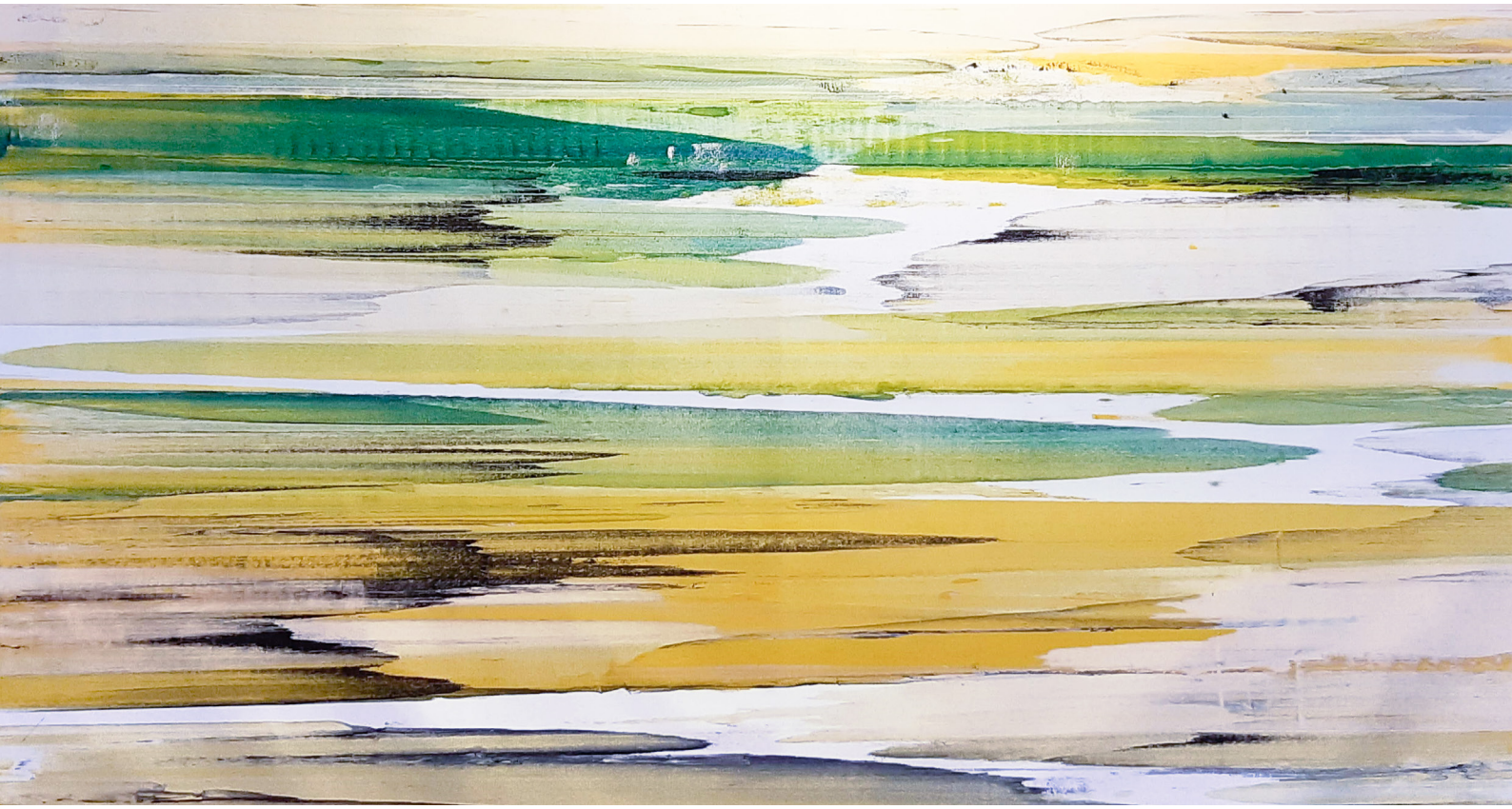
HSE UNIVERSITY

Human Capital
Multidisciplinary
Research Center

Context-Dependent Outcome Encoding in Human Reinforcement Learning

Research Digest № 8 (13) • 2022



The world-class Human Capital Multidisciplinary Research Center was founded in November 2020 as part of the National Science Project as a consortium of four leading organizations in human capital research: the National Research University Higher School of Economics, the Russian Presidential Academy of National Economy and Public Administration, the Moscow State Institute of International Relations (University) of the Ministry of Foreign Affairs of the Russian Federation, and the Russian Academy of Sciences N.N. Mikloukho-Maklay Institute of Ethnology and Anthropology.

The creation of the Center has become the largest initiative in Russia in the field of social sciences and humanities in recent decades. Among its main tasks are not only conducting world–class research in the field of human development, but also establishing cooperation with foreign leading organizations, launching educational programs, creating advanced scientific infrastructure, ensuring the transfer of the results obtained into the practice of public administration and education.

The Center implements 78 research projects. The research program covers key aspects of human potential that are relevant today on the global agenda:

Social Science and Humanities Aspects of Human Capital

Neurocognitive Mechanisms of Social Behavior

Demography and Active Aging

Natural and Climatic Factors Affecting Sustainable Development

Employment and Development of Skills and Competencies

Human Capital and Security in the Global Context

Humans in the Era of Technological Transformations

Research Digest is prepared within the project Context-Dependent Outcome Encoding in Human Reinforcement Learning.

Project supervisor: Olga Voron
Authors: S. Palminteri, M. Lebreton
Editor: A. Andrianova

Introduction

The view that perceptions 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 these 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] involving either description- or experience-based choices. In description-based paradigms, decision variables (i.e. payoffs and probabilities. 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 the 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 (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:

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

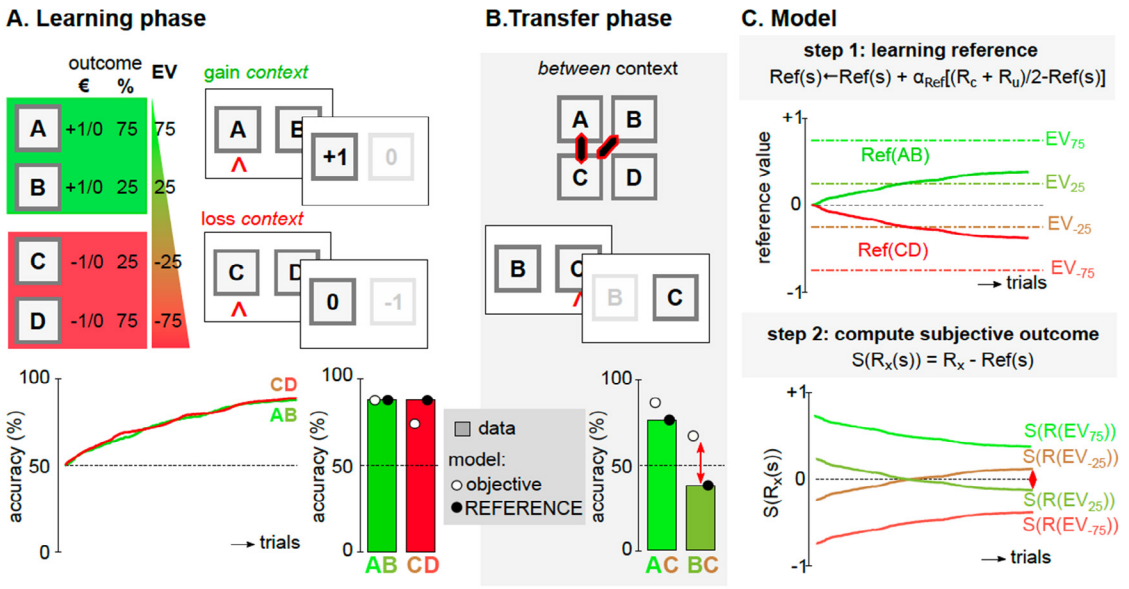
where J_i is the judgement of a particular stimulus i on a specific attribute, S_i is the objective value of the same stimulus in the perceptual attribute under consideration, and $\bar{S}$ is the norm, namely the arithmetic mean of all stimuli relevant to defining the context. The norm constitutes a *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 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] specifically in research on decision-making under risk. Kahneman and Tversky's 1979 study tested financial choices under risk, concluding that such judgements deviate significantly from the assumptions of expected utility theory, which had remarkable impacts on science, policy and industry. Though substantial evidence supports prospect theory, many presumed canonical theories have drawn scrutiny for recent replication failures. In response, we directly test the original methods in a multinational study (n = 4, 098 participants, 19 countries, 13 languages).

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-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 objective values. 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 option contingencies, arrangements and manipulations [*20–22].

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 value of the outcomes of the EV_{25} and EV_{-25} options started to be subjectively 'perceived' as negative and positive, respectively.

Outcome range-adaptation in reinforcement learning

In the late 20th century, Allen Parducci revealed the presence of context-dependence in affective assessments 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:

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

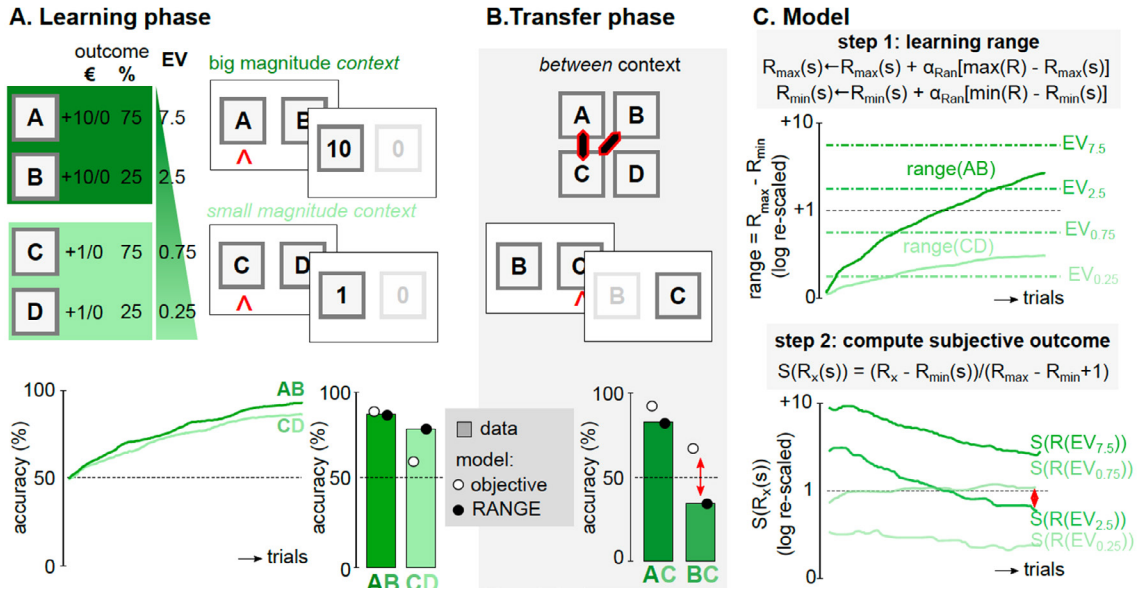
where S_i is the objective value of the stimulus i in the perceptual attribute under consideration, while S_{max} and S_{min} are the highest and lowest values presented in the relevant context, bounding the *range* of possible outcomes. Essentially, the range principle states that subjective valuation 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-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 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 satisfactorily captures the key behavioral effects. In our last study [27] range adaptation has been shown to lead to suboptimal choices, particularly notable in reinforcement learning (RL, we also modulated the difficulty of the learning phase in two ways: by manipulating outcome information (partial vs. complete feedback) and by manipulating the task structure (blocked vs interleaved design). 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] when people learn the odds and outcomes from experience, the extreme outcomes (best and worst. 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.

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. 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.

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 (Figure 1-2). Our work shows that context dependency can, of course, improve learning performance in specific conditions (loss avoidance, small magnitude). 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 – e.g. in environments where the resources are highly volatile [32, 33].

Open questions

The present demonstration of context-dependent outcome encoding (Figure 1 and Figure 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 [34], future studies should investigate address whether the same mechanism generalizes to other learning (Pavlovian, instrumental, goal directed) and representational (declarative, episodic) systems [35, 36]. 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.

Deciphering the mechanisms and properties of reference-point dependence and range adaptation may also be key to appreciating the neurobiological encoding of learning and decision-related variables [**13, *37, 38].

References

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; 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.
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.

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 Gen in press, doi:10.31234/osf.io/7d4tc.

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.

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.

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.

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.

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

33. McNamara JM, Fawcett TW, Houston AI: An Adaptive Response to Uncertainty Generates Positive and Negative Contrast Effects. Science 2013, 340:1084–1086.

34. 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.

35. 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.

36. Squire LR: Memory systems of the brain: A brief history and current perspective. Neurobiol Learn Mem 2004, 82:171–177.

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

38. Cox KM, Kable JW: BOLD Subjective Value Signals Exhibit Robust Range Adaptation. J Neurosci 2014, 34:16533–16543.