervasiveness of perceptual illusions and decision-making biases, combined with decades of research in psychology, economics and neurosciences, consolidated the notion that perceptual and economic decisions are highly susceptible to contextual effects [3]. A significant fraction of contextual effects seems to result from two fundamental computations: reference-point centring and range adaptation [4–6].

In most ecological and real-life situations, decisions are arguably strongly influenced by the retrospective recollection of past outcomes experienced in similar situations [7]. Yet, in these experience-based decisions — realm of the reinforcement-learning framework — the notion of outcome context-dependence has been mostly neglected, until recent times [8,9]. Here, we review recent experimental work demonstrating that, in human reinforcement learning, outcomes are encoded and remembered as a function of the learning context.

By building on earlier work in perceptual decision-making, we consider outcome context-dependence as a manifestation of adaptive coding. Adaptive coding formalizes the idea that the (neural) representation of a variable is constrained by its underlying statistical distribution (i.e., the *context* [4,5]). Analogously, in reinforcement learning, outcome encoding is influenced by the distribution of outcomes experienced in the same or similar contexts.

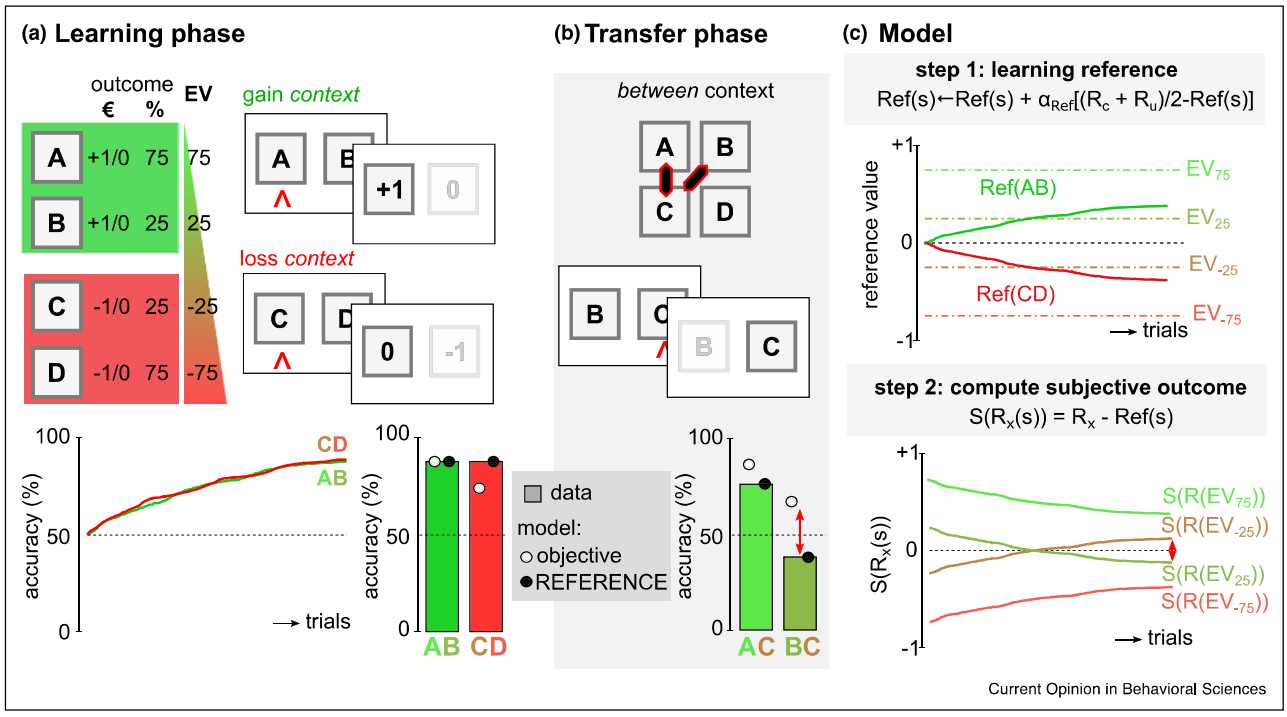
Outcome reference point-dependence in reinforcement learning

Harry Helson (1898–1977)'s adaptation-level (AL) theory constitutes the first systematic empirical investigation and theoretical formalization of the reference point-dependence of perceptual judgments [10]. AL theory postulates that perceptual features (such as luminosity, loudness and weight) are evaluated relative to a norm (or adaptation level) as follows:

$$J_i = S_i - \overline{S}$$

where J_i is the judgement of a particular stimulus i on a specific attribute, S_i is the objective value of the same stimulus in the perceptual attribute under consideration, and $\overline{S}$ is the norm, namely the arithmetic mean of all stimuli relevant to defining the context. The norm constitutes a *reference point*, usually defined as the running average of similar stimuli recently or simultaneously sampled, which is used as a point of comparison to judge the currently experienced stimulus (centring). By importing the AL core intuition into the realm of economic judgment and decision-making, Kahneman and Tversky proposed that the utility of an expected outcome does not

Figure 1



Reference point-dependence in RL: task, results and model variables.
(a) Learning phase contexts (top panel) and typical behavior (bottom panel). Subjects are presented for several trials with two learning contexts: AB (gain-maximization context) and CD (loss-minimization context). Feedback is probabilistic. Accuracy typically starts at chance level and progressively increases, reaching a similar plateau in both learning contexts. **(b)** Transfer phase contexts (top panel) and typical behavior (bottom panel). After the learning phase, symbols are re-arranged in new combinations. Here, we focus on the most informative combinations (AC and BC). The hallmark of outcome reference-point dependence is the preference for C over B in the BC comparison (green bar). While these behavioral signatures observed in both the learning and the transfer phase strikingly contrast with a model assuming objective outcome encoding (white dots), they are well captured by the REFERENCE model (black dots). Of note, choice pattern in the AC is also informative and indicates that the centering is only partial. **(c)** Evolution of the contextual variables (top panel) and subjective outcomes (bottom panel). The top panel illustrates the canonical temporal evolution of the reference points in the gain and loss contexts. Halfway through the learning phase, the reference points cross the expect value of the small gain/loss options. The bottom panel illustrates the resulting evolution of the average subjective outcomes for each option. Symmetrically to the top panel, roughly halfway through the learning phase, the subjective values of the outcomes of the EV_{25} and EV_{-25} options start to be subjectively 'perceived' as negative and positive, respectively.

reflect its objective value, but rather a sense of gain or loss, relative to a reference point. Reference-point dependence is therefore an intrinsic feature of prospect theory (PT [11,12]).

In a recent study, we tested if reference point-dependence affects the way outcomes are encoded (and stored in memory) in human reinforcement learning [13••]. Our behavioral paradigm joins a learning phase with a transfer phase [14,15]. Initially, during the learning phase, participants had to choose between options presented as fixed pairs of cues that were associated with a probabilistic outcome. The type of outcome defined the *learning context*: 'gain' (i.e. reward maximization) or 'loss' (i.e. punishment minimization) (Figure 1a). In the transfer phase, participants were required to express their option preference for each pairwise possible combination, including hybrid combinations of options from different learning contexts (Figure 1b). Two key behavioral results emerged: i) during learning phase, accuracy was well above chance and remarkably similar in the gain and the loss contexts; ii) option preferences in the transfer phase violated the strictly monotonic ranking dictated by their expected values (Figure 1a and b). More specifically, we found a significant preference for the small-loss option over the small-gain option. Crucially, these two key effects violate the predictions of outcome encoding by a standard Q-learning algorithm. In the learning phase, the standard model predicts lower performance in the loss condition: a phenomenon due to an intrinsic asymmetry in reinforcement rate in the gain and loss contexts (a.k.a. the punishment learning paradox [16–18]). In the transfer phase the standard model predicts a strictly monotonic ranking of option preferences as a function of their

Box 1 Reinforcement-learning models with outcome context dependence.

Both the REFERENCE [13••] and the RANGE [27••] models build on a standard Q-learning model, applied to a two-armed bandits task with complete feedback information [57]. For each pair of cues (i.e. state s), the REFERENCE model learns a reference point $Ref(s)$, often referred to as context-value or state-value $V(s)$, which is updated as follows:

$$Ref(s) \leftarrow Ref(s) + \alpha_{Ref} * \left(\frac{R(c) + R(u)}{2} - Ref(s) \right)$$

$R(c)$ and $R(u)$ are the outcome of the chosen and unchosen option respectively, while α_{Ref} is a learning rate ($0 < \alpha_{Ref} < 1$)[8] (see Figure 1c for the temporal evolution of the variable). $Ref(s)$ is then used to calculate the subjective outcome for each option a as follows:

$$s(R(a)) = R(a) - Ref(s)$$

The RANGE model infers two context-level variables: $Rmax(s)$ and $Rmin(s)$, which are updated as follows:

$$Rmax(s) \leftarrow Rmax(s) + \alpha_{Ran} * (\max(R(:)) - Rmax(s)),$$
$$if \max(R(:)) > Rmax$$
$$Rmin(s) \leftarrow Rmin(s) + \alpha_{Ran} * (\min(R(:)) - Rmin(s)),$$
$$if \min(R(:)) < Rmin$$

where $max(R(:))$ and $min(R(:))$ are respectively the highest and lowest possible outcomes observed in a given trial, while α_{Ran} is a learning rate ($0 < \alpha_{Ran} < 1$) (see Figure 2c for the temporal evolution of the variable). In this formulation, $Rmax(s)$/$Rmin(s)$ can only increase/decrease: it suits only tasks where the range does not change over time. The model could easily accommodate situations where the range changes over time, by simply assuming that $Rmax(s)$ is updated at a smaller rate when the observed outcome is smaller than the current estimation of $Rmax(s)$ (the opposite holds true for $Rmin(s)$) [23]. This variable is then used to calculate the subjective outcome for each option a as follows[9]:

$$s(R(a)) = \frac{R(a) - Rmin(s)}{Rmax(s) - Rmin(s) + 1}$$

Finally, both models assume that option values $Q(:,s)$ are updated following the standard update rule:

$$Q(a,s)_{t+1} = Q(a,s)_t + \alpha_A * (s(R(a)_t) - Q(a,s)_t)$$

Where α_A is a learning rate ($0 < \alpha_A < 1$). Both models make decisions with a standard 'softmax' decision rule with a fixed temperature parameter. These models have been shown to satisfactorily account for the behavioral patterns in both the learning phase and transfer phase (see Figures 1c and 2c), which falsify several plausible alternative formulations in reinforcement learning (actor-critic, habit learning [17,58]) and in neuroeconomics (subjective utility, divisive normalization [2,59]). As a final remark, even if the two models are not mathematically nested, the subtraction of $Rmin(s)$ at the numerator of the range normalization rules indicates that the RANGE model also implies the possibility that objectively negative outcomes can be reframed as subjectively positive.

objective values (see Box 1). By following the intuition of AL and PT theories, we proposed a model that learns the value of a reference-point and uses it to dynamically center the outcomes before computing the option-specific prediction error (Figure 1c). We refer to this model as the REFERENCE model. This model successfully explains symmetrical gain-loss performance in the learning phase and the suboptimal preference pattern in the transfer phase. Moreover, it outperforms the standard Q-learning model in a broad range of conditions, arguing in favor of outcome reference-point dependence in reinforcement-learning. This result has been replicated not only in our laboratory, but also in other studies and featuring different designs, including social learning [19] and different contingencies, options' arrangements and manipulations [20•,21,22].

Outcome range-adaptation in reinforcement learning

In the late 20th century, Allen Parducci revealed the presence of context-dependence in affective judgements of happiness, pleasure and pain, and formalized his findings in the range frequency (RF) theory [23]. Of particular interest to our review is Parducci's 'range principle', which describes the subjective judgement of a stimulus J_i as:

$$J_i = \frac{S_i - S_{min}}{S_{max} - S_{min}}$$

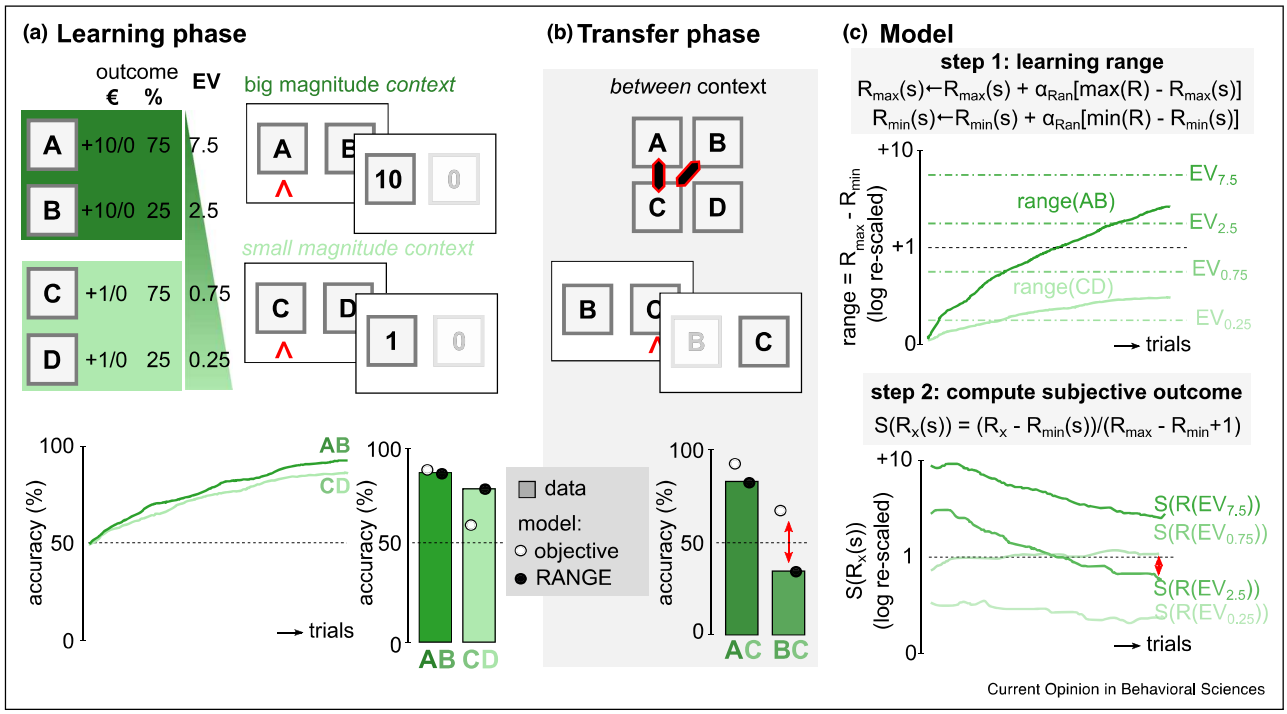
where S_i is the objective value of the stimulus i in the perceptual attribute under consideration, while S_{max} and S_{min} are the highest and lowest values presented in the relevant context, bounding the *range* of possible values. Essentially, the range principle states that subjective valuation is adapted to the underlying distribution of stimuli through a normalization rule. Recently, Kontek and Lewandoswky translated this idea into description-based decision-making by proposing the range-dependent utility model as an alternative to PT [24•]. The model assumes that the prospective valuation of the expected payoff of lotteries is range-adapted and accounts for several known behavioral paradoxes [25].

In a couple of recent studies, we tested if the range principle also applies to outcome encoding and retrospective retrieval from memory in reinforcement learning [26••,27••]. We built upon the previous behavioral paradigms to include systematic manipulation of outcome magnitudes, generating learning contexts with different outcome ranges. As in the previous study, the learning phase was followed by a transfer phase, which included new combinations of options (Figure 2a and b). Again, two key results emerged from these studies: i) accuracy was very similar in the small and the big magnitude contexts; ii) in the transfer phase, participants' choice-elicited preferences were not consistent with the objective outcome values. Notably, options that were locally

[8] Of note, the model proposed by Klein *et al.* [20•] is a special case of the REFERENCE model described (when $\alpha_{Ref} = 1$).

[9] In the denominator '+1' is added for computational convenience. It could be replaced by a free-parameter to account for task-specific differences, akin to a semi-saturation term governing the efficiency of the normalization [5].

Figure 2



Range adaption in RL: task, results and model variables.
(a) Learning phase contexts (top panel) a typical behavior (bottom panel). Subjects are presented for several trials with two learning contexts: AB (big-magnitude context) and CD (small magnitude context). Feedback is probabilistic. Accuracy typically starts at chance and progressively increases reaching a quite similar plateau in both learning contexts. **(b)** Transfer phase contexts (top panel) and typical behavior (bottom panel). After the learning phase, symbols are re-arranged in new combinations. Here, we focus on the most informative combinations (AC and BC). The hallmark signature of outcome range adaptation is the preference for C over B in the BC comparison (green bar). While these behavioral signatures observed in both the learning and the transfer phases strikingly contrast with a model assuming objective outcome encoding (white dots), they are well captured by the RANGE model (black dots). Of note, choice pattern in the AC is also informative, as it indicates that the range adaptation is only partial. **(c)** Evolution of the contextual variables (top panel) and subjective outcomes (bottom panel). The top panel illustrates the canonical temporal evolution of the ranges in the big and small magnitudes contexts. To the end of the learning phase, the ratio between the expected value of the options and the range values become similar in the big and small magnitude contexts. Crucially, R_{max} and R_{min} updates are conditional of $R > R_{max}$ and $R < R_{min}$, respectively (see Box 1). The bottom panel illustrates the evolution of the average subjective outcomes for each option. Notably, approximately halfway through the learning phase, the subjective value of the outcomes of the $EV_{2.5}$ and $EV_{0.75}$ cross over.

correct in the small magnitude contexts were systematically preferred to options that were locally incorrect in the big magnitude contexts, despite their objective expected values having the opposite ranking. A standard Q-learning model (with objective outcomes and softmax decision rule [28]) fails to predict this pattern, because its choice probabilities (and therefore accuracy) are strongly affected by the relative magnitudes of the option values. In line with RF theory, we proposed a model that learns the range of possible outcomes and uses it to dynamically rescale the outcomes before computing the option-specific prediction error (Figure 2c). This model, referred to as the RANGE model (see Box 1), satisfactorily captures the key behavioral effects. In our last study [27••], we also modulated the difficulty of the learning phase in two ways: by manipulating outcome information (partial versus complete feedback) and by manipulating the task structure (blocked versus interleaved trials). We found that outcome range adaptation was more pronounced in the easiest settings (block design, complete feedback), consistent with the idea that these manipulations enabled the participants to identify the context-relevant variables more easily. Crucially, as predicted by the RANGE model, this result was accompanied by a reduction in the subjects' ability to successfully extrapolate option values in the transfer phase. This finding is in striking opposition to the almost universally shared intuition that reducing task difficulty should lead, if anything, to more accurate and rational behavior [29,30].

Another recent study investigated choices in a reinforcement learning paradigm featuring repeated choices between a deterministic (i.e., risk-free) and a probabilistic (i.e., risky) option. Results showed that the outcome range

matters in subjective outcome values [31•]. Specifically, the authors convincingly demonstrated that risk preferences were strongly driven by an increased saliency of the extreme (i.e., the highest and the lowest possible) outcomes presented locally, in a given context, rather than being attached to any specific objective outcome value.

Dependence of irrelevant alternatives in human reinforcement learning

In the first part of this review, we provided evidence generalizing two manifestations of contextual influences to reinforcement learning: reference-point dependence and range normalization. However, the notion of context in psychology and economics is much richer, encompassing many other dimensions [32]. Particularly relevant for economic decision-making is the notion of the choice set or menu. In fact, standard normative theories assume that decision-makers (should) evaluate options in a way that the relative probability of choosing, say, option A over B, should be independent of the presence of a third alternative, say C (the so-called independence of irrelevant alternatives axiom, IIA [28]). Despite being intuitive, IIA is quite often violated in description-based decisions [33]. A recent study actually demonstrated that such contextual effects induced by the choice set also occur in reinforcement learning [34••]. The authors designed several behavioral paradigms, where participants choose between three options whose expected values were designed to elicit classic violations of IIA. However, the expected values were to be learned by trial-and-error. Their results reveal that similar outcomes 'inhibit' each other, whereas the most dissimilar (hence salient) outcomes 'stand out'. This bottom-up attentional bias can be captured by a computational model that decreases the subjective value of an outcome as a function of the cumulative sum of the perceived similarity between a given outcome and others concomitantly presented.[7] As a result, this study illustrates that contextual effects created by the choice set also extend to reinforcement learning scenarios. In these scenarios, contextual effects produce preference patterns that sometime oppose those observed in decision among lotteries, thus providing a new instance of the experience-description gap [8,9,35].

Relation to behavioral economics research and alternative computational frameworks

We reviewed converging evidence in support of the idea that the subjective value of an outcome is strongly influenced by the learning context, derived from the distribution of other and past outcomes [13••,19,20•,26••, 27••,31•,34••]. This body of work suggests that context plays a role in virtually all types of decision-making, possibly via the recycling of similar neural computational processes and constraints [3–5].

Contextual factors, such as the reference point, are central to theories of description-based decisions (such as PT). Although experimental evidence suggests that the reference point is dynamically updated by the choice history, the exact algorithm and mechanisms remain to be specified [36,37], thus weakening the theory [38]. Importing learning models into the decision-by-description framework and leveraging functional neuroimaging methods could prove useful in bridging this gap, both at the normative and descriptive levels [13••,39].

We also proposed that range adaptation can be implemented via a range normalization mechanism, based on the learned maximum and minimum possible outcomes [27••,31•]. Although reinforcement learning traditionally relies on behavioral paradigms featuring unidimensional outcomes (the 'numeric' reward), multi-attribute choice is another canonical situation in which the choice menu has been shown to be critical [33]. In this context, range normalization could apply to each attribute separately, generating and explaining the 'decoy' effects observed in classical description-based decisions with a computationally tractable mechanism [40].

Context effects in choice can be understood through an alternative computational formalism: the decision-by-sampling (DbS) framework [41]. The DbS framework supposes that the subjective value of an option comes from a series of ordinal comparisons between outcomes drawn from memory. Since subjective values come from comparisons with other options, context-effects arise naturally from the DbS formalism [42]. Furthermore, DbS could provide a unified framework for description-based (sampling from distributions) and experience-based (sampling from memory) decision-making. Particularly relevant for our treatment, a recent elegant study showed that DbS concomitantly generates range effects and achieves efficient coding of information [43•].

What are the functional roles of outcome context-dependence in reinforcement learning?

Converging evidence shows that outcome context-dependence systematically induces suboptimal choices when options are extrapolated beyond their original learning contexts in the transfer phase (Figures 1 and 2). This is particularly striking as similar behavioral findings have been found in distant species, such as rodents [44] and birds [45•,46]. Identifying predictable sources of biases is always puzzling, because evolutionary forces should have, in principle, negatively selected processes leading to suboptimal choices. Our work shows that context dependency can, of course, improve learning performance in specific conditions (loss avoidance, small magnitude).

[7] Of note, while the authors investigated only the complete feedback case, they argue that their model could be easily extended to the partial feedback case, by assuming that outcome comparison occurs between the presented outcome and previous outcomes stored in memory. To our knowledge, this hypothesis remains to be empirically verified.

However, most of these beneficial learning effects could be achieved by normalizing value signals at the choice phase, rather than at the learning and memorization phase, without bearing the costs of irrational preferences in the transfer phase. We speculate two possible functional roles for this learning bias. First, outcome context-dependence could simply result from adaptive and efficient (neural) coding principles, thereby optimizing information processing during learning [4,5]. Alternatively, while context-dependent learning induces suboptimal choices in our laboratory setting, they may be evolutionarily rational, meaning that they generate, on average, optimal performance in the environments where they evolved — for example, in environments where the resources are highly volatile [47,48].

Option value learning or direct policy learning?

A whole spectrum of models exists in the reinforcement learning framework, ranging from models assuming that expected values are learned for individual options (such as Q-learning), to models assuming that choice policies are learnt without intermediate option values representations (such as direct policy learning) [49•]. The latter hypothesis was backed up by evidence from a couple of studies on humans, where direct policy learning methods better explained subjects' choices in complete feedback tasks, at least in some critical trials [20•,50,51]. However, while the empirical data reviewed here clearly falsify the Q-learning's assumption that option-values are learned on a context-independent (or objective) scale, they also reject the equally extreme predictions of direct policy learning, by showing residual effect of outcome valence and magnitude in option value preferences (Figures 1 and 2) [13••,26••,27••]. We therefore favour a hybrid scenario where option-specific values are computed, but based on subjective outcomes that are encoded in a context-dependent manner.

Open questions

The present demonstration of context-dependent outcome encoding (Figures 1 and 2) relies on a combination of an instrumental learning phase and of a transfer phase eliciting preference as instrumental choices (e.g. in a procedural manner). Whereas recent evidence suggests that the Pavlovian learning system presents similar outcome encoding constraints [52], future studies should investigate address whether the same mechanism generalizes to other learning (Pavlovian, instrumental, goal directed) and representational (declarative, episodic) systems [53,54]. Finally, although we focused our review on situations, where context-dependent reinforcement learning concurrently benefits the learning phase and undermines generalization, an exhaustive investigation of learning and transfer environments could potentially identify situations where this trade-off can be tipped in favor of better generalization.

To conclude, investigating the effects of past outcomes on learning opens up a promising window, not only to define and formalize contextual effects (Box 2), but also to understand how the subjective, hedonic perception of outcomes shapes preferences. Deciphering the mechanisms and properties of reference-point dependence and range adaptation may also be key to appreciating the

Box 2 multiple definitions of context.

This quote from Parducci clearly illustrates how broadly the term *context* can be interpreted [23]:

"The term context refers to a conceptual representation of a set of events, real or imaginary, determining the dimensional judgement of any particular event" (p. 36).

In this box we clarify the meaning of *context* in the main studies reviewed here, building on an analogy to its definition in perceptual and value-based decision-making [6,60].

In visual cognition, 'spatial' context refers to objects simultaneously presented with a target stimulus. Effects due to the spatial context do not require the inference of hidden contextual variable because all relevant information is immediately available. Concering outcome encoding, the spatial context can be seen as the simultaneous presentation of multiple outcomes. While this definition of context becomes ambiguous when the outcome of the chosen option is displayed alone, it has been used to build models of complete feedback paradigms [20•,34••].

'Temporal' context refers to objects presented in the (more or less recent) past. This is how contextual effects are defined in the REFERENCE and the RANGE models [13••,27••]. In this case, outcome context-dependence is driven by hidden contextual variables (such as reference point, or the maximum possible reward), whose values are inferred from the history of past outcomes. This definition of context applies to both partial and feedback tasks, since it does not require simultaneous presentation of all feedback information. A corollary question concerns the time horizon for temporal context integration. Evidence suggests that contextual variables can be computed simultaneously over different time scales [61,62].

Both the 'spatial' and the 'temporal' definitions of context given above are implicit, that is, they are not attached to any particular cue. However, evidence suggests that contextual information can be attached to explicit value-neutral stimuli (e.g. visual cues, background colors) that can then be used to generalize contextual information to new options, without the need to experience the relevant outcomes [63,64•].

Finally, we mainly focused on external contexts (i.e. derived from stimuli). However, the internal state of an organism can also contribute to define a context. There is ample evidence that the level of satiation or the current energetic budget strongly influences memory and decision-making [65]. Accordingly, the recently developed framework of homeostatic reinforcement learning postulates that the subjective value of an outcome is determined by whether a given outcome moves toward (or away from) a homeostatic set-point [66,67•], defining an alternative formulation of outcome context-dependence in reinforcement learning.

It is important to disambiguate the term 'state', which has different meanings in different fields. In ethology and foraging research 'state' refers to the internal (physiological) status of the organism [47], while in animal and reinforcement learning literature, it mainly refers to a node in a Markov decision-process, roughly synonym of what we refer to as 'external context' [57].

neurophysiological encoding of learning and decision-related variables [13••,39,55•,56].

Conflict of interest statement

Nothing declared.

Acknowledgements

SP thanks Benedetto de Martino, Vincent Lenglin and Mikhail Spektor for helpful clarifications. SP and ML thank Sophie Bavard, Chih-Chung Ting, Nahuel Salem-Garcia and Laura Fontanesi for stimulating discussions and for leading most of the experimental work that nurtured these ideas over the last years. SP and ML thank Stefano Vrizzi for proof-reading the manuscript. SP is supported by an ATIP-Avenir grant (R16069JS), the Programme Emergence(s) de la Ville de Paris, the Fondation Fyssen, the Fondation Schlumberger pour l'Education et la Recherche and the Institut de Recherche en Santé Publique (IRESP, grant number: 20II138-00). The article was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2020-928) and the French National Agency of Research (ANR; FrontCog ANR-17-EURE-0017). ML is supported by an SNSF Ambizione grant (PZ00P3_174127) and an ERC Starting Grant (948671).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fechner GT: *Elemente der Psychophysik*. Breitkopf und Härtel; 1860.

2. Bernoulli D: **Specimen theoriae novae de mensura sortis.** *Comment Acad Sci Imp Petropolitanae* 1738, **5**:175-192.

3. Kahneman D: **Maps of bounded rationality: psychology for behavioral economics.** *Am Econ Rev* 2003, **93**:1449-1475.

4. Carandini M, Heeger DJ: **Normalization as a canonical neural computation.** *Nat Rev Neurosci* 2012, **13**:51-62.

5. Louie K, Glimcher PW: **Efficient coding and the neural representation of value.** *Ann N Y Acad Sci* 2012, **1251**:13-32.

6. Rangel A, Clithero JA: **Value normalization in decision making: theory and evidence.** *Curr Opin Neurobiol* 2012, **22**:970-981.

7. Rangel A, Camerer C, Montague PR: **A framework for studying the neurobiology of value-based decision making.** *Nat Rev Neurosci* 2008, **9**:545-556.

8. Garcia B, Cerrotti F, Palminteri S: **The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty.** *Philos Trans R Soc B Biol Sci* 2021, **376**:20190665.

9. Hertwig R, Erev I: **The description–experience gap in risky choice.** *Trends Cogn Sci* 2009, **13**:517-523.

10. Helson H: *Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior*. New York: Haper; 1964.

11. Kahneman D, Tversky A: **Prospect theory: an analysis of decision under risk.** *Econometrica* 1979, **47**:263.

12. Ruggeri K, Alí S, Berge ML, Bertoldo G, Bjørndal LD, Cortijos-Bernabeu A, Davison C, Demić E, Esteban-Serna C, Friedemann M *et al.*: **Replicating patterns of prospect theory for decision under risk.** *Nat Hum Behav* 2020, **4**:622-633.

13. •• Palminteri S, Khamassi M, Joffily M, Coricelli G: **Contextual modulation of value signals in reward and punishment learning.** *Nat Commun* 2015, **6**:8096
This study investigates reference-point dependence in human reinforcement learning, showing that the REFERENCE model represents an efficient computational solution to the punishment avoidance problem. The paper also present neurophysiological evidence supporting the computational assumptions of the REFERENCE model.

14. Frank MJ, Seeberger LC, O'Reilly RC: **By carrot or by stick: cognitive reinforcement learning in parkinsonism.** *Science* 2004, **306**:1940-1943.

15. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD: **Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans.** *Nature* 2006, **442**:1042-1045.

16. Moutoussis M, Bentall RP, Williams J, Dayan P: **A temporal difference account of avoidance learning.** *Netw Comput Neural Syst* 2008, **19**:137-160.

17. Maia TV: **Two-factor theory, the actor-critic model, and conditioned avoidance.** *Learn Behav* 2010, **38**:50-67.

18. Mowrer OH: *Learning Theory and Behavior*. John Wiley & Sons; 1960.

19. Burke CJ, Baddeley M, Tobler PN, Schultz W: **Partial adaptation of obtained and observed value signals preserves information about gains and losses.** *J Neurosci* 2016, **36**:10016-10025.

20. • Klein TA, Ullsperger M, Jocham G: **Learning relative values in the striatum induces violations of normative decision making.** *Nat Commun* 2017, **8**:16033
This study investigates context-dependent reinforcement learning in humans coupling behavioral and neural analyses. Transfer phase performance is undermined by outcome context-dependent learning as in [13,25,26]; Neurophysiological data are consistent with the notion of relative outcome encoding.

21. Lebreton M, Bacily K, Palminteri S, Engelmann JB: **Contextual influence on confidence judgments in human reinforcement learning.** *PLoS Comput Biol* 2019, **15**:e1006973.

22. Ting C-C, Palminteri S, Lebreton M, Engelmann J: **The elusive effects of incidental anxiety on reinforcement-learning.** *J Exp Psychol Learn Mem Cogn* 2021 http://dx.doi.org/10.1037/xlm0001033.

23. Parducci A: *Happiness, Pleasure, and Judgment: The Contextual Theory and its Applications*. Lawrence Erlbaum Associates, Inc.; 1995.

24. • Kontek K, Lewandowski M: **Range-dependent utility.** *Manag Sci* 2017, **64**:2812-2832
This study provides an axiomatic basis for a range-dependent utility model, adapting Parducci's range-frequency theory to decision-making under risk. The model is shown to account for several known decision paradoxes (including Allais' paradox).

25. Tversky A, Kahneman D: **Advances in prospect theory: cumulative representation of uncertainty.** *J Risk Uncertain* 1992, **5**:297-323.

26. •• Bavard S, Lebreton M, Khamassi M, Coricelli G, Palminteri S: **Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences.** *Nat Commun* 2018, **9**:4503
This study reveals the co-existence of reference-point dependence and range adaptation in human reinforcement learning. Among other results, it also shows that contextual biases are positively correlated with an explicit understanding of the task structure.

27. •• Bavard S, Rustichini A, Palminteri S: **Two sides of the same coin: beneficial and detrimental consequences of range adaptation in human reinforcement learning.** *Sci Adv* 2021, **7**:eabe0340
This study investigates range adaptation in human reinforcement learning. It provides evidence in favor of the RANGE model (described in Box 1) over eight experiment and by re-analyzing data from [25]. The paper also explores the paradoxical relation between learning and transfer phase performance.

28. Luce RD: *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation; 2012.

29. Day RH: **Rational choice and economic behavior.** *Theory Decis* 1971, **1**:229-251.

30. McFadden DL: **Rationality for economists?** *J Risk Uncertain* 1999, **19**:73-105.

31. • Ludvig EA, Madan CR, McMillan N, Xu Y, Spetch ML: **Living near the edge: how extreme outcomes and their neighbors drive risky choice.** *J Exp Psychol Gen* 2018, **147**:1905-1918

By investigating risk preferences in human reinforcement learning, this study found that they are not determined by the value of the outcomes *per se*, but rather they are driven by over representation in memory of the contextually more salient outcomes.

32. Louie K, De Martino B: **Chapter 24 - the neurobiology of context-dependent valuation and choice**. In *Neuroeconomics*, edn 2. Edited by Glimcher PW, Fehr E.Academic Press; 2014:455-476.

33. Busemeyer JR, Gluth S, Rieskamp J, Turner BM: **Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions.** *Trends Cogn Sci* 2019, **23**:251-263.

34. •• Spektor MS, Gluth S, Fontanesi L, Rieskamp J: **How similarity between choice options affects decisions from experience: the accentuation-of-differences model**. *Psychol Rev* 2019, **126**:52

This compelling study investigates classical IIA violations in human reinforcement learning and provides clear evidence for outcome context-dependence in several experiments. The results are consistent with an attentional bias, where similar outcomes are reciprocally inhibited.

35. Ert E, Lejarraga T: **The effect of experience on context-dependent decisions.** *J Behav Decis Mak* 2018, **31**:535-546.

36. Arkes HR, Hirshleifer D, Jiang D, Lim SS: **A cross-cultural study of reference point adaptation: evidence from China, Korea, and the US.** *Organ Behav Hum Decis Process* 2010, **112**:99-111.

37. Baucells M, Weber M, Welfens F: **Reference-point formation and updating.** *Manag Sci* 2011, **57**:506-519.

38. Baillon A, Bleichrodt H, Spinu V: **Searching for the reference point.** *Manag Sci* 2019, **66**:93-112.

39. Rigoli F, Friston KJ, Dolan RJ: **Neural processes mediating contextual influences on human choice behaviour.** *Nat Commun* 2016, **7**:12416.

40. Soltani A, Martino BD, Camerer C: **A range-normalization model of context-dependent choice: a new model and evidence**. *PLoS Comput Biol* 2012, **8**:e1002607.

41. Stewart N, Chater N, Brown GD: **Decision by sampling**. *Cognit Psychol* 2006, **53**:1-26.

42. Vlaev I, Chater N, Stewart N, Brown GD: **Does the brain calculate value?** *Trends Cogn Sci* 2011, **15**:546-554.

43. • Bhui R, Gershman SJ: **Decision by sampling implements efficient coding of psychoeconomic functions.** *Psychol Rev* 2018, **125**:985

This theory paper shows how decision-by-sampling can normatively emerge as a way to efficiently represent information in a noisy channel. The paper also shows that the resulting model displays behavior compatible with Parducci's range-frequency principle.

44. Flaherty CF: *Incentive Relativity*. Cambridge University Press; 1996.

45. • Pompilio L, Kacelnik A: **Context-dependent utility overrides absolute memory as a determinant of choice**. *Proc Natl Acad Sci U S A* 2010, **107**:508-512

To our knowledge, this study (in European starlings) is the first one to demonstrate context-dependent outcome encoding in an instrumental learning paradigm combining a learning (training) and a test (transfer) phase. This design served as a base for later studies in humans [26,27].

46. Vasconcelos M, Monteiro T, Kacelnik A: **Context-dependent preferences in starlings: linking ecology, foraging and choice**. *PLoS One* 2013, **8**:e64934.

47. McNamara JM, Trimmer PC, Houston AI: **The ecological rationality of state-dependent valuation.** *Psychol Rev* 2012, **119**:114.

48. McNamara JM, Fawcett TW, Houston AI: **An adaptive response to uncertainty generates positive and negative contrast effects.** *Science* 2013, **340**:1084-1086.

49. • Hayden BY, Niv Y: **The case against economic values in the orbitofrontal cortex (or anywhere else in the brain)**. *Behav Neurosci* 2021, **135**:192-201 http://dx.doi.org/10.1037/bne0000448

Thought-provoking opinion paper challenging the common assumption that the brain represents expected value and arguing in favor of heuristic decision-making and direct policy learning.

50. Li J, Daw ND: **Signals in human striatum are appropriate for policy update rather than value prediction.** *J Neurosci* 2011, **31**:5504-5511.

51. Hayes W, Wedell D: **Regret in experience-based decisions: the effects of expected value differences and mixed gains and losses**. *PsyArXiv Preprints* 2020 http://dx.doi.org/10.31234/osf.io/xaeyn.

52. Fontanesi L, Palminteri S, Lebreton M: **Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling.** *Cogn Affect Behav Neurosci* 2019, **19**:490-502.

53. Balleine BW, Daw ND, O'Doherty JP: **Chapter 24 - multiple forms of value learning and the function of dopamine.** In *Neuroeconomics.* Edited by Glimcher PW, Camerer CF, Fehr E, Poldrack RA. Academic Press; 2009:367-387.

54. Squire LR: **Memory systems of the brain: a brief history and current perspective**. *Neurobiol Learn Mem* 2004, **82**:171-177.

55. • Lebreton M, Bavard S, Daunizeau J, Palminteri S: **Assessing inter-individual differences with task-related functional neuroimaging**. *Nat Hum Behav* 2019, **3**:897-905

This perspective papers claims (and mathematically demonstrates) that hypotheses concerning the normalization of the outcome signals bear heavy consequences on how neural data (especially functional magnetic resonance) should be analyzed and interpreted in model-based fMRI approaches.

56. Cox KM, Kable JW: **BOLD subjective value signals exhibit robust range adaptation.** *J Neurosci* 2014, **34**:16533-16543.

57. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. Cambridge University Press; 1998.

58. Miller KJ, Shenhav A, Ludvig EA: **Habits without values**. *Psychol Rev* 2019, **126**:292.

59. Webb R, Glimcher PW, Louie K: **The normalization of consumer valuations: context-dependent preferences from neurobiological constraints**. *Manag Sci* 2020, **67**:93-125.

60. Louie K, Glimcher PW, Webb R: **Adaptive neural coding: from biological to behavioral decision-making.** *Curr Opin Behav Sci* 2015, **5**:91-99.

61. Zimmermann J, Glimcher PW, Louie K: **Multiple timescales of normalized value coding underlie adaptive choice behavior**. *Nat Commun* 2018, **9**:3206.

62. Holper L, Brussel LDV, Schmidt L, Schulthess S, Burke CJ, Louie K, Seifritz E, Tobler PN: **Adaptive value normalization in the prefrontal cortex is reduced by memory load**. *eNeuro* 2017, **4**.

63. Freidin E, Kacelnik A: **Rational choice, context dependence, and the value of information in European starlings (*Sturnus vulgaris*)**. *Science* 2011, **334**:1000-1002.

64. • Madan CR, Spetch ML, Machado FMDS, Mason A, Ludvig EA: **Encoding context determines risky choice**. *Psychol Sci* 2021, **32**:743-754

This paper clearly shows that information about outcome range can be attached to external cues (background images in this case), thus influencing the valuation of newly presented options. It also shows that contextual endpoints (Rmax and Rmin) are preferentially remembered.

65. Schuck-Paim C, Pompilio L, Kacelnik A: **State-dependent decisions cause apparent violations of rationality in animal choice**. *PLoS Biol* 2004, **2**:e402.

66. Juechems K, Summerfield C: **Where does value come from?** *Trends Cogn Sci* 2019, **23**:836-850.

67. • Keramati M, Gutkin B: **Homeostatic reinforcement learning for integrating reward collection and physiological stability**. *eLife* 2014, **3**:e04811

This paper presents a comprehensive computational framework of how reinforcement learning can be integrated with homeostatic principles, where outcomes are processed as a function of the internal state of the agent.