Science Advances
AAAS

Manuscript Template

FRONT MATTER

Title

The other side of the coin: the paradoxical consequences of range adaptation in human reinforcement learning

Authors
Sophie Bavard[1,2,3], Aldo Rustichini[4], Stefano Palminteri[1,2,3]

Affiliations
1 Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et Recherche Médicale, 29 rue d'Ulm 75005 Paris, FR.
2 Ecole normale supérieure, 29 rue d'Ulm 75005 Paris, FR.
3 Université de Recherche Paris Sciences et Lettres, 60 rue Mazarine 75006 Paris, FR.
4 University of Minnesota, 1925 4th Street South 4-101, Hanson Hall, Minneapolis, USA.

Abstract (150)

Evidence suggests that economic values are rescaled as a function of the range of the available options. Critically, although locally adaptive, range adaptation has been shown to lead to suboptimal choices. This is particularly striking in reinforcement learning (RL) situations when options are extrapolated from their original context. Range adaptation can be seen as the result of an adaptive coding process aiming at increasing the signal-to-noise ratio. However, this hypothesis leads to a counterintuitive prediction: decreasing task difficulty should increase range adaptation and, consequently, extrapolation errors. Here, we tested the paradoxical relation between range adaptation and performance in a large sample of participants performing variants of a RL task, where we manipulated task difficulty. Results confirmed that range adaptation induces systematic extrapolation errors and is stronger when decreasing task difficulty. Finally, we propose a range-adapting model and show that it is able to parsimoniously capture all the behavioral results.

MAIN TEXT

Introduction

In the famous Ebbinghaus illusion, two circles of identical size are placed near to each other, and larger circles surround one, while smaller circles surround the other. As a result, the central circle surrounded by larger circles appears smaller than the central circle surrounded by smaller circles, indicating that the subjective estimation of size of an object is affected by its the surroundings.

Beyond perceptual decision-making, wealth of evidence in neuroscience and in economics suggests that the subjective economic value of one option is not estimated in isolation, but is highly dependent of the context in which the options are presented (1,2). The vast majority of neuroeconomic studies of context-dependent valuation in humans considered situations where subjective values are triggered by explicit cues, that is stimuli whose value can be directly inferred, such as lotteries or snacks (3–5). However, in a series of recent papers, we and other groups demonstrated that contextual adjustments also permeate reinforcement learning situations, i.e., when option values have to be inferred from the history of past outcomes (6–8). We showed that an option, whose small objective value (7.5c) is learned in a context of smaller outcomes, is preferred to an option whose objective value (25c) is learned in a context of bigger outcomes, thus providing an economic equivalent of the Ebbinghaus illusion. Similar observations in birds suggest that this is a feature of decision-making broadly shared across vertebrates (9,10).

Although (as illustrated in the previous example) value context-dependence may lead to suboptimal decisions, it could be normatively understood as an adaptive process aimed at rescaling the behavioral response as a function of the range of the available options. Specifically, it could be seen as the result of an adaptive coding process aiming at increasing the signal-to-noise ratio by a system (the brain) constrained by the fact that behavioral variables have to be encoded by finite firing rates. In other terms, such *range adaptation* would be a consequence of how the system adjusts and optimizes the function associating the firing rate to the objective value to put its slope its maximum for each context (11,12).

If range adaptation is an automatic consequence of how the brain adapts its response to the distributions of the available outcomes, factors that facilitate the identifiability of these distributions should make it more pronounced. This would translate into a bigger difference between the objective option values (context-independent or *absolute*) and their corresponding subjective values (context-dependent or *relative*).

This leads to a counterintuitive prediction in the context of reinforcement learning. In fact, this is in striking contrast with the intuition embedded in virtually all learning algorithms, that making a learning-problem easier (by facilitating the identification of the outcome distributions) should, if anything, lead to more accurate and objective internal representations. In the present study, we aim at testing this hypothesis, while concomitantly gaining a better understanding of range adaptation at the computational level.

To empirically test this hypothesis, we build on previous research, and used a task featuring a *learning phase* and a *transfer phase* (6). In the learning phase, participants had to determine by trial-and-error the most favorable option in four fixed pairs of options (contexts), with different outcome ranges. In the transfer phase, the original options were rearranged, thus creating new contexts. This setup allowed us to quantify learning (or acquisition) errors during the first phase, and transfer (or extrapolation) errors during the second phase. Crucially, the task contexts were designed such that the correct responses in the transfer phase presented an overall higher expected value. We varied this paradigm in eight different versions where we manipulated the task difficulty in complementary ways. First, some of the experiments (E3, E4, E7, E8) featured *complete* feedback information, meaning that participants were informed about the outcome of the forgone option. This manipulation reduces task difficulty by resolving the uncertainty

concerning the counterfactual outcome. Accordingly, it has been repeatedly shown to improve learning performance (8,13). Second, some of the experiments (E5, E6, E7, E8) featured a block (instead of interleaved design), meaning that all the trials featuring one context were presented in a row. This manipulation reduces task difficulty by reducing working memory demand and has also been shown to improve learning performance (14). Finally, in some of the experiments (E2, E4, E6, E8), feedback was also provided in the transfer phase, thus allowing to assess if and how the values learned during the learning phase can be revised.

Behavioral analyses backed up our prediction and indicate that acquisition error rate in the learning phase is largely dissociable from extrapolation error rate in the transfer phase. Critically (and paradoxically), transfer phase error rate was higher when the learning phase was easier. Accordingly, the estimated deviation between the objective values and the subjective values increased in the complete feedback and block design tasks. The deviation was corrected only in the experiments featuring complete feedback in the transfer test.

To complement choice rate analysis, we developed a computational model that implements range adaption as a range normalization process, by tracking the maximum and the minimum possible reward in each learning context. Model simulations parsimoniously captured performance in the learning and the transfer phase, including the range adaptation-induced suboptimal preferences. Model simulations also allowed us to rule out alternative interpretations of our results that could come from two prominent psychological and economic theories: habit formation and risk aversion (15,16). Model comparison results were confirmed by checking out-of-sample likelihood as a quantitative measure of goodness of fit.

Results

Experimental protocol

We designed a series of learning and decision-making experiments involving variants of a main task. The main task was composed of two phases: the *learning* and the *transfer* phase. During the learning phase, participants were presented with eight abstract pictures, organized in four stable choice contexts. In the learning phase, each choice context featured only two possible outcomes: either 10pt/0pt or 1pt/0pt. The outcomes were probabilistic (75% or 25%) and we labeled the choices contexts as a function of the difference in expected value between the most and the least rewarding option: ΔEV=5 and ΔEV=0.5 (**Figure 1A**). In the subsequent transfer phase, the eight options were re-arranged into new choice contexts, where options associated with 10pt were compared to options associated with 1pt (see (7,10) for similar designs in humans and starlings). The resulting new four contexts were labeled ΔEV=7.25, ΔEV=6.75, ΔEV=2.25, ΔEV=1.75 (**Figure 1B**). In our between-subjects study, we developed eight different variants of the main paradigm where we manipulated whether we provided trial-by-trial feedback in the transfer phase (with / without), the quantity of information provided at feedback (partial: only the outcome of the chosen option is shown / complete: both outcomes are shown) and the temporal structure of choice contexts presentation (interleaved: choice contexts appear in a randomized order / block: all trials belonging to the same choice contexts are presented in a row) (**Figure 1C**). All the experiments reported in the main text were conducted online (N=100 participants in each version of the experiment); we report in the supplementary information the results concerning a similar experiment realized in the lab (see **Supplementary Materials**).

Overall correct response rate

The main dependent variable in our study was the correct response rate, i.e., the proportion of expected value-maximizing choices in the learning and the transfer phase (crucially our task design allowed to identify an expected value-maximizing choice in all choice contexts). In the learning phase, the average correct response rate was significantly higher than chance level 0.5 (0.69±0.16, $t(799) = 32.49$, $p < .0001$, $d = 1.15$; **Figure 2A-B**). Replicating previous findings, in the learning phase, we also observed a moderate but significant effect of the choice contexts, where the correct choice rate was higher in the ΔEV=5.0 compared to the ΔEV=0.5 contexts (0.71±0.18 vs 0.67±0.18; $t(799) = 6.81$, $p < .0001$, $d = 0.24$; **Figure 2C**)(6).

Correct response rate was also higher than chance in the transfer phase (0.62±0.17, $t(799) = 20.29$, $p < .0001$, $d = 0.72$, **Figure 2D-E**), but it was also strongly modulated by the choice context ($F(2.84,2250.66) = 271.68$, $p < .0001$, $\eta^2 = .20$, Huynh–Feldt corrected). In the transfer phase, the ΔEV=1.75 choice context is of particular interest, since the expected value maximizing option was the least favorable option of a ΔEV=5.0 context in the learning phase, and, conversely, the expected value minimizing option was the most favorable option of a ΔEV=0.5 context of the learning phase. In other words, for subject relying on expected values calculated in a context-independent scale, the $EV_{2.5}$ option is preferred compared to the $EV_{0.75}$ option. On the other side, a subject encoding the option values on a fully context-independent manner (which is equivalent to encode the rank between two options in a given context), will be perceive the $EV_{2.5}$ option as the less favorable option compared to the $EV_{0.75}$. Therefore, preferences in the ΔEV=1.75 context are diagnostic of

whether values are learned and encoded in an absolute or relative scale. Crucially, in the ΔEV=1.75 context, we found that participants' average correct choice rate was significantly *below* chance level (0.42±0.30, $t(799) = -7.25$, $p < .0001$, $d = -0.26$; **Figure 2F**), thus demonstrating that participants express suboptimal preferences in this context, i.e., they do not choose the option with the highest objective expected value.

Between-experiments comparisons: learning phase

In this section we analyze the correct response rate as a function of the experimental factors manipulated across the eight experiments (the quantity of provided information, that could be either partial or complete; the temporal structure of choice contexts presentation, that could be block or interleaved; and whether feedback was provided in the transfer phase). In the main text we report the significant results, but please see **Tables 1 and 2** for all results and effect sizes.

First, we analyzed the correct choice rate in the learning phase (**Figure 2B**). As expected, increasing feedback information had a significant effect on correct choice rate in the learning phase ($F(1,792) = 55.57$, $p < .0001$, $\eta_p^2 = .18$); similarly, performance in the block design experiments was significantly higher ($F(1,792) = 87.22$, $p < .0001$, $\eta_p^2 = .25$). We found a significant interaction between feedback information and task structure, reflecting that the difference of performance between partial and complete feedback was higher in block design ($F(1,792) = 5.05$, $p = .02$, $\eta_p^2 = .02$). We found no other significant main effect, double or triple interaction (**Table 1**).

We also analyzed the difference in performance between the ΔEV=5.0 and ΔEV=0.5 choice contexts across experiments (**Figure 2C**). We found a small but significant effect of temporal structure, the differential being smaller in the block compared to interleaved experiments ($F(1,792) = 7.71$, $p = .006$, $\eta_p^2 = .01$), and found no other significant main effect, nor interaction.

To sum up, as expected (8,13,14), increasing feedback information and clustering the choice contexts had a beneficial effect on correct response rate in the learning phase. Designing the choice contexts in blocks also blunted the difference in performance between the small (ΔEV=0.5) and big (ΔEV=5.0) magnitude contexts.

Between-experiments comparisons: transfer phase

We then analyzed the correct choice rate in the transfer phase (**Figure 2E**). Unsurprisingly, showing trial-by-trial feedback in the transfer phase led to significantly higher performance ($F(1,792) = 137.18$, $p < .0001$, $\eta_p^2 = .07$). Increasing feedback information from partial to complete also had a significant effect on transfer phase correct choice rate ($F(1,792) = 22.36$, $p < .0001$, $\eta_p^2 = .01$). Interestingly, we found no significant main effect of task structure in the transfer phase (see **Table 1**).

We found a significant interaction between feedback information and the presence of feedback in the transfer phase, showing that the increase in performance due to the addition of feedback information is higher when both outcomes were displayed during the learning phase ($F(1,792) = 20.18$, $p < .0001$, $\eta_p^2 = .01$). We also found a significant interaction between transfer feedback and task structure, reflecting that the increase in performance due to the addition of feedback information was even higher in block design ($F(1,792) = 42.22$, $p < .0001$, $\eta_p^2 = .02$). Finally, we found a significant triple interaction between feedback information, the presence of feedback in the

transfer phase, and task structure ($F(1,792) = 5.02$, $p = .03$, $\eta_p^2 = .003$). We found no other significant double interaction. We also separately analyzed the correct choice rate in the ΔEV=1.75 context (**Figure 2F**). Overall, the statistical effects presented a similar pattern as the correct choice rate across all conditions (see **Table 2**), indicating that overall correct choice rate and the correct choice rate in the key comparison ΔEV=1.75 provided a coherent picture. Furthermore, comparing the ΔEV=1.75 to chance level (0.5) revealed that participants, overall, significantly expressed *reward minimizing preferences* in this choice context. Crucially, the lowest correct choice rate was observed in the experiment featuring complete feedback, clustered choice contexts and no feedback in the transfer phase (E7; 0.27±0.32 $t(99) = -7.11$, $p < .0001$, $d = -0.71$); the addition of feedback in the transfer phase reversed the situation, since the only experiment where participants expressed reward maximizing preference was E8 (0.59±0.29, $t(99) = 2.96$, $p = .0038$, $d = 0.30$).

Between-phase comparison

Interestingly, we found a significant interaction between the phase (learning or transfer) and transfer feedback (without/with) on correct choice rate ($F(1,792) = 82.30$, $p < .0001$, $\eta_p^2 = .09$). This interaction is shown in **Figure 3** and reflects the fact that while adding transfer feedback information had a significant effect on transfer performance ($F(1,792) = 137.18$, $p < .0001$, $\eta_p^2 = .05$, **Figure 3A-B**), it was not sufficient to outperform learning performance (with transfer feedback: learning performance 0.69±0.16 vs transfer performance 0.68±0.15, $t(399) = 0.89$, $p = .38$, $d = 0.04$, **Figure 3B**).

Finally, close inspection of the learning curves revealed that in experiments where feedback was not provided in the transfer phase (E1, E3, E5 and E7), choice rates (and therefore option preferences) were stationary (**Figure 3A** and **Figure 3B**). This observation rules out the possibility that reduced performance in the transfer phase was induced by progressively forgetting the values of the options (in which case we should have observed a non-stationary and decreasing correct response rate).

In conclusion, comparison between the learning and the transfer phase reveals two inter-related and intriguing facts: i) despite the fact that the transfer phase happens immediately after an extensive learning phase, performance is, if anything, lower compared to the learning phase; ii) factors that improve performance (by intrinsically or extrinsically reducing task difficulty) in the learning phase have either no (feedback information) or a negative (task structure) impact on the transfer phase performance.

Inferred option values

To visualize and quantify how much observed choices deviate from the experimentally determined true option values, we optimized the four possible option values as free parameters.

More precisely, we initialized each subjective value at their true value (we labeled the four possible expected values as follows: $EV_{7.5}$, $EV_{2.5}$, $EV_{0.75}$, and $EV_{0.25}$), and optimized these values by gradient descent in order to maximize the likelihood of observing participants' choices using the logistic function (for, say, options $EV_{2.5}$ and $EV_{0.75}$):

$$P(\mathrm{EV}_{2.5}) = \frac{1}{1 + e^{(V(\mathrm{EV}_{0.75}) - V(\mathrm{EV}_{2.5}))}} \quad (1)$$

So that, if a participant chose indifferently between the $EV_{2.5}$ and the $EV_{0.75}$ option, their fitted values would be very similar: $V(EV_{2.5}) \approx V(EV_{0.75})$. Conversely, a participant with a sharp (optimal) preference for $EV_{2.5}$ over $EV_{0.75}$ would lead different fitted values: $V(EV_{2.5}) > V(EV_{0.75})$. In a first step, in the experiments where feedback was not provided in the transfer phase (E1, E3, E5 and E7), we optimized a set of subjective values per participant.

Consistent with the correct choice rate results described above, we found a value inversion of the two intermediary options ($EV_{2.5}$ 4.46±1.2, $EV_{0.75}$ 5.26±1.2, $t(399) = -7.82$, $p < .0001$, $d = -0.67$), which were paired in the ΔEV=1.75 context (**Figure 3C**). The differential was also strongly modulated across experiments ($F(3,396) = 18.9$, $p < .0001$, $\eta_p^2 = .13$, **Figure 3C**) and reached its highest value in E7 (complete feedback and block design).

As a second step, in the experiments where feedback was provided in the transfer phase (E2, E4, E6 and E8), we optimized a set of subjective values per trial. This fit allows us to estimate the trial-by-trial evolution of the subjective values over task time. The results of this analysis clearly show that suboptimal preferences progressively arise during the learning phase and disappear during the transfer phase (**Figure 3D**). However, the suboptimal preference was completely corrected only in E8 (complete feedback, block design) by the end of the transfer phase.

The analysis of inferred option values clearly confirms that participants' choices do not follow the true underlying objective monotonic ordering of the option values. Furthermore, it also clearly illustrates that in choice contexts that are supposed to facilitate the learning of the option values (complete feedback, block design), the deviation from monotonic ordering, at least at the beginning of transfer phase, is paradoxically greater. Monotonicity was fully restored only in E8, where complete feedback was provided in the transfer phase.

Computational formalization of the behavioral results

To formalize context-dependent reinforcement learning and account for the behavioral results, we designed a modified version of a standard model, where option-dependent Q-values are learnt from a range-adapted reward term. In the present study we implemented range adaptation as a range normalization process, which one among other possible implementations (17). At each trial *t*, the relative reward, $R_{\mathrm{RAN,t}}$, is calculated as follows:

$$R_{\mathrm{RAN,t}} = \frac{R_{\mathrm{ABS,t}} - R_{\mathrm{MIN,t}}(s)}{R_{\mathrm{MAX,t}}(s) - R_{\mathrm{MIN,t}}(s) + 1} \quad (2)$$

where *s* is the decision context (i.e., a combination of options) and R_{MAX} is a state-level variable, initialized to 0 and updated at each trial *t* if the outcome is greater than its current value. As such, R_{MAX} will converge to the maximum outcome value in each decision context, which in our task is either 1pt or 10pt. In the first trial $R_{\mathrm{RAN}} = R_{\mathrm{ABS}}$ (because $R_{\mathrm{MAX},0}(s) = 0$), and in later trials it is progressively normalized between 0 and 1 as the range value $R_{\mathrm{MAX}}(s)$ converges to its true value. Since in our task the minimum possible outcome is always zero $R_{\mathrm{MIN,t}}$ update was omitted while fitting the main experiments (but included in a ninth dataset analyzed below).

We refer to this model as the RANGE model and we compared it to a benchmark model (ABSOLUTE) which updates option values based the absolute reward values (note that the ABSOLUTE is nested within the RANGE model).

For each model, we estimated the optimal free parameters by likelihood maximization. We used the out-of-sample likelihood to compare goodness-of-fit and parsimony of the different models. To calculate the out-of-sample likelihood in the learning phase, the optimization was performed on half of the trials (one ΔEV=5.0 and one ΔEV=0.5 decision context) in the learning phase, and the best fitting parameters in this first set were used to predict choices in the remaining half of trials. In the learning phase, we found that the RANGE model significantly outperformed the ABSOLUTE model (out-of-sample LL_{RAN} vs. LL_{ABS}, $t(799) = 6.89$ $p < .0001$, $d = 0.24$, **Table 3**). To calculate the out-of-sample likelihood in the transfer phase, the optimization was performed on all trials of the learning phase and the best fitting parameters in the learning phase were used to predict choices in the transfer phase. Thus, the resulting likelihood is not only out-of-sample, but also cross-learning phase. This analysis revealed that the RANGE model outperformed the ABSOLUTE model (out-of-sample LL_{RAN} vs. LL_{ABS}, $t(799) = 8.56$, $p < .0001$, $d = 0.30$).

To study the behaviors of our computational model and confirm the behavioral reasons underlying the out-of-sample likelihood results, we simulated the two models (using the individual best fitting parameters)(18). In the learning phase, only the RANGE model managed to reproduce the observed correct choice rate. Specifically, the ABSOLUTE model predicts very poor performance in the ΔEV=0.5 context (ABS vs. data, $t(799) = -16.90$, $p < .0001$, $d = 0.60$, RAN vs. data, $t(799) = -1.79$, $p = .07$, $d = -0.06$, **Figure 4A**).

In the transfer phase, and particularly in the ΔEV=1.75 context, only the RANGE model manages to account for the observed correct choice rate, while the ABSOLUTE model fails (ABS vs. data $t(799) = 13.20$, $p < .0001$, $d = 0.47$, RAN vs. data $t(799) = 0.36$, $p = .72$, $d = 0.01$, **Figure 4C-D**). In general, the ABSOLUTE model tends to overestimate the correct choice rate in the transfer phase.

In addition to looking at the choice patterns, we submitted the RANGE model simulations to the inferred subjective option values analysis. Not only is the RANGE model able to capture the value inversion that we observed in the data, as well as the estimated options (RAN vs data, $t(799) = 1.55$, $p = .12$, $d = 0.06$, **Figure 4E**), but it is also able to predict its dynamic emergence and its trial-by-trial evolution (**Figure 4F**).

Ruling out habit formation

One of the distinguishing behavioral signatures of the RANGE model compared the ABSOLUTE one is the preference for the suboptimal option in the ΔEV=1.75 context. Since the optimal option in the ΔEV=1.75 context is not often chosen during the learning phase (where it is locally suboptimal), it could be argued that this result arises from taking decisions based on a weighted average between their absolute values and past choice propensity (a sort of *habituation* or choice trace). To rule out this interpretation, we fitted and simulated a version of a HABIT model, which takes decisions based on a weighted sum of the absolute Q-values and a habitual choice trace (16,19). The habitual choice trace component is updated with an additional learning rate parameter that gives a bonus to the selected action. Decisions are taken comparing option-specific decision-weights D_t :

$$D_t(s,c) = (1-\omega) * Q_t(s,c) + \omega * H_t(s,c) \quad (3)$$

where at each trial t, state s and chosen option c, ω is the arbiter, Q is the absolute Q-value H is the habitual choice trace component. The weight ω is fitted as an additional parameter (for ω=0 the model reduces to the ABSOLUTE model) and governs the relative influence of each controller.

We found that the HABIT model, similarly to the ABSOLUTE model, fails to perfectly match the participants' behavior, especially in the ΔEV=0.5 and ΔEV=1.75 contexts (**Figure 5A**). Indeed, in the learning phase, the addition of a habitual component is not enough to cope for the difference in option values, and therefore the model simulations in the transfer phase fail to match the observed choice pattern (**Figure 5B**). This is because the HABIT model encodes values on an absolute scale and does not manage to develop a strong preference for the correct response in the ΔEV=0.5 context, in the first place (**Figure 5A**). Thus, it does not carry a choice trace strong enough to overcome the absolute value of the correct response in the ΔEV=1.75 context (**Figure 5B, Supp. Figure 2 A-B**, **Table 3**).

To summarize, a model assuming absolute value encoding coupled with a habitual component could not fully explain observed choices in both the learning and transfer phase.

Ruling out diminishing marginal utility

One of the distinguishing behavioral signatures of the RANGE model is that it predicts very similar correct choice rates in the ΔEV=5.00 and the ΔEV=0.50 contexts compared to the behavioral data, while both the ABSOLUTE and the HABIT predict a huge drop in performance in the ΔEV=0.50 that directly stems from its small difference in expected value. It could be argued that this result arises from the fact that expected *utilities* (and not expected *values*) are learned in our task. Specifically, a diminishing marginal utility parameter would blunt differences in outcome magnitudes and would suppose that choices are made by comparing outcome probabilities. The process could also explain the preference for the suboptimal option in the ΔEV=1.75 context, since the optimal option in the ΔEV=1.75 context is rewarded (10pt) only the 25% of the time, while the suboptimal option is rewarded (1pt) 75% of the time. To rule out this interpretation, we fitted and simulated a UTILITY model, which updates Q-value based reward utilities calculated from absolute reward as follows:

$$R_{\mathrm{UTI,t}} = \left(R_{\mathrm{ABS,t}}\right)^{\nu} \quad (4)$$

where the exponent ν is the utility parameter ($0 < \nu < 1$, for ν =1 the model reduces to the ABSOLUTE model).

We found that the UTILITY model, similarly to the RANGE model, captures quite well the participants' behavior in the ΔEV=0.5 context (**Figure 5C**). However, concerning the transfer phase (especially the ΔEV=1.75 context), it fails to capture the observed pattern (**Figure 5C-D**). Additional analyses suggest that this is specifically driven by the experiments where the feedback was provided during the transfer phase (**Figure 5D**). Indeed, the static nature of the UTILITY fails to match the fact that the preferences in the ΔEV=1.75 context can be reversed by providing complete feedback (**Supp. Figure 2 C-D**). Model comparison showed that the RANGE model also outperformed the UTILITY model (out-of-sample LL_{RAN} vs. LL_{UTY}, $t(799) = 3.21$, $p = .001$, $d = 0.06$, **Table 3**). The comparison between the RANGE and the UTILITY model went in the same direction with a similar magnitude, reaching marginal statistical significance (out-of-sample LL_{RAN} vs. LL_{HAB}, $t(799) = 1.77$, $p = .07$, $d = 0.05$). To summarize, a model assuming diminishing marginal utilities could not fully explain observed choices in the transfer phase.

Sub-optimality of range-adaptation in our task

The RANGE model is computationally more complex compared to the ABSOLUTE model, as it presents an additional internal variable (R_{MAX}), which is learnt with a dedicated parameter. Here we wanted to assess whether this additional computational complexity really paid off in our task.

We split the participants according to the sign of out-of-sample likelihood difference between the RANGE and the ABSOLUTE model: if positive, the RANGE model better explains the participant's data (RAN>ABS), if negative, the ABSOLUTE model does (ABS>RAN). Reflecting our overall model comparison result, we found more participants in the RAN>ABS, compared to the ABS>RAN category (N=545 vs. N=255).

We found no main effect of winning model on overall (both phases) performance ($F(1,798) = 0.03$, $p = .87$, $\eta_p^2 = 0$). Interestingly, we found that while RANGE encoding is beneficial and allows for better performances in the learning phase, it leads to worst performance in the transfer phase ($F(1,798) = 187.3$, $p < .0001$, $\eta_p^2 = .19$, **Figure 6A**). In other terms, in our task, it seems that the learning phase and the transfer phase are playing the game tug of war: when performance are pulled in favor of the learning phase (RANGE model participants) this will be at the cost of the transfer phase (and vice versa).

A second question is whether overall in our study, behaving as a RANGE model revealed economically advantageous. To answer this question, we compared the final monetary payoff in the real data, following the simulations using the participant-level best fitting parameters. Consistently with the task design, we found that the monetary outcome was higher in the transfer phase than in the learning phase (transfer gains $M = 2.16 \pm 0.54$, learning gains $M = 1.99 \pm 0.35$, $t(799) = 8.71$, $p < .0001$, $d = 0.31$). Crucially, we found that the simulation of the RANGE model induces significantly lower monetary earnings (ABS vs RAN, $t(799) = 19.39$, $p < .0001$, $d = 0.69$, **Figure 6B**). This result indicates that despite being locally adaptive (in the learning phase), in our task, range adaptation is economically disadvantageous, thus supporting the idea that it is the consequence of automatic, uncontrolled, process.

Validation of range adaptation in previous dataset

Our main experiments only featured positive outcomes, in addition to 0. Since in our model the state-level variables (R_{MAX} and R_{MIN}) are initialized at 0, R_{MAX} converges to the maximum outcome value in each choice context, while R_{MIN} remains 0 in every trial and choice context. This is set-up is not ideal to test the full normalization rule we are proposing here. To obviate this limitation, we re-analyzed a ninth dataset (N=60) from a previously published study on a related topic (6). Crucially, in addition to an outcome magnitude manipulation ('10c vs '1€', similar to our learning phase), this study also manipulated the valence of the outcomes (gain vs loss). This latter manipulation allows to assess situations where the value of R_{MIN} can change and converge to a negative values, thus allowing us to compare the full range normalization rule to a simplified version:

$$\frac{R_{\mathrm{ABS}} - R_{\mathrm{MIN}}}{R_{\mathrm{MAX}} - R_{\mathrm{MIN}}} \quad \text{vs} \quad \frac{R_{\mathrm{ABS}}}{R_{\mathrm{MAX}}}$$

We later refer to the simplified version of the model as the RMAX model. Model simulations show that while the RMAX model can capture the learning and transfer patterns for the gain-related options, it fails to do so for the loss-related options (**Figure 5E-F**). Indeed, the lack of update of R_{MAX} in the loss contexts induces the RMAX model to encode values on an absolute scale (without normalization) and therefore, to fail to account for the range adaptation process. On the other hand, by updating both R_{MAX} in the gain contexts and R_{MIN} in the loss contexts, the RANGE model adapts with a full range in all contexts and is able to match participants' behavior even in the loss-related options as well as the gain-related options, in both the learning and transfer phases (**Figure 5E-F**). To conclude, this final analysis provides crucial support to the idea that range adaptation is consistent with a full range normalization rule, which takes into account both the maximum and the minimum possible outcomes.

Discussion

In the present paper we investigate context-dependent reinforcement learning, more specifically range adaptation, in a large cohort of human participants tested online over eight different variants of a behavioral task. Building on previous studies of context-dependent learning, the core idea of the task is to juxtapose an initial learning phase with fixed pairs of options (featuring either small or big outcomes) to a subsequent transfer phase where options are rearranged in new pairs (mixing up small and big outcomes)(6,7,10). In some experiments, we directly reduced task difficulty by reducing outcome uncertainty providing complete feedback. In some experiments, we indirectly modulated task difficulty by clustering in time the trials of a given contexts; therefore, reducing working memory demand. Finally, in some experiments, feedback was also provided in the transfer phase.

Behavioral findings

As expected, correct choice rate in the learning phase was higher when the feedback was complete, which indicates that participants integrated the outcome of the forgone option when it is presented (8,14). Also expectedly, in the learning phase participants displayed a higher correct choice rate when the trials of a given context were all blocked together, indicating that reducing working memory demands facilitate learning (15). Replicating previous findings, we also found that, overall, correct response rate was slightly but significantly higher in the big magnitude contexts ($\Delta EV=5.0$), but the difference was much smaller compared to what one would expect assuming absolute value learning and representation (as showed by the ABSOLUTE model simulations (6)): a pattern consistent with a *partial* range adaptation. The outcome magnitude-induced difference in correct choice rate was significantly smaller and not different from zero in block experiments (*full* adaptation), thus providing a first hint that reducing task difficulty increases range adaptation. Despite learning phase performance being fully consistent with our hypothesis, the crucial evidence comes from the results of the transfer phase. First, overall correct response rate pattern in the transfer phase did not follow that of the learning phase. Complete feedback and block design factors have no direct beneficial effect on transfer phase performance. In fact, if anything, the *worst* possible transfer phase performance was obtained in a complete feedback and block experiment. This was particularly striking in the $\Delta EV=1.75$ condition, where participants significantly preferred the suboptimal option and, again, worst score was obtained in a complete feedback and block design experiment. Second, we ruled out that the comparably low performance in the transfer phase

was due to having forgotten the value of the options. Indeed, since the transfer phase is, by definition, after the learning phase, although very unlikely (the two phases were only few seconds apart), it is conceivable that a drop in performance is due to the progressive forgetting of the option values. Two features of the correct choice rate curves allowed to reject this interpretation: i) correct choice rate abruptly decreases just after the learning phase; ii) when feedback is not provided the choice rate remains perfectly stable with no sign of regression to chance level. On the other side, i.e., when feedback was provided in the transfer phase, the correct choice rate increased to reach (on average) the level research at the end of the learning phase. The results are therefore consistent with the idea that in the transfer phase, participants express context-dependent option values acquired during the learning phase, which entails a first counterintuitive phenomenon: even if the transfer phase is performed immediately after the learning phase, the correct choice rate drops. This is due to the rearrangement of the options in new choice contexts, where options that were previously optimal solutions (in the small magnitude contexts) become suboptimal solutions. We also observed a second counterintuitive phenomenon: factors that increase performance during the learning phase (i.e., increasing feedback information and reducing working memory load), paradoxically further undermined transfer phase correct choice rate. The conclusions based on these behavioral observations were confirmed by inferring the most plausible option values based on the observed choices, where we could compare the objective ranking of the options to their subjective estimation. The only experiment where we observed an almost monotonic ranking was the partial feedback / interleaved experiment, even if we observed no significant difference between the EV=2.5 and the EV=0.75 options. In all the other experiments, the EV=0.75 option was valued more compared to the EV=2.5 option, with the highest difference observed in the complete feedback / block design. Thus, in striking opposition with the almost universally shared intuition that reducing task difficulty should lead to more accurate subjective estimates, here we present a rare instance where the opposite is true.

Computational mechanisms

The observed behavioral results were satisfactorily captured by a parsimonious model (the RANGE model) that instantiated a dynamic range normalization process. Specifically, the RANGE model learns in parallel context-dependent variables (R_{MAX} and R_{MIN}) that are used to normalize the outcomes. The R_{MAX} and R_{MIN} are learnt incrementally and the speed determines the extent of the normalization, leading to partial or full range adaptation as a function of the contextual learning rate. Developing a new model was necessary, as previous models of context-dependent reinforcement learning did not include range adaptation and focused on different dimensions of context-dependence (reference point-centering and outcome comparison) (7,8). The model also represents an improvement over a previous study where we instantiated partial arrange adaptation assuming a perfect and innate knowledge about the outcome ranges and a static hybridization between relative and absolute outcome values (6).

One limitation is that in present formulation R_{MAX} and R_{MIN} can only grow. Again, this is a feature that is well suited for our task, but may not correspond to many other laboratory-based and real-life situations, where the range can drift over time. This limitation could be overcome by assuming, for example, that R_{MAX} is also updated at a smaller rate when the observed outcome is smaller than the current R_{MAX} (the opposite could be true R_{MIN}). Finally, we note that our model applied to the main eight experiments (where R_{MIN} was irrelevant) can also be seen as special case of a divisive normalization process (temporal normalization (20)). To verify the relevance of the full range normalization rule, we re-analysed a previous dataset involving negative outcomes, where we were

able to show that both the R_{MAX} and R_{MIN} were important to explain the full spectrum of the behavioral results. Finally, it is worth noting that range normalization has been shown to perform poorly in explaining context-dependent decision-making in other (i.e., not reinforcement learning) paradigms (17,21,22), opening to the possibility that the normalization algorithm is different in experience-based and description-based choices. Future research contrasting different outcome ranges, multiple-options tasks are required to firmly determine which the functional form of normalization follows is better suited for experience-based and description-based choices (23).

We compared and ruled out another plausible computational interpretation derived from prominent psychological theory (24,25). First, we considered a habit formation model (16). We reasoned that our transfer phase results (and particularly the value inversion in the ΔEV=1.75 context) could derive from the participants choosing based on a weighted average between absolute values and past choice propensities. In fact, the suboptimal option in the ΔEV=1.75 context (EV=0.75) was chosen more frequently than the optimal option (EV=2.5) in the learning phase. However, model simulations showed that the HABIT model was not capable to explain the observed pattern. In fact, in the learning phase, the HABIT model, just like the ABSOLUTE model, did not develop a preference for the EV=0.75 option strong enough to generate a habitual trace sufficient to explain the transfer phase pattern. Beyond model simulation comparisons, we believe that this interpretation could have been rejected based on a priori arguments. The HABIT model can be conceived as a way to model habitual behavior, i.e., responses automatically triggered by stimulus-action associations. However, both in real life and laboratory experiments, habits have been shown to be acquired over time scales (days, months, year) order of magnitudes bigger compared to the timeframe of our tasks (26,27). Indeed, it is even debatable whether in our task participants developed even a sense of familiarity toward the (never seen before) abstract cues we used as stimuli. The HABIT model can also be conceived as a way to model choice hysteresis, sometimes referred to as choice repetition of perseveration bias, that could arise from a form of sensory-motor facilitation, where recently performance actions become facilitated (19,28). However, the screen position of the stimuli was randomized in a trial-by-trial basis; most of the experiments involved inter-leaved design, thus precluding any strong role for sensory-motor facilitation-induced choice inertia.

We compared and ruled out a plausible computational interpretation derived from economic theory (29). Since Bernoulli (1700-1782), risk aversion is explained by assuming diminishing marginal utility of objective outcomes (30). At the limit, if diminishing marginal utility was applied in our case, the utility of 10 points could be perceived as the utility of 1 point. In this extreme scenario, choices would be only based on the comparison between the outcome probabilities. This could explain most aspects of the choice pattern. Indeed, the UTILITY model did a much better job compared to HABIT model. However, compared to the RANGE model, it failed to reproduce the observed behavior of the experiments were feedback was provided in the transfer phase. This naturally results from the fact that the model assumes diminishing marginal utility as being a static property of the model and therefore cannot account for experience-dependent correction of context-dependent biases. However, also in this case, a priori considerations could have ruled out the UTILITY interpretation. Our experiment involves stakes small enough to make diminishing marginal utility not reasonable. Rabin provides a full treatment of this issue, and shows that the explaining risk aversion for small stakes (as those used in the laboratory) using diminishing marginal utility leads to extremely unlikely prediction, such as turning down gambles with infinite positive expected values (15). Indeed, following H. Markowitz intuition, most realistic models of the utility function suppose risk neutrality (or risk seeking) for small gains (31).

Our results contribute to the old and still ongoing debate about whether the brain computes option-oriented values independently from the decision-process itself (2,32). On one side of the spectrum, decision theories such as expected utility theory and prospect theory, postulate that a value is attached to each option independently of the other options simultaneously available (32). On the other side of the spectrum, other theories, such as regret theory, postulate that the value of an option is primarily determined by the comparison with other available options (33). A similar gradient exists in the reinforcement learning framework, between methods such as the Q-learning, on one side, and direct policy learning without value computations, on the other side (34). Recent studies in humans, coupling imaging to behavioral modeling, provided some support for direct policy learning in humans, by showing that, in complete feedback tasks, participants' learning was driven by a teaching signal, essentially determined by the comparison between the obtained and the forgone outcomes (regret/relief)(7,35). Beyond behavioral model comparison, analysis of neural activity in the ventral striatum (a brain system traditionally thought to encode option-specific prediction errors (36)), was also consistent with direct policy learning. However, while our findings clearly falsify the Q-learning's assumption that option-values are learned in an absolute (or context-independent) scale, model simulations also reject the other extreme view of direct policy learning (see **Supplementary Materials**). Indeed, our results are rather consistent with a hybrid scenario where option-specific values are initially encoded in an absolute scale and are progressively normalized to eventually represent the context-specific rank of each option. This view is also consistent with previous results using tasks including loss-related options that clearly showed that option valence was taken into account in transfer learning performance (6,8). In addition to that other studies illustrate that in similar paradigms, other behavioral measures, such as reaction times and confidence, are strongly affected by the learning context valence (valence is a construct that is absent in direct policy learning methods)(13,37). Finally, consistent with our intermediate view, other imaging studies found value related representations more consistent with a partial normalization process (38,39).

Conclusions

To conclude, we demonstrated that in humans, reinforcement learning values are learnt in a context-dependent manner that is compatible with range adaptation (instantiated as a range normalization process) (40). Specifically, we tested the possibility that these results from the way outcome information is automatically processed to achieve adaptive coding (41), by showing that the lower outcome uncertainty, the fuller range adaptation. This leads to a paradoxical result: reducing task difficulty can, in some occasions, decrease choice optimality. This surprising result can be understood with a perceptual analogy. Going into a dark room forces us to adapt our retinal response to dark, so that when we go back into a light condition we do not see very well. The longer we are exposed to dim light, the stronger the effect when we go back to normal.

Our findings fit in the debate aimed at deciding whether the computational processes leading to suboptimal decisions have to be considered flaws or feature of human cognition (42,43). Range adapting reinforcement learning is clearly adaptive in the learning phase. We could hypothesize that the situations in which the process is adaptive are more frequent in real life. In other terms the performance of the system has to be evaluated as a function of the tasks it has been selected to solve. It is true that we may be hit by a bus when we exit a dark room because we do not see well, but on average, the benefit of a sharper perception in a dark room is big enough to compensate for the (rare) event of a bus waiting for us outside the dark room. Ultimately, whether context-

474 dependent reinforcement learning should be considered a flaw or a desirable feature of human
475 cognition should be determined comparing the real life frequency of the situations where it is
476 adaptive (as in the learning phase) to that where it is maladaptive (as in the transfer phase). While
477 our study does not settle this issue, our findings do demonstrate that this process induce, at least in
478 some circumstances, economically suboptimal choices.

Materials and Methods

- **Participants**

For the laboratory experiment, we recruited 40 participants (28 females, aged 24.28±3.05 years old) via Internet advertising in a local mailing-list dedicated to cognitive science-related activities. For the online experiments, we recruited 8x100 participants (414 females, aged 30.06±10.10 years) from the Prolific platform (www.prolific.co). We based the online sample size on a power analysis that was based on the behavioral results of the lab experiment. In the ΔEV=1.75 context, lab participants reached a difference between choice rate and chance (0.5) of 0.11 ± 0.30 (mean±s.e.m.). To obtain the same with a power of 0.95, the Matlab function 'samsizepwr.m' indicated a value of 99 participants that we rounded to 100. The research was carried out following the principles and guidelines for experiments including human participants provided in the declaration of Helsinki (1964, revised in 2013). The Inserm Ethical Review Committee / IRB00003888 approved the study on November 13th, 2018 and participants were provided written informed consent prior to their inclusion. To sustain motivation throughout the experiment, participants were given a bonus depending on the number of points won in the experiment (average money won in pounds: 4.14±0.72, average performance against chance: M = 0.65±0.13, $t(799)=33.91$, $p<0.0001$). A laboratory-based experiment was originally performed (N=40) to ascertain that online testing would not significantly affecting the main conclusions. The results are presented in the **Supplementary Materials**.

- **Behavioral tasks**

Participants performed an online version of a probabilistic instrumental learning task adapted from previous studies (6). After checking the consent form, participants received written instructions explaining how the task worked and that their final payoff would be affected by their choices in the task. During the instructions the possible outcomes in points (0pt, 1pt and 10pt) were explicitly showed as well as their conversion rate (1pt=0.005£). The instructions were followed by a short training session of 12 trials aiming at familiarizing the participants with the response modalities. Participants could repeat the training session up to two times and then started the actual experiment.

In our task, options were materialized by abstract stimuli (cues) taken from randomly generated identicons, colored such that the subjective hue and saturation were very similar according to the HSL_{UV} color scheme (www.hsluv.org).On each trial, two cues were presented on both sides of the screen. The side in which a given cue was presented was pseudo-randomized, such that a given cue was presented an equal number of times on the left and the right. Participants were required to select between the two cues by clicking on one cue. The choice window was self-paced. A brief delay after the choice was recorded (500 ms), the outcome was displayed for 1000 ms. There was no fixation screen between trials. The average reaction time was 1.36±0.04 seconds (median: 1.16), the average experiment completion time was 325.24±8.39 seconds (median: 277.30).

As in previous studies, the full task consisted in one *learning* phase followed by a *transfer* phase (6–8,44). During the learning phase, cues appeared in four fixed pairs. Each pair was presented 30 times, leading to a total of 120 trials. Within each pair, the two cues were associated to a zero and a non-zero outcome with reciprocal probabilities (0.75/0.25 and 0.25/0.75). At the end of the trial, the cues disappeared and the selected one was replaced by the outcome ("10", "1", or "0") (**Figure**

1A). In experiments E3, E4, E7 and E8, the outcome corresponding to the forgone option (sometimes referred to as the *counterfactual* outcome) was also displayed (**Figure 1C**). Once they had completed the learning phase, participants were displayed with the total points earned and their monetary equivalent.

During the transfer phase after the learning phase, the pairs of cues were rearranged into four new pairs. The probability of obtaining a specific outcome remained the same for each cue (**Figure 1B**). Each new pair was presented 30 times, leading to a total of 120 trials. Before the beginning of the transfer phase, participants were explained that they would be presented with the same cues, only that the pairs would not have been necessarily displayed together before. In order to prevent explicit memorizing strategies, participants were not informed that they would have to perform a transfer phase until the end of the learning phase. After making a choice, the cues disappeared. In experiments E1, E3, E5 and E7, participants were not informed of the outcome of the choice on a trial-by-trial basis and the next trial began after 500ms. This was specified in the instruction phase. In experiments E2, E4, E6 and E8, participants were informed about the result of their choices in a trial-by-trial basis and the outcome was presented for 1000ms. In all experiments they were informed about the total points earned at the end of the transfer phase. In addition to the presence / absence of feedback, experiments differed in two other factors. Feedback information could be either partial (experiments E1, E2, E5, E6) or complete, (experiments E3, E4, E7, E8; meaning the outcome of the forgone option was also showed). When the transfer phase included feedback, the information factor was the same as in the learning phase. Trial structure was also manipulated, such that in some experiments (E5, E6, E7, E8), all trials of a given choice context were clustered ('blocked'), and in the remaining experiments (E1, E2, E3, E4) they were interleaved, in both the learning phase and the transfer phase (**Figure 1C**).

Re-analysis of a previous experiment

In the present paper we also include new analyses of previously published experiments (6). The general constructional principle of the previous experiments is similar to that used in the present experiments, as they involved a learning phase and a transfer phase. However, the previous design different from the present one in several important respects. First, in addition to an outcome magnitude manipulation ('10c vs '1€', similar to our learning phase), this study also manipulated the valence of the outcomes (gain vs loss), generating to a 2x2 factorial design. Second, the organization of the transfer phase was quite different. Indeed, each option was compared with all other possible options. The mean dependent variable extracted from the transfer phase is therefore not the correct response rate, but simply choice rate per option (which is proportional to its subjective value). The data were pooled across two experiments featuring partial (N=20) and partial and complete feedback trials (N=40). In both experiments the choice contexts were interleaved. Other differences include the fact that these previous experiments were laboratory-based and featured a slightly different number of trials, different stimuli and timing (see the original publication for more details).

- **Analyses**

Behavioral analyses.

The main dependent variable was the *correct* choice rate, i.e., choices directed toward the option with the highest expected value. Statistical effects were assessed using multiple-way repeated measures ANOVAs with choice context (labeled in the manuscript by their difference in expected values: ΔEV) as within-subject factor, and feedback information, feedback in the transfer phase and task structure as between-subjects factors. Post-hoc tests were performed using one-sample and two-sample t-tests for respectively within- and between-experiment comparisons. To assess overall performance, additional one sample t-tests were performed against chance level (0.5).We report the *t*-statistic, *p*-value, and Cohen's *d* to estimate effect size (two-sample t-test only). Given the large sample size (N=800), central limit theorem allows us to assume normal distribution of our overall performance data and to apply properties of normal distribution in our statistical analyses, as well as sphericity hypotheses. Concerning ANOVA analyses, we report the uncorrected statistical, as well as Huynh–Feldt correction for repeated measures ANOVA when applicable (45), *F*-statistic, *p*-value, partial eta-squared η_p^2 and generalized eta-squared η^2 (when Huynh-Feldt correction is applied) to estimate effect size. All statistical analyses were performed using Matlab (www.mathworks.com) and R (www.r-project.org). For visual purposes, learning curves were smoothed using a moving average filter (span of 5 in Matlab's *smooth* function).

- **Models**

We analyzed our data with variation of simple associative learning models (46,47). The goal of all models is to estimate in each choice context (or *state*) the expected reward (R) of each option and pick the one that maximizes this expected reward *R*.

At trial *t*, option values of the current context *s* are updated with the delta rule:

$$Q_{t+1}(s,c) = Q_t(s,c) + \alpha_c \delta_{c,t} \quad (5)$$

$$Q_{t+1}(s,u) = Q_t(s,u) + \alpha_u \delta_{u,t} \quad (6)$$

where α_c is the learning rate for the chosen (c) option and α_u the learning rate for the unchosen (u) option, i.e., the counterfactual learning rate. δ_c and δ_u are prediction error terms calculated as follows:

$$\delta_{c,t} = R_{c,t} - Q_t(s,c) \quad (7)$$

$$\delta_{u,t} = R_{u,t} - Q_t(s,u) \quad (8)$$

δ_c is calculated in both partial and complete feedback contexts and δ_u is calculated in the experiments with complete feedback only.

We modelled participants' choice behavior using a softmax decision rule representing the probability for a participant to choose one option *a* over the other option *b*:

$$P_t(s,a) = \frac{1}{1 + e^{(Q_t(s,b) - Q_t(s,a)) * \beta)}} \quad (9)$$

where β is the inverse temperature parameter. High temperatures ($\beta \to 0$) cause the action to be all (nearly) equiprobable. Low temperatures ($\beta \to +\infty$) cause a greater difference in selection probability for actions that differ in their value estimates (46).

We compared three alternative computational models: the ABSOLUTE model, which encodes outcomes in an absolute scale independently of the choice context in which they are presented, the RANGE model which tracks the value of the maximum reward in each context and normalizes the actual reward accordingly, rescaling rewards between 0 and 1, and the HABIT model, which integrates action weights into the decision process.

ABSOLUTE model

The outcomes are encoded as the participants see them. A positive outcome is encoded as its actual positive value (in points): $R_{\mathrm{ABS,t}} \in \{10, 1, 0\}$.

RANGE model

The outcomes (both chosen and unchosen) are encoded on a context-dependent relative scale. On each trial the relative reward $R_{\mathrm{RAN,t}}$ is calculated as follows:

$$R_{\mathrm{RAN,t}} = \frac{R_{\mathrm{ABS,t}} - R_{\mathrm{MIN,t}}(s)}{R_{\mathrm{MAX,t}}(s) - R_{\mathrm{MIN,t}}(s) + 1} \quad (2)$$

As R_{MIN} is initialized at zero and never changes, in our task this model can be reduced as:

$$R_{\mathrm{RAN,t}} = \frac{R_{\mathrm{ABS,t}}}{R_{\mathrm{MAX,t}}(s) + 1} \quad (10)$$

where *s* is the decision context (i.e., a combination of options) and R_{MAX} is a context-dependent variable, initialized to 0 and updated at each trial *t* if the outcome is greater than its current value:

$$R_{\mathrm{MAX},t+1}(s) = R_{\mathrm{MAX},t}(s) + \alpha_R \left(R_{\mathrm{ABS,t}} - R_{\mathrm{MAX},t}(s)\right) \quad \text{if } R_{\mathrm{ABS,t}} > R_{\mathrm{MAX},t}(s) \quad (11)$$

Accordingly, outcomes are progressively normalized so that eventually $R_{\mathrm{RAN,t}} \in [0,1]$. The chosen and unchosen option values and prediction errors are updated with the same rules as in the ABSOLUTE model. Note that the ABSOLUTE model is nested within the RANGE model ($\alpha_R = 0$).

HABIT model

The outcomes are encoded on an absolute scale, but decisions integrate a habitual component (16,19). To do so, in addition to the Q-values, a habitual (or choice trace) component *H* is tracked and updated (with a dedicated learning rate parameter) that takes into account the selected action (1 for chosen option, 0 for the unchosen option). The choice is performed with a softmax rule based on decision weights *D* that integrate Q-values and decision weights *H*:

$$D_t(s, c) = (1 - \omega) * Q_t(s, c) + \omega * H_t(s, c) \quad (3)$$

where at each trial *t*, state *s* and chose option *c*, *D* is the arbiter, *Q* is the goal directed component (Q-values matrix), *H* is the habitual component. The weight ω is fitted as an additional parameter and governs the relative weights of values and habits (for ω=0 the model reduces to the ABSOLUTE model).

655

656 UTILITY model

657 The outcomes are encoded as an exponentiation of the absolute reward, leading to a curvature of
658 the value function (29):

659 $$R_{\mathrm{UTI,t}} = \left(R_{\mathrm{ABS,t}}\right)^{\nu} \quad (4)$$

660 where the exponent ν is the utility parameter, with $0 < \nu < 1$ (for ν =1 the model reduces to the
661 ABSOLUTE model).

References and Notes

1. Louie K, Glimcher PW. Efficient coding and the neural representation of value. Ann N Y Acad Sci. 2012 Mar;1251:13–32.

2. Vlaev I, Chater N, Stewart N, Brown GDA. Does the brain calculate value? Trends Cogn Sci. 2011 Nov;15(11):546–54.

3. Cox KM, Kable JW. BOLD Subjective Value Signals Exhibit Robust Range Adaptation. J Neurosci. 2014 Dec 3;34(49):16533–43.

4. Nieuwenhuis S, Heslenfeld DJ, Alting von Geusau NJ, Mars RB, Holroyd CB, Yeung N. Activity in human reward-sensitive brain areas is strongly context dependent. NeuroImage. 2005 May 1;25(4):1302–9.

5. Elliott R, Agnew Z, Deakin JFW. Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. Eur J Neurosci. 2008 May;27(9):2213–8.

6. 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 29;9(1):4503.

7. Klein TA, Ullsperger M, Jocham G. Learning relative values in the striatum induces violations of normative decision making. Nat Commun. 2017 Jun 20;8(1):16033.

8. Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in reward and punishment learning. Nat Commun. 2015 Aug 25;6:8096.

9. Freidin E, Kacelnik A. Rational Choice, Context Dependence, and the Value of Information in European Starlings (Sturnus vulgaris). Science. 2011 Nov 18;334(6058):1000–2.

10. Pompilio L, Kacelnik A. Context-dependent utility overrides absolute memory as a determinant of choice. Proc Natl Acad Sci. 2010 Jan 5;107(1):508–12.

11. Rustichini A, Conen KE, Cai X, Padoa-Schioppa C. Optimal coding and neuronal adaptation in economic decisions. Nat Commun. 2017 Oct 31;8(1):1208.

12. Webb R, Glimcher PW, Louie K. The Normalization of Consumer Valuations: Context-Dependent Preferences From Neurobiological Constraints. Manag Sci [Internet]. 2020 May 27 [cited 2020 Jul 27]; Available from: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2019.3536

13. 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 Jun 1;19(3):490–502.

14. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012 Apr;35(7):1024–35.

15. Rabin M. Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion. 2000 Jun 9 [cited 2020 Jul 27]; Available from: https://escholarship.org/uc/item/61d7b4pg

16. Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychol Rev. 2019;126(2):292–311.

17. Landry P, Webb R. Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice [Internet]. Rochester, NY: Social Science Research Network; 2019 Jan [cited 2020 Nov 30]. Report No.: ID 2963863. Available from: https://papers.ssrn.com/abstract=2963863

18. Palminteri S, Wyart V, Koechlin E. The Importance of Falsification in Computational Cognitive Modeling. Trends Cogn Sci. 2017 Jun;21(6):425–33.

19. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018 Dec 1;87:31–45.

20. Louie K, Glimcher PW, Webb R. Adaptive neural coding: from biological to behavioral decision-making. Curr Opin Behav Sci. 2015 Oct 1;5:91–9.

21. Dumbalska T, Li V, Tsetsos K, Summerfield C. A map of decoy influence in human multialternative choice. Proc Natl Acad Sci. 2020 Oct 6;117(40):25169–78.

22. Daviet R, Webb R. A Double Decoy Experiment to Distinguish Theories of Dominance Effects [Internet]. Rochester, NY: Social Science Research Network; 2019 Mar [cited 2020 Nov 30]. Report No.: ID 3374514. Available from: https://papers.ssrn.com/abstract=3374514

23. Gluth S, Kern N, Kortmann M, Vitali CL. Value-based attention but not divisive normalization influences decisions with multiple alternatives. Nat Hum Behav. 2020 Jun;4(6):634–45.

24. Goodwin PB. Habit and Hysteresis in Mode Choice. Urban Stud. 1977;14(1):95–8.

25. Dickinson A, Weiskrantz L. Actions and habits: the development of behavioural autonomy. Philos Trans R Soc Lond B Biol Sci. 1985 Feb 13;308(1135):67–78.

26. Lally P, Jaarsveld CHM van, Potts HWW, Wardle J. How are habits formed: Modelling habit formation in the real world. Eur J Soc Psychol. 2010;40(6):998–1009.

27. Thrailkill EA, Trask S, Vidal P, Alcalá JA, Bouton ME. Stimulus Control of Actions and Habits: A Role for Reinforcer Predictability and Attention in the Development of Habitual Behavior. J Exp Psychol Anim Learn Cogn. 2018 Oct;44(4):370–84.

28. Akaishi R, Umeda K, Nagase A, Sakai K. Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron. 2014 Jan 8;81(1):195–206.

29. Neumann J von, Morgenstern O. Theory of Games and Economic Behavior. Princeton University Press; 1953. 774 p.

30. Bernoulli D. Exposition of a New Theory on the Measurement of Risk. Econometrica. 1954;22(1):23–36.

31. Markowitz H. The Utility of Wealth. J Polit Econ. 1952 Apr 1;60(2):151–8.

32. Kahneman D, Tversky A. Subjective probability: A judgment of representativeness. Cognit Psychol. 1972 Jul 1;3(3):430–54.

33. Loomes G, Sugden R. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. Econ J. 1982;92(368):805–24.

34. Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Massachusetts Institute of Technology Press; 2001. 460 p.

35. Li J, Daw ND. Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci Off J Soc Neurosci. 2011 Apr 6;31(14):5504–11.

36. Palminteri S, Pessiglione M. Chapter 23 - Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. In: Dreher J-C, Tremblay L, editors. Decision

Neuroscience [Internet]. San Diego: Academic Press; 2017 [cited 2020 Nov 30]. p. 291–303. Available from: http://www.sciencedirect.com/science/article/pii/B9780128053089000233

37. Lebreton M, Bacily K, Palminteri S, Engelmann JB. Contextual influence on confidence judgments in human reinforcement learning. PLOS Comput Biol. 2019 avr;15(4):e1006973.

38. 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 Sep 28;36(39):10016–25.

39. Pischedda D, Palminteri S, Coricelli G. The Effect of Counterfactual Information on Outcome Value Coding in Medial Prefrontal and Cingulate Cortex: From an Absolute to a Relative Neural Code. J Neurosci. 2020 Apr 15;40(16):3268–77.

40. Conen KE, Padoa-Schioppa C. Partial Adaptation to the Value Range in the Macaque Orbitofrontal Cortex. J Neurosci. 2019 May 1;39(18):3498–513.

41. Padoa-Schioppa C, Rustichini A. Rational Attention and Adaptive Coding: A Puzzle and a Solution. Am Econ Rev. 2014 May;104(5):507–13.

42. Gigerenzer G. The Bias Bias in Behavioral Economics. Rev Behav Econ. 2018 Dec 30;5(3–4):303–36.

43. Haselton MG, Nettle D, Andrews PW. The Evolution of Cognitive Bias. In: The Handbook of Evolutionary Psychology [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2020 Jul 27]. p. 724–46. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470939376.ch25

44. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004 Dec 10;306(5703):1940–3.

45. Girden ER. ANOVA: Repeated Measures. SAGE; 1992. 88 p.

46. Sutton RS, Barto AG. Reinforcement Learning - An Introduction [Internet]. Mit Press; 1998 [cited 2017 Jun 8]. Available from: http://gen.lib.rus.ec/book/index.php?md5=C5BC41EB7F60C05D37473B4A9AC77A2A

47. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Class Cond II Curr Res Theory. 1972;2:64–99.

48. Hergueux J, Jacquemet N. Social preferences in the online laboratory: a randomized experiment. Exp Econ. 2015 Jun 1;18(2):251–83.

49. Kahneman D. Maps of Bounded Rationality: Psychology for Behavioral Economics. Am Econ Rev. 2003 Dec;93(5):1449–75.

50. Shavit T, Sonsino D, Benzion U. A comparative study of lotteries-evaluation in class and on the Web. J Econ Psychol. 2001 Aug 1;22(4):483–91.

51. Schoeffler M, Stöter F-R, Bayerlein H, Edler B, Herre J. An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results. In: ISMIR. 2013.

52. Reinecke K, Gajos KZ. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing [Internet]. Vancouver, BC, Canada: Association for Computing Machinery; 2015 [cited 2020 Jul 27]. p. 1364–1378. (CSCW '15). Available from: https://doi.org/10.1145/2675133.2675246

Acknowledgments

Funding: S.P. is supported by an ATIP-Avenir grant (R16069JS), the Programme Emergence(s) de la Ville de Paris, the Fondation Fyssen and the Fondation Schlumberger pour l'Education et la Recherche. S.B. is supported by MILDECA (Mission Interministérielle de Lutte contre les Drogues et les Conduites Addictives) and the EHESS (Ecole des Hautes Etudes en Sciences Sociales). The funding agencies did not influence the content of the manuscript.

Figures and Tables

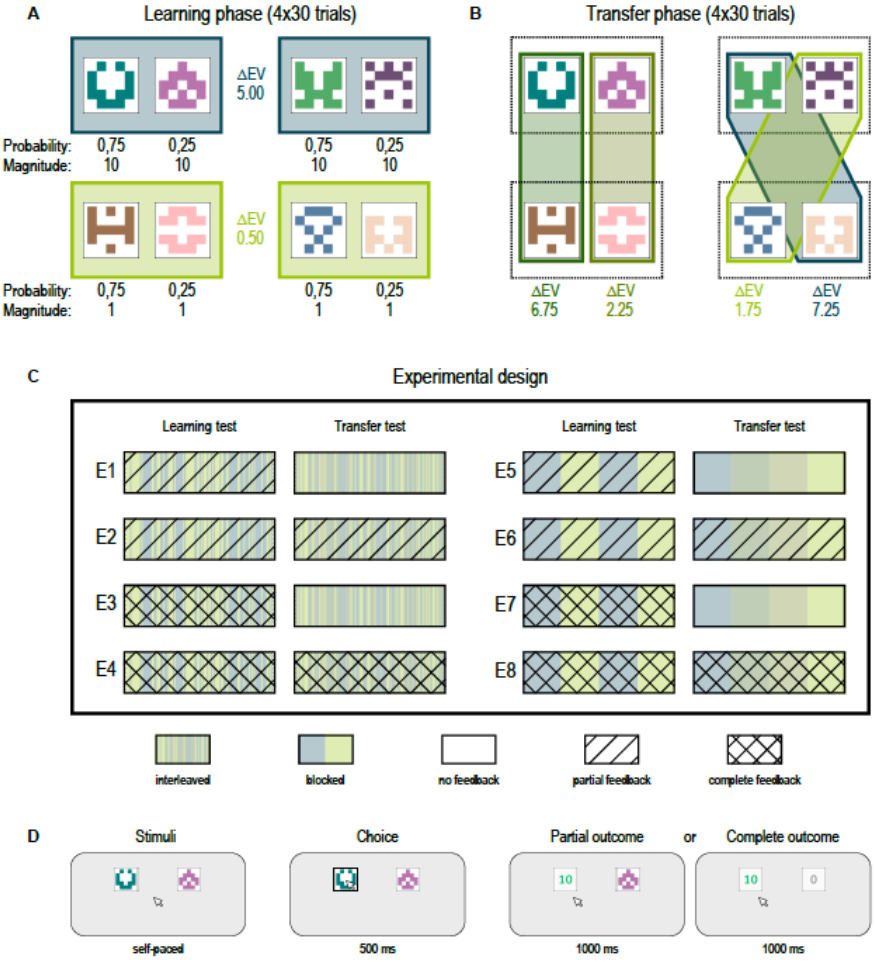


Fig. 1. Experimental design. (A) Choice contexts in the learning phase. During the learning phase, participants were presented with four choice contexts, including high magnitude (ΔEV = 5.0 contexts) and low magnitude (ΔEV = 0.5 contexts). **(B)** Choice contexts in the transfer phase. The four options were re-arranged into four new choice contexts, each involving both the 1pt and the 10pt outcome. **(C)**. Experimental design. The eight experiments varied in the temporal arrangement of choice contexts (interleaved or block) and the quantity of feedback in the learning phase (partial or complete) and the transfer phase (without or with). (**D**) Successive screens of a typical trials (complete feedback; durations are given in milliseconds).

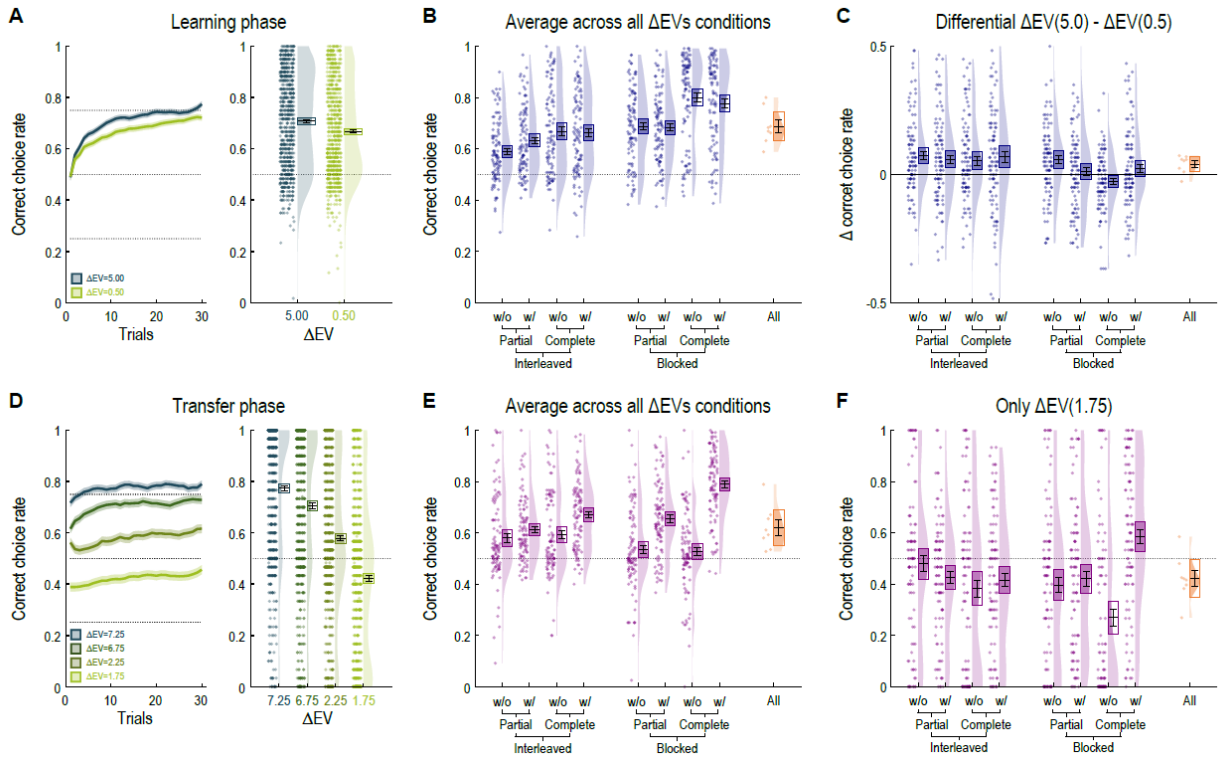


Figure 2. Behavioral results. (**A**) Correct choice rate in the learning phase as a function of the choice context (ΔEV=5.0 or ΔEV=0.5). Leftmost panel: learning curves; rightmost panel: average across all trials. (**B**) Average correct response rate in the learning phase per experiment (in blue: N=800 participants) and meta-analytical (in orange: N=8 experiments). (**C**) Difference in correct choice rate between the ΔEV=5.0 and the ΔEV=0.5 contexts per experiment (in blue: N=800 participants) and meta-analytical (in orange: N=8 experiments). (**D**) Correct choice rate in the transfer phase as a function of the choice context (ΔEV=7.25, ΔEV=6.75, ΔEV=2.25 or ΔEV=1.75). Leftmost panel: learning curves; rightmost panel: average across all trials. **(E)** Average correct response rate in the transfer phase per experiment (in pink: N=800 participants) and meta-analytical (in orange: N=8 experiments). **(F)** Correct choice rate for the ΔEV=1.75 context only (in pink: N=800 participants) and meta-analytical (in orange: N=8 experiments). In all panels: points indicate individual average, areas indicate probability density function, boxes indicate 95% confidence interval and errors bars indicate s.e.m.

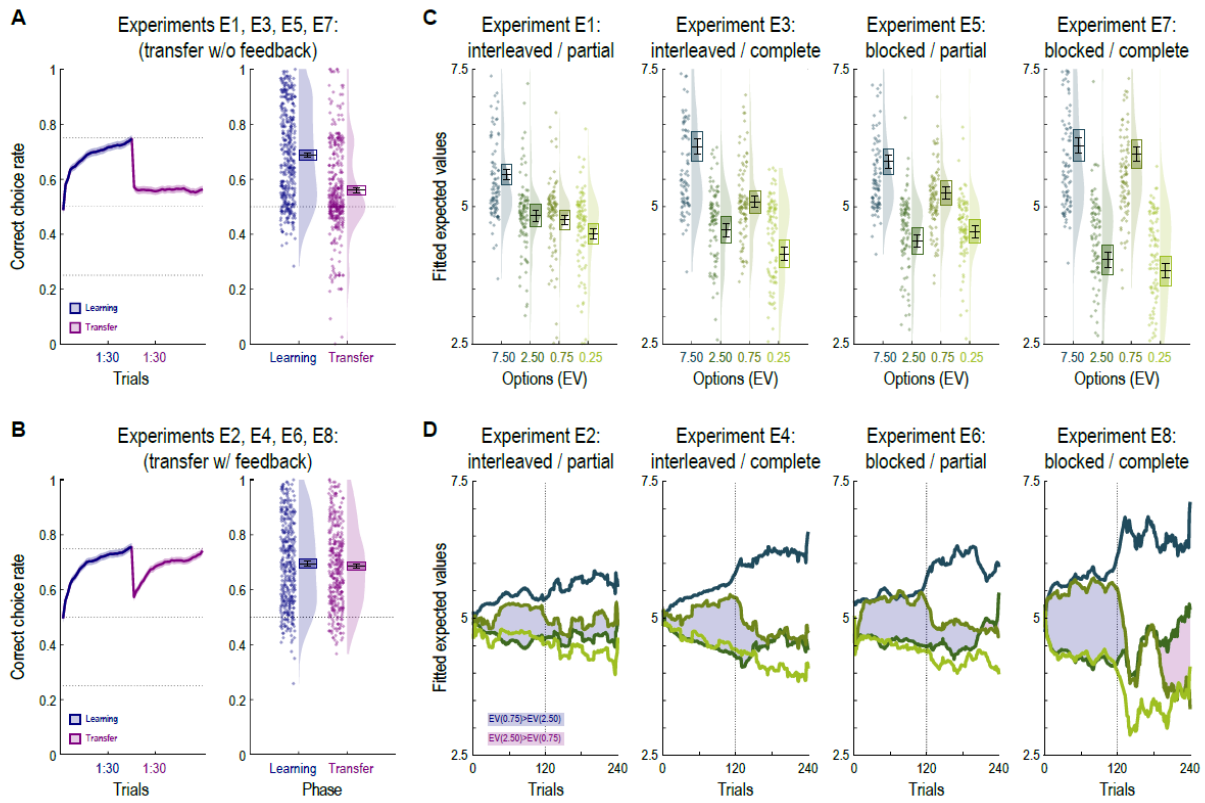


Figure 3. Learning versus transfer comparison and inferred option values. (A-B) Average response rate in the learning (blue) and transfer (pink) phase for experiments without **(A)** and with **(B)** trial-by-trial transfer feedback. Leftmost panel: learning curves; rightmost panel: average across all trials. **(C)** Average inferred option values for the experiments without trial-by-trial transfer feedback. **(D)** Trial-by-trial inferred option values for the experiments with trial-by-trial transfer feedback. In all panels: points indicate individual average, areas indicate probability density function, boxes indicate 95% confidence interval and errors bars indicate s.e.m.

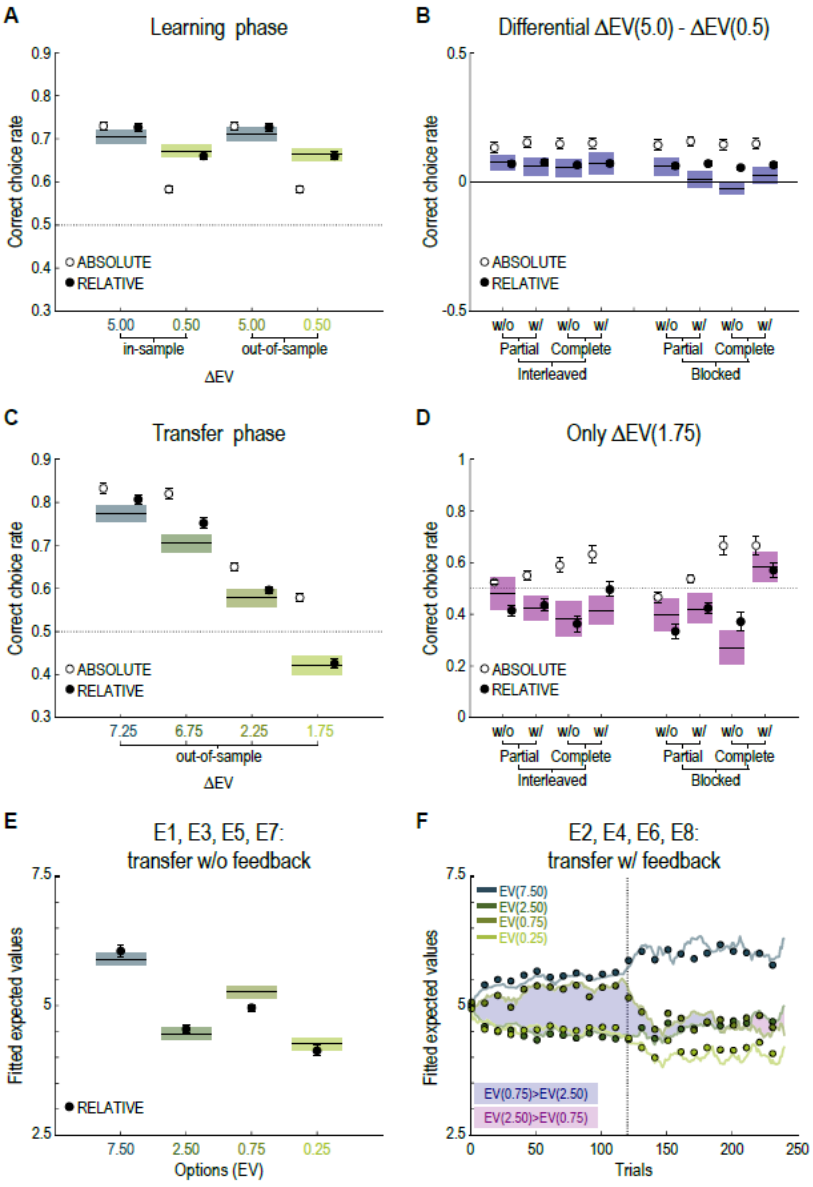


Figure 4. Model comparison. Model simulations of ABSOLUTE (white) and RANGE (black) models over the behavioral data (mean and 95% confidence interval) in each context. **(A)** Simulated data in the learning phase were obtained with the parameters fitted in half the contexts (ΔEV=5.0 and ΔEV=0.5) of the learning phase (in-sample and out-of-sample predictions). **(B)** Data and simulations of the differential between high magnitude (ΔEV=5.0) and low magnitude (ΔEV=0.5) contexts. **(C)** Simulated data in the transfer phase were obtained with the parameters fitted in all the contexts of the learning phase (out-of-sample predictions). **(D)** Data and simulations in the context ΔEV=1.75 only. **(E)** Average inferred option values for the behavioral data and simulated data (black dots: RANGE model only) for the experiments without trial-by-trial transfer feedback. **(F)** Trial-by-trial inferred option values for the behavioral data and simulated data (colored dots: RANGE model only) for the experiments with trial-by-trial transfer feedback, where curves indicate trial-by-trial fit of each inferred option value, and colored dots indicate RANGE model simulations.

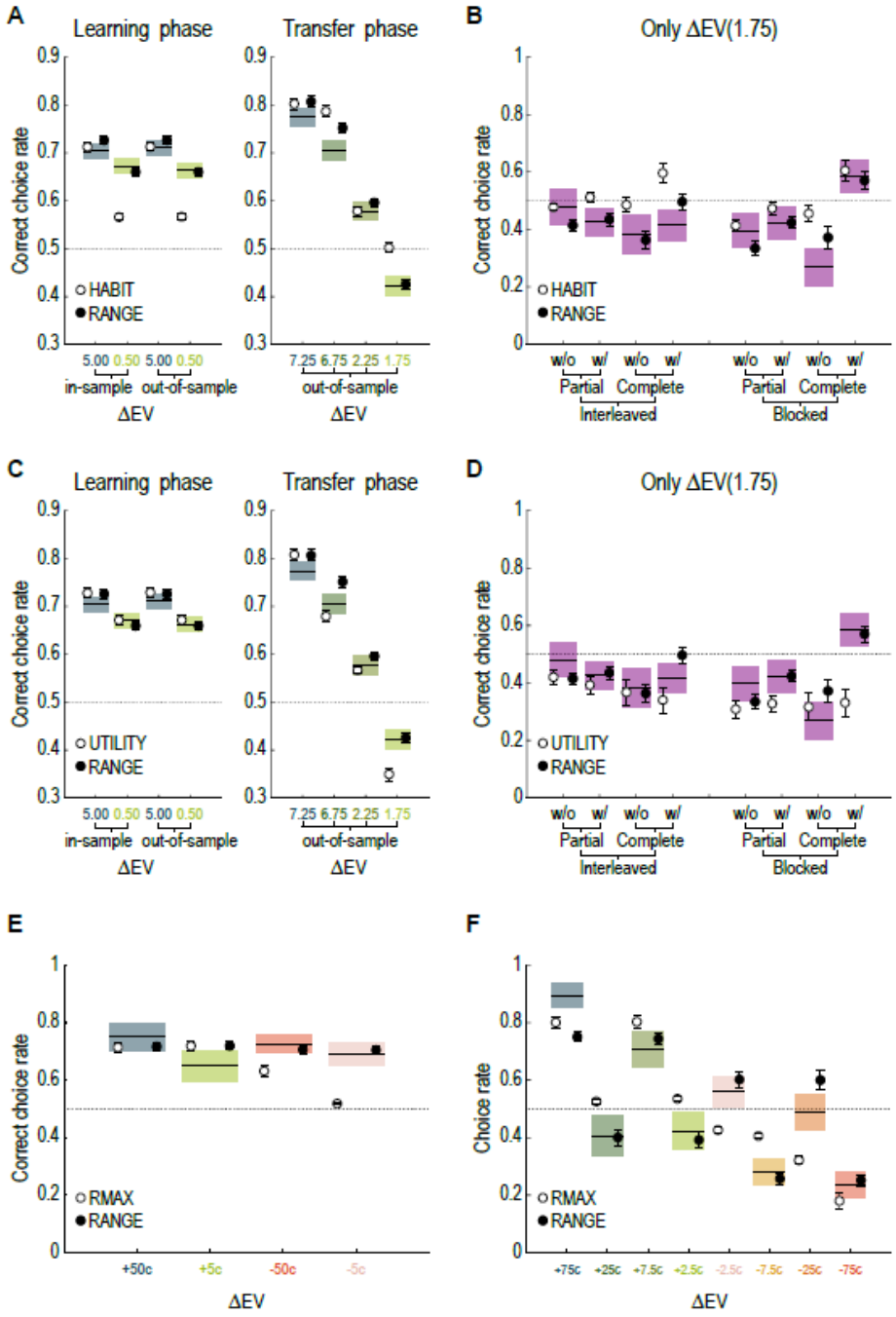


Figure 5. Ruling out alternative models. Model simulations of HABIT, resp. UTILITY (white) and RANGE (black) models over the behavioral data (mean and 95% confidence interval) in each context. **(A, C)** Simulated data in the learning phase were obtained with the parameters fitted in half the contexts (ΔEV=5.0 and ΔEV=0.5) of the learning phase (in-sample and out-of-sample predictions). Simulated data in the transfer phase were obtained with the parameters fitted in all the contexts of the learning phase (out-of-sample predictions). **(B, D)** Data and simulations in the context ΔEV=1.75 only. **(E,F)** Behavioral data from Bavard, Lebreton et al (2018). Comparing the full RANGE model to its simplified version RMAX in the learning phase (correct choice rate per choice context) and in the transfer test (choice rate per symbol). This study included gain -related

contexts (with +50c and +5c as average outcomes) and loss-related contexts (with -50c and -5c as average outcomes) in the learning phase. Choice rates in the transfer phare are presented as a function of decreasing option expected values.

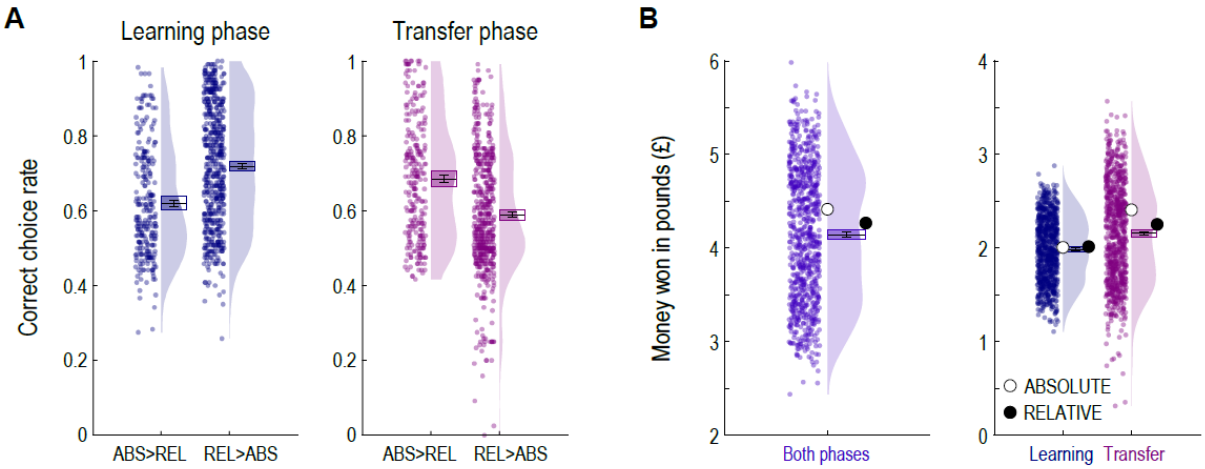


Figure 6. The financial cost of relative value learning. (A) Correct choice rate in the learning phase (blue) and the transfer phase (pink) as a function of the difference in log-likelihood between the ABSOLUTE and the RANGE models. ABS>RAN: positive difference, N=255. RAN>ABS: negative difference, N=545. **(B)** Actual and simulated money won in pounds over the whole task (purple), the learning phase only (blue) and the transfer phase only (pink). Points indicate individual participants, areas indicate probability density function, boxes indicate confidence interval and errors bars indicate s.e.m. Dots indicate model simulations of ABSOLUTE (white) and RANGE (black) models.

			Learning performance				Transfer performance				Overall performance			
	DFn	DFd	F-val		Diff	η_p^2	F-val		Diff	η_p^2	F-val		Diff	η_p^2
LF-Learningfeedback Complete>Partial	1	792	55,57	***	0,079	0,18	22,36	***	0,050	0,01	61,68	***	0,064	0,11
TF-Transferfeedback With>Without	1	792	0,04		0,002	0,00	137,18	***	0,12	0,07	58,11	***	0,063	0,10
BE-Blockeffe	1	792	87,22	***	0,099	0,25	1,53		0,013	0,00	46,82	***	0,056	0,08

ct Block > Interleaved													
PE - Phase effect Learning > Transfer	1	792	-		-	-	-		-	-	103,07	***	0,067	0,12
LF x TF	1	792	2,61			0,01	20,18	***		0,01	3,33			0,01
LF x BE	1	792	5,05	*		0,02	1,66			0,00	5,20	*		0,01
TF x BE	1	792	2,43			0,01	42,22	***		0,02	9,89	**		0,02
LF x PE	1	792	-		-	-	-		-	-	4,97	*		0,01
TF x PE	1	792	-		-	-	-		-	-	82,30	***		0,09
BE x PE	1	792	-		-	-	-		-	-	42,09	***		0,05
LF x TF	1	792	0,55			0,00	5,02	*		0,00	3,65			0,01

x BE											
LF x TF x PE	1	792	-	-	-	-	-	-	23,37	***	0,03
LF x BE x PE	1	792	-	-	-	-	-	-	0,61		0,00
TF x BE x PE	1	792	-	-	-	-	-	-	40,58	***	0,05
LF x TF x BE x TE	1	792	-	-	-	-	-	-	1,39		0,00

Table 1. Statistical effects of the ANOVA on the choice rate as a function of task factors. LF: learning feedback (complete/partial). TF: transfer feedback (with/without). BE: block effect (block/interleaved). PE: phase effect (learning/transfer). DFn: degrees of freedom numerator. DFd: degrees of freedom denominator. η_p^2: portion of variance explained. *p<.05, **p<.01, ***p<.001

	Experiment 1 N=100		Experiment 2 N=100		Experiment 3 N=100		Experiment 4 N=100		Experiment 5 N=100		Experiment 6 N=100		Experiment 7 N=100		Experiment 8 N=100		Total N=800	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Age	30,48	10,70	27,23	8,30	32,01	10,51	31,57	9,80	33,04	10,48	28,46	10,20	28,73	9,89	28,84	9,60	30,06	10,10
% correct																		
Learning phase	0,59	0,12	0,63	0,13	0,67	0,17	0,66	0,16	0,69	0,15	0,68	0,14	0,80	0,17	0,78	0,16	0,69	0,16
ΔEV=5.0	0,63	0,16	0,66	0,17	0,70	0,19	0,70	0,20	0,72	0,17	0,69	0,17	0,79	0,19	0,79	0,18	0,67	0,18
ΔEV=0.5	0,55	0,13	0,60	0,14	0,64	0,19	0,63	0,19	0,66	0,17	0,68	0,14	0,81	0,17	0,76	0,18	0,71	0,18
Transfer phase	0,58	0,17	0,61	0,12	0,59	0,16	0,67	0,13	0,54	0,16	0,66	0,14	0,53	0,16	0,79	0,14	0,62	0,17
ΔEV=7.25	0,67	0,28	0,76	0,22	0,75	0,29	0,85	0,19	0,66	0,30	0,84	0,18	0,76	0,31	0,93	0,14	0,77	0,26
ΔEV=6.75	0,64	0,29	0,68	0,26	0,70	0,31	0,81	0,19	0,62	0,32	0,76	0,27	0,55	0,37	0,89	0,16	0,71	0,30
ΔEV=2.25	0,54	0,27	0,58	0,19	0,54	0,34	0,61	0,28	0,47	0,32	0,60	0,18	0,54	0,36	0,76	0,22	0,58	0,29
ΔEV=1.75	0,48	0,30	0,43	0,23	0,38	0,33	0,42	0,27	0,40	0,31	0,42	0,28	0,27	0,32	0,59	0,29	0,42	0,30

Table 2. Participants' age and correct choice rate as a function of experiments and task factors.

Model	Learning phase	Transfer phase
ABSOLUTE	-42.74 ± 1.27***	-161.19 ± 11.41***
RANGE	-37.72 ± 0.96	-96.79 ± 4.79
HABIT	-36.68 ± 0.91	-104.62 ± 6.01$^{\$}$
UTILITY	-36.31 ± 0.53***	-104.94 ± 5.24**

Table 3. Quantitative model comparing. Values **reported here represent out-of-sample likelihood after two-fold cross validation. Comparison to the RANGE model: ***p<0.001; **p<0.01; $^{\$}$p<0.08.**

Supplementary Materials

Laboratory-based replications and robustness to outliers

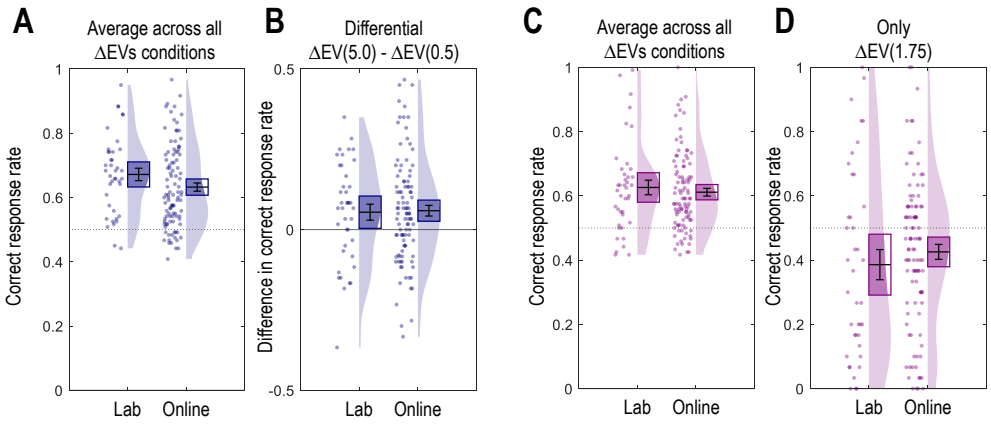
To ascertain that online testing would not significantly affect the main conclusions, a laboratory-based experiment was originally performed. We recruited 40 participants (28 females, aged 24.28±3.05) via Internet advertising in mailing list dedicated to cognitive science-related activities. The experimental design was that of experiment E2 of the main text (partial feedback information in both the learning phase and the transfer phase, and trials in an interleaved order; see **Figure 1**).

In order to characterize learning behavior of participants, we analyzed the correct response rate in both phases, i.e., choices directed toward the most favorable option at each trial. To assess successful learning, we first tested participants' correct response rate against chance level. We found it to be above chance level in both the learning phase ($t(39) = 8.88$, $p < .0001$, $d = 1.40$, **Supp. Figure 1A**) and the transfer phase ($t(39) = 5.55$, $p < .0001$, $d = 0.88$, **Supp. Figure 1C**). We found a significant effect of magnitude in the learning phase ($t(39) = 2.18$, $p = .036$, $d = 0.34$, **Supp. Figure 1B**), and the correct choice rate in the ΔEV=1.75 context was significantly below chance level ($t(39) = -2.43$, $p = .020$, $d = -0.38$, **Supp. Figure 1D**). Of note, the effect sizes were really comparable with the ones observed in the corresponding online experiment (learning performance $d = 1.04$ vs 1.40, transfer performance $d = 0.93$ vs 0.88, magnitude effect $d = 0.35$ vs 0.34, value inversion $d = -0.32$ vs -0.38).

In addition to checking that the same significant results were present, to formally assess the similarity between online- and laboratory-based experiments, we explicitly compared their scores. Correct choice rate in the learning phase did not significantly differ between laboratory and online datasets ($t(138) = 1.67$, $p = .10$, $d = 0.31$, **Supp. Figure 1A**), neither did the magnitude effect ($t(138) = -0.15$, $p = .88$, $d = -0.03$, **Supp. Figure 1B**). Concerning the transfer phase, overall correct choice rate was not significantly different ($t(138) = 0.62$, $p = .54$, $d = 0.12$, **Supp. Figure 1C**) and the same result was obtained looking specifically at the ΔEV=1.75 context ($t(138) = -0.84$, $p = .40$, $d = -0.16$, **Supp. Figure 1D**). Of note, although the control over the measure of reaction times is arguably limited in online experiments, also this measure did not differ between laboratory- and online-based experiments ($t(138) = -0.50$, $p = .62$, $d = -0.09$), This similarity between laboratory- and online-based results supports the usefulness of online-based experiments as a way to target larger, more diversified populations with reduced administrative and financial costs (48). The limitations that can be encountered with online-based experiments - such as lower data quality, faster reaction time, lack of engagement from the participants (49,50) – were not significantly present in our data.

However, to further check the robustness of our results, we run analyses of the online data excluding subjects presenting unusual task completion time. We approximated participants' total reaction time over the whole task by a normal distribution and removed outliers at a significance level of $p<0.05$. This led to a removal of 30 participants for the eight online experiments leading to a final sample of 770 participants. We found that the totality of the statistically significant results described in the **Results** section were observable without these reaction time outliers, thus we decided to include all participants in the statistics reported in the **Results** section. In conclusion, our results successfully replicate in the laboratory. Moreover, our results confirm the findings of

recent studies comparing both experimental methods and showing that they produce comparable data quality (51,52).



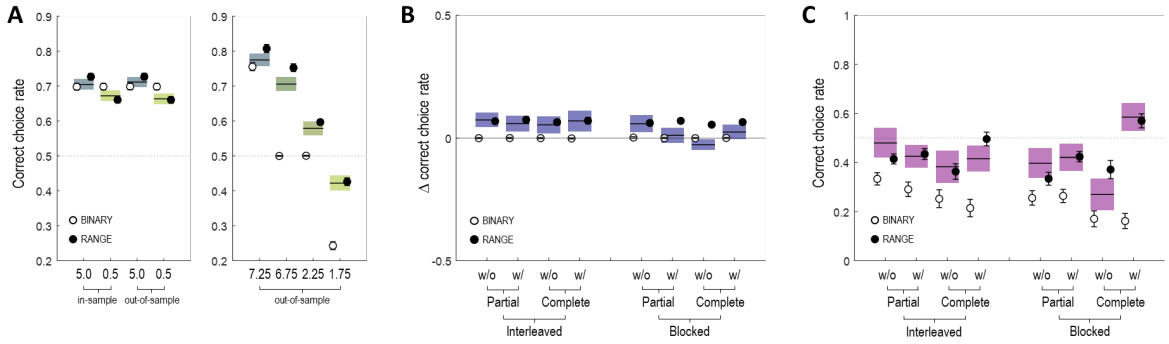
Supp. Figure 1. Comparing laboratory and online experiments. (**A**) Average correct response rate in the learning phase per experiment. (**B**) Difference in correct choice rate between the ΔEV=5.0 and the ΔEV=0.5 contexts. **(C)** Average correct response rate in the transfer phase. **(D)** Correct choice rate for the ΔEV=1.75 context only.

Additional model comparisons

The computational results presented here follow the same fitting and simulation methods presented in the main text for the main computational models.

BINARY model

We analyzed the generative performances of a "full-adaptation" model encoding non-zero outcomes as ones, regardless of their actual magnitude (10pt, 1pt), that we refer to as the BINARY model. At least three behavioral features allow us to reject the BINARY model. Of note, the modek is a special case of the UTILITY model for extremely diminishing marginal utility ($\nu = 0$; $R_{\mathrm{UTI,t}} = \left(R_{\mathrm{ABS,t}}\right)^{\nu}$). First, it is not able to capture participants' behavior in the learning phase by failing to accurately predict the outcome magnitude difference (**Supp. Figure 2A** and **Supp. Figure 2B**); second, the model predicts perfect indifference in the ΔEV=6.75 and the ΔEV=2.25 contexts in the transfer phase, while behavioral results show, respectively, a strong and moderate preference for the high EV options in these contexts; third, the BINARY model predicts an exaggerated rate of suboptimal preferences in the ΔEV=1.75 context in the transfer phase (**Supp. Figure 2A** and **Supp. Figure 2C**). This is true in all 8 experiments and even more striking in E8 where the participants were able to correct their bias.



Supp. Figure 2: generative performance of the RANGE model (black dots) compared to a full-adaptation model encoding rewards as 1's or 0's (white dots: BINARY model). Black lines represent the empirical averages. Colored squares indicate the s.e.m. around the empirical averages.

REFERENCE model

We also analyzed the generative performances of a previous context-dependence model (8) that we call here REFERENCE because of its distinctive feature is to apply reference point dependence to outcome encoding:

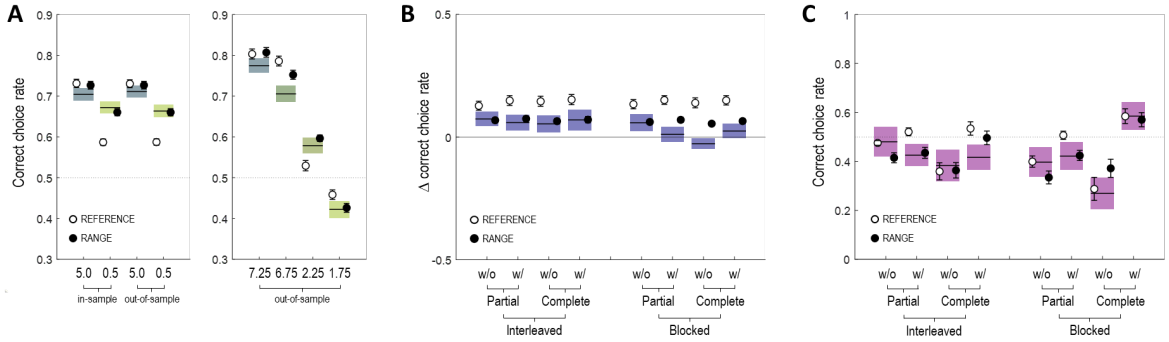
$$Q(s,a) \leftarrow Q(s,a) + \alpha_Q * (R_{\mathrm{ABS}} - V(s) - Q(s,a))$$

Where s is the state (or context, pair of options), $V(s)$ is the state value (or reference point), $Q(s,a)$ is the Q-value (estimated expected value). $V(s)$ is also learnt iteratively, as follows:

$$V(s) \leftarrow V(s) + \alpha_V * \left(\frac{R_{\text{ABS}} + Q(s,u)}{2} - V(s) \right)$$

When the feedback is complete, $Q(s,u)$ (the Q-value of the unchosen option) is replaced by the outcome of the unchosen option. α_V is an additional free parameter for the state value $V(s)$.

Concerning the learning phase, model simulation analysis (**Supp. Figure 3**) showed that, while the REFERENCE model matches the performance in the high-magnitude contexts in the learning phase (ΔEV=5), it fails to capture the performance in low-magnitude contexts (ΔEV=0.5). This is expected as the model does not implement range adaptation in any form. Concerning the transfer phase, the REFERENCE model reproduces a pattern that is qualitatively close to the observed results, but still less accurate compared to the RANGE model (out of sample likelihood comparison LL_{RAN} = -96.79 vs LL_{REF} = -186.68, $t(799) = 8.26$, $p < .0001$). To sum up, the REFERENCE model is strongly rejected by the learning phase results (where it essentially behaves like to the ABSOLUTE model) and weakly rejected by the transfer phase results, where it manages to capture the overall pattern, but in a less accurate manner.



Supp. Figure 3: generative performance of the RANGE model (black dots) compared to the REFERENCE model (white dots). Black lines represent the empirical averages. Colored squares indicate the s.e.m. around the empirical averages.

GLOBAL model

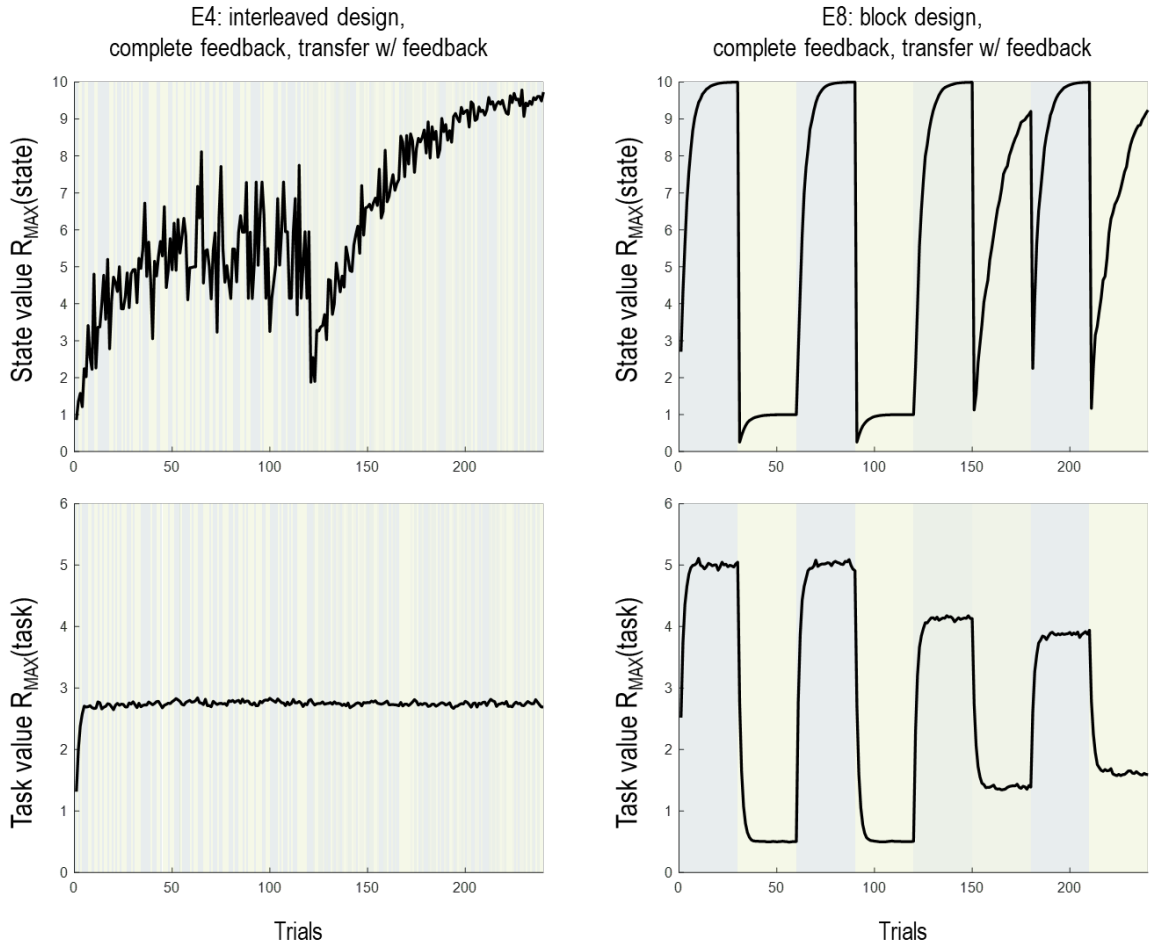
The RANGE model as we presented in the main text, does not contain any element to account for the block/interleaved effect. Here we propose a possible computational interpretation to account for the effects of this manipulation (more precisely the fact that contextual effects are exacerbated in block experiments). The key idea of this model is that the notion of 'context' can be break down into two components. The 'local' context is what we referred to as simply the context in the paper (essentially a pair of cues, or a state in the reinforcement learning framework). In addition to the local context we also postulate a 'global' context that integrate over a time scale larger than a trial (it could be understood as the current 'value' of the task). To instantiate these ideas, we built an alternative model (GLOBAL) that includes both "global" (or task-level) and local (or pair of options-level) contextual variables: $R_{\text{MAX}}(\text{task})$ and $R_{\text{MAX}}(\text{state})$. The $R_{\text{MAX}}(\text{state})$ is learnt similarly to the state-value in the REFERENCE model, except that it is not bounded to any particular pair of options:

$$R_{\mathrm{MAX}}(\mathrm{task}) \leftarrow R_{\mathrm{MAX}}(\mathrm{task}) + \alpha_T * \left(\frac{R_{\mathrm{ABS}} + Q(s,u)}{2} - R_{\mathrm{MAX}}(\mathrm{task})\right)$$

When the feedback is complete, $Q(s,u)$ (the Q-value of the unchosen option) is replaced by the outcome of the unchosen option. α_T is an additional free parameter for the $R_{\mathrm{MAX}}(\mathrm{task})$. The range normalization rule (that we write here in its simplified manner that takes into account that $R_{\mathrm{MIN}} = 0$ everywhere in our task) in the option value update rule of the GLOBAL model is as follows:

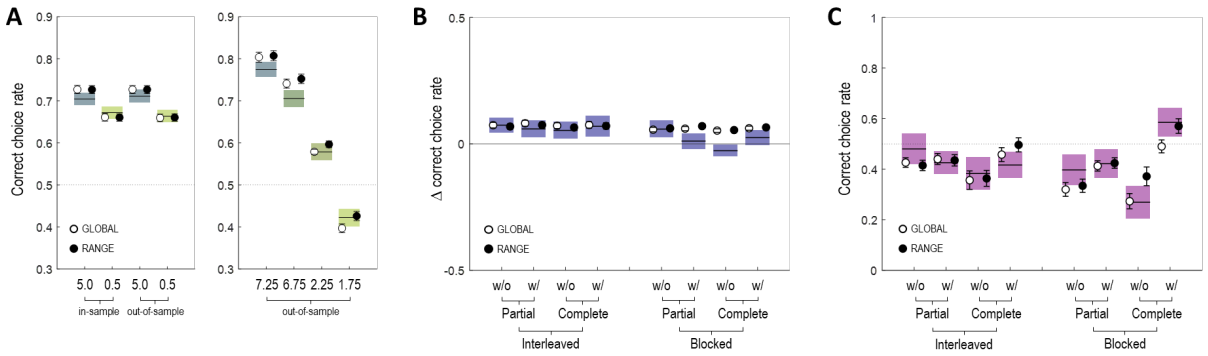
$$Q(s,a) \leftarrow Q(s,a) + \alpha_Q * \left(\frac{R_{\mathrm{ABS}}}{R_{\mathrm{MAX}}(\mathrm{state}) + R_{\mathrm{MAX}}(\mathrm{task}) + 1} - Q(s,a)\right)$$

This simple model accounts for increased contextual effects in block design, because in the block design, $R_{\mathrm{MAX}}(\mathrm{task})$ and the $R_{\mathrm{MAX}}(\mathrm{state})$ remain coherent for longer time periods (**Supp. Figure 4**), thus allowing the summation of their effects. As shown in **Supp. Figure 5**, the model seems qualitatively equal than the RANGE model, if not better at matching performance in most of the 8 different versions of the ΔEV=1.75 context.



Supp. Figure 4: The figure illustrates the evolution across the experiment of the hidden variables $R_{\mathrm{MAX}}(\mathrm{state})$ and $R_{\mathrm{MAX}}(\mathrm{task})$. Simulations concern E4 (interleaved design, complete feedback,

transfer with feedback) and E8 (block design, complete feedback, transfer with feedback). Background colors show the choice context (color code as in Figure 1).



Supp. Figure 5: generative performance of the RANGE model (black dots) compared to the GLOBAL model (white dots). Black lines represent the empirical averages. Colored squares indicate the s.e.m. around the empirical averages.

REGRET model

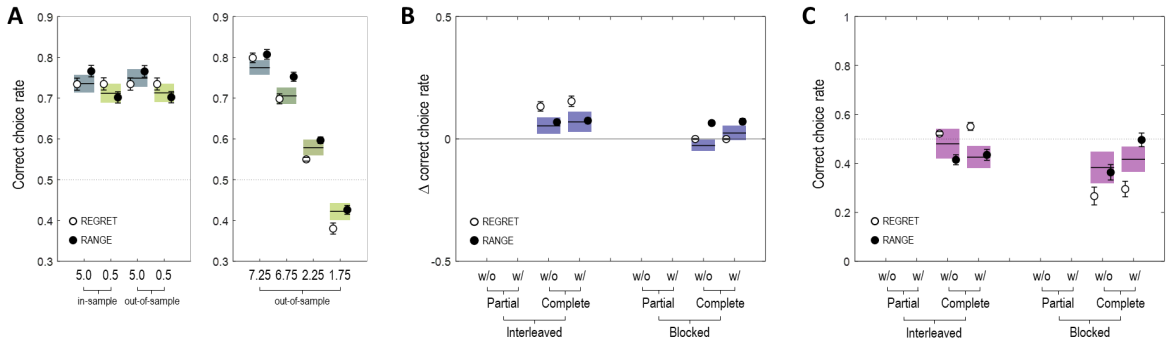
Finally, we analyzed a model assuming that option values are purely encoded by outcome comparison (akin to a relief/regret signal). A similar idea has been put forward by other studies (7,35) where it proved successful in explain striatal neural activity and, to some extent, behavioral data. This model has the strong handicap that it cannot be straight-forwardly extended to the partial feedback case, where the outcome of the unchosen option is not showed. We therefore tested the proposed model in the 4 experiments featuring complete feedback.

Option values in the REGRET model are updated as follows, with R_C and R_U the outcomes of the chosen option and unchosen option, respectively:

$$R_{\mathrm{REG},t} = \begin{cases} 1 & \text{if } R_C > R_U \\ 0 & \text{if } R_C = R_U \\ -1 & \text{if } R_C < R_U \end{cases}$$

$$Q_{t+1}(s,c) = Q_t(s,c) + \alpha_c * (R_{\mathrm{REG},t} - Q_t(s,c))$$
$$Q_{t+1}(s,u) = Q_t(s,u) + \alpha_u * (R_{\mathrm{REG},t} - Q_t(s,u))$$

As clearly illustrated by the model simulations (**Supp. Figure 6**), the REGRET model does not fit well the behavioral data, especially in the transfer phase, where it overestimates value inversion in the ΔEV=1.75 context. In other terms through a different mechanism, the REGRET model suffers from the same problem the BINARY model: they predict to much option value context dependence.



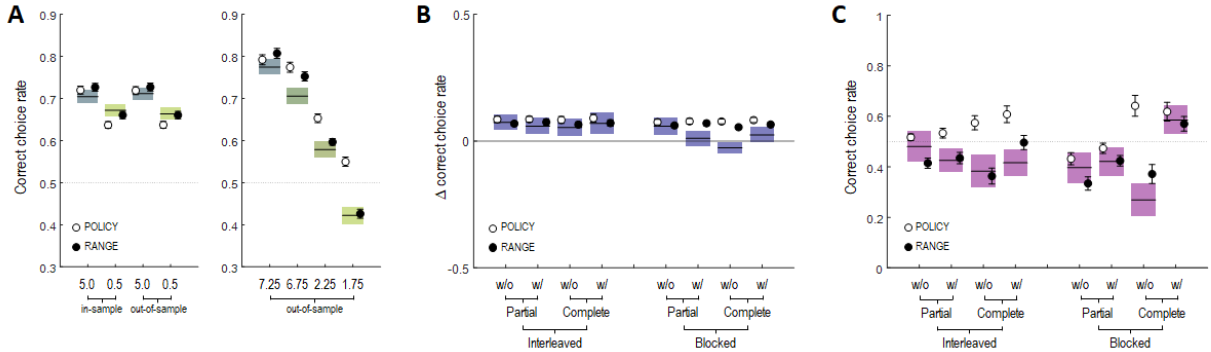
Supp. Figure 7: generative performance of the RANGE model (black dots) compared to the REGRET model (white dots). Black lines represent the empirical averages. Colored squares indicate the s.e.m. around the empirical averages.

POLICY model

Finally, we considered a model that applies range normalization at the decision step, i.e., in the softmax, instead of the outcome encoding stage as in the RANGE model. In this model (POLICY) the probability of choosing option a over option b is defined by:

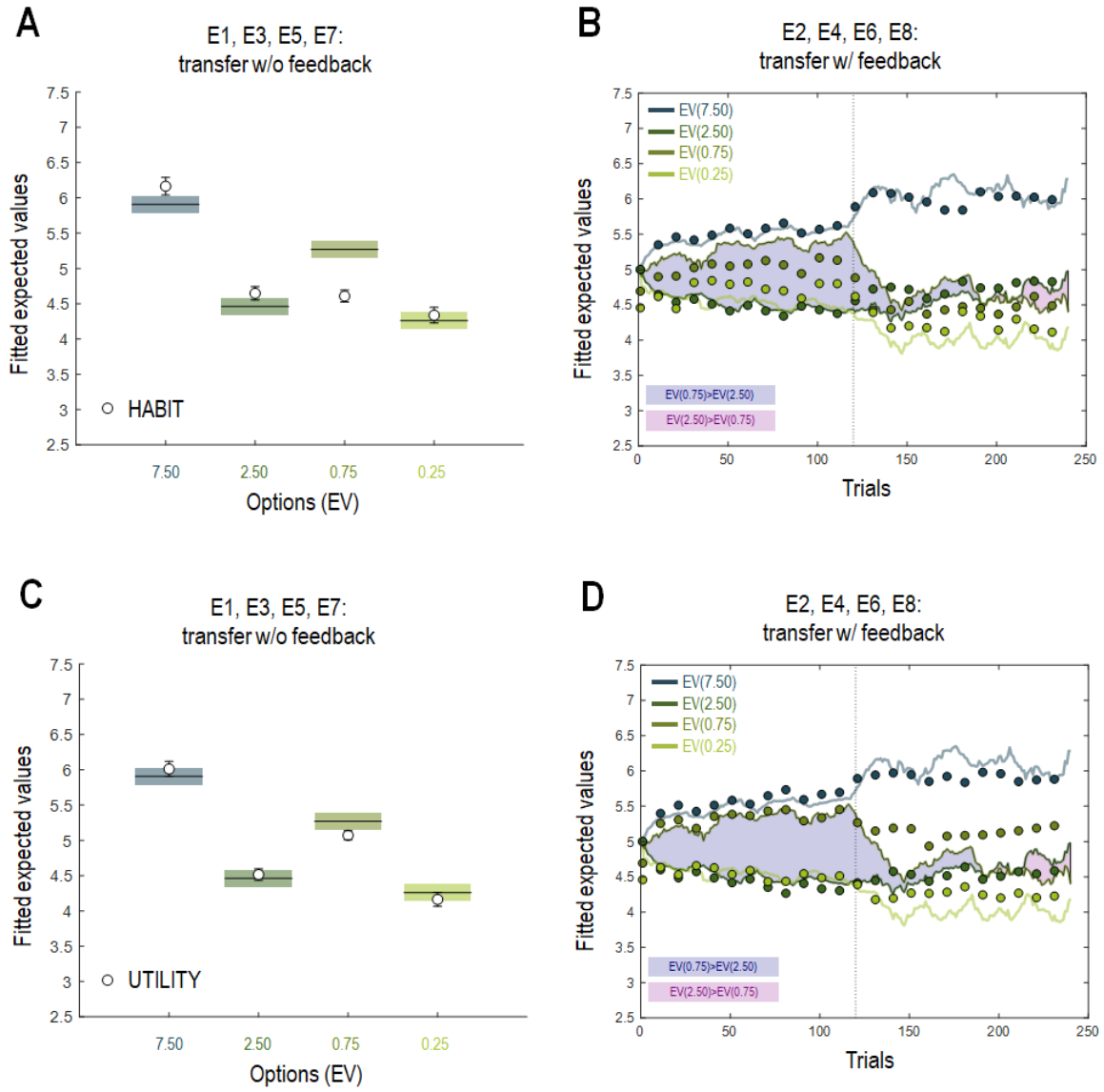
$$P_t(s,a) = \frac{1}{1 + e^{\left(\beta * \frac{Q_t(s,b) - Q_t(s,a)}{1 + \max\{Q_t(s,:)\} - \min\{Q_t(s,:)\}}\right)}}$$

Similarly to the RANGE model, the POLICY model is able to capture the magnitude difference in the learning phase. In the transfer phase however, the POLICY model fails to predict the value inversion in the ΔEV=1.75 context. This is due to the fact that, despite the normalization process within the softmax function, option values remain encoded in an absolute scale. Whereas in the learning phase the POLICY model predicts a behavior compatible with the RANGE model, in the transfer phase it predicts a behavior consistent with the ABSOLUTE model (**Supp. Fig. 6**).



Supp. Figure 6: generative performance of the RANGE model (black dots) compared to the POLICY model (white dots). Black lines represent the empirical averages. Colored squares indicate the s.e.m. around the empirical averages.

1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099



Supp. Figure 8. Ruling out habitual learning and marginally decreasing utility. (A-C) Average inferred option values for the behavioral data and simulated data for the experiments without trial-by-trial transfer feedback (white dots: HABIT (resp. UTILITY) model). **(B-D)** Trial-by-trial inferred option values for the behavioral data and simulated data for the experiments with trial-by-trial transfer feedback, where curves indicate trial-by-trial fit of each inferred option value, and colored dots indicate HABIT (resp. UTILITY) model simulations.

Supplementary references

1. Louie K, Glimcher PW. Efficient coding and the neural representation of value. Ann N Y Acad Sci. 2012 Mar;1251:13–32.

2. Vlaev I, Chater N, Stewart N, Brown GDA. Does the brain calculate value? Trends Cogn Sci. 2011 Nov;15(11):546–54.

3. Cox KM, Kable JW. BOLD Subjective Value Signals Exhibit Robust Range Adaptation. J Neurosci. 2014 Dec 3;34(49):16533–43.

4. Nieuwenhuis S, Heslenfeld DJ, Alting von Geusau NJ, Mars RB, Holroyd CB, Yeung N. Activity in human reward-sensitive brain areas is strongly context dependent. NeuroImage. 2005 May 1;25(4):1302–9.

5. Elliott R, Agnew Z, Deakin JFW. Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. Eur J Neurosci. 2008 May;27(9):2213–8.

6. 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 29;9(1):4503.

7. Klein TA, Ullsperger M, Jocham G. Learning relative values in the striatum induces violations of normative decision making. Nat Commun. 2017 Jun 20;8(1):16033.

8. Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in reward and punishment learning. Nat Commun. 2015 Aug 25;6:8096.

9. Freidin E, Kacelnik A. Rational Choice, Context Dependence, and the Value of Information in European Starlings (Sturnus vulgaris). Science. 2011 Nov 18;334(6058):1000–2.

10. Pompilio L, Kacelnik A. Context-dependent utility overrides absolute memory as a determinant of choice. Proc Natl Acad Sci. 2010 Jan 5;107(1):508–12.

11. Rustichini A, Conen KE, Cai X, Padoa-Schioppa C. Optimal coding and neuronal adaptation in economic decisions. Nat Commun. 2017 Oct 31;8(1):1208.

12. Webb R, Glimcher PW, Louie K. The Normalization of Consumer Valuations: Context-Dependent Preferences From Neurobiological Constraints. Manag Sci [Internet]. 2020 May 27 [cited 2020 Jul 27]; Available from: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2019.3536

13. 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 Jun 1;19(3):490–502.

14. Collins AGE, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012 Apr;35(7):1024–35.

15. Rabin M. Diminishing Marginal Utility of Wealth Cannot Explain Risk Aversion. 2000 Jun 9 [cited 2020 Jul 27]; Available from: https://escholarship.org/uc/item/61d7b4pg

16. Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychol Rev. 2019;126(2):292–311.

17. Landry P, Webb R. Pairwise Normalization: A Neuroeconomic Theory of Multi-Attribute Choice [Internet]. Rochester, NY: Social Science Research Network; 2019 Jan [cited 2020 Nov 30]. Report No.: ID 2963863. Available from: https://papers.ssrn.com/abstract=2963863

18. Palminteri S, Wyart V, Koechlin E. The Importance of Falsification in Computational Cognitive Modeling. Trends Cogn Sci. 2017 Jun;21(6):425–33.

19. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. J Math Psychol. 2018 Dec 1;87:31–45.

20. Louie K, Glimcher PW, Webb R. Adaptive neural coding: from biological to behavioral decision-making. Curr Opin Behav Sci. 2015 Oct 1;5:91–9.

21. Dumbalska T, Li V, Tsetsos K, Summerfield C. A map of decoy influence in human multialternative choice. Proc Natl Acad Sci. 2020 Oct 6;117(40):25169–78.

22. Daviet R, Webb R. A Double Decoy Experiment to Distinguish Theories of Dominance Effects [Internet]. Rochester, NY: Social Science Research Network; 2019 Mar [cited 2020 Nov 30]. Report No.: ID 3374514. Available from: https://papers.ssrn.com/abstract=3374514

23. Gluth S, Kern N, Kortmann M, Vitali CL. Value-based attention but not divisive normalization influences decisions with multiple alternatives. Nat Hum Behav. 2020 Jun;4(6):634–45.

24. Goodwin PB. Habit and Hysteresis in Mode Choice. Urban Stud. 1977;14(1):95–8.

25. Dickinson A, Weiskrantz L. Actions and habits: the development of behavioural autonomy. Philos Trans R Soc Lond B Biol Sci. 1985 Feb 13;308(1135):67–78.

26. Lally P, Jaarsveld CHM van, Potts HWW, Wardle J. How are habits formed: Modelling habit formation in the real world. Eur J Soc Psychol. 2010;40(6):998–1009.

27. Thrailkill EA, Trask S, Vidal P, Alcalá JA, Bouton ME. Stimulus Control of Actions and Habits: A Role for Reinforcer Predictability and Attention in the Development of Habitual Behavior. J Exp Psychol Anim Learn Cogn. 2018 Oct;44(4):370–84.

28. Akaishi R, Umeda K, Nagase A, Sakai K. Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron. 2014 Jan 8;81(1):195–206.

29. Neumann J von, Morgenstern O. Theory of Games and Economic Behavior. Princeton University Press; 1953. 774 p.

30. Bernoulli D. Exposition of a New Theory on the Measurement of Risk. Econometrica. 1954;22(1):23–36.

31. Markowitz H. The Utility of Wealth. J Polit Econ. 1952 Apr 1;60(2):151–8.

32. Kahneman D, Tversky A. Subjective probability: A judgment of representativeness. Cognit Psychol. 1972 Jul 1;3(3):430–54.

33. Loomes G, Sugden R. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. Econ J. 1982;92(368):805–24.

34. Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Massachusetts Institute of Technology Press; 2001. 460 p.

35. Li J, Daw ND. Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci Off J Soc Neurosci. 2011 Apr 6;31(14):5504–11.

36. Palminteri S, Pessiglione M. Chapter 23 - Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. In: Dreher J-C, Tremblay L, editors. Decision

Neuroscience [Internet]. San Diego: Academic Press; 2017 [cited 2020 Nov 30]. p. 291–303. Available from: http://www.sciencedirect.com/science/article/pii/B9780128053089000233

37. Lebreton M, Bacily K, Palminteri S, Engelmann JB. Contextual influence on confidence judgments in human reinforcement learning. PLOS Comput Biol. 2019 avr;15(4):e1006973.

38. 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 Sep 28;36(39):10016–25.

39. Pischedda D, Palminteri S, Coricelli G. The Effect of Counterfactual Information on Outcome Value Coding in Medial Prefrontal and Cingulate Cortex: From an Absolute to a Relative Neural Code. J Neurosci. 2020 Apr 15;40(16):3268–77.

40. Conen KE, Padoa-Schioppa C. Partial Adaptation to the Value Range in the Macaque Orbitofrontal Cortex. J Neurosci. 2019 May 1;39(18):3498–513.

41. Padoa-Schioppa C, Rustichini A. Rational Attention and Adaptive Coding: A Puzzle and a Solution. Am Econ Rev. 2014 May;104(5):507–13.

42. Gigerenzer G. The Bias Bias in Behavioral Economics. Rev Behav Econ. 2018 Dec 30;5(3–4):303–36.

43. Haselton MG, Nettle D, Andrews PW. The Evolution of Cognitive Bias. In: The Handbook of Evolutionary Psychology [Internet]. John Wiley & Sons, Ltd; 2015 [cited 2020 Jul 27]. p. 724–46. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470939376.ch25

44. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004 Dec 10;306(5703):1940–3.

45. Girden ER. ANOVA: Repeated Measures. SAGE; 1992. 88 p.

46. Sutton RS, Barto AG. Reinforcement Learning - An Introduction [Internet]. Mit Press; 1998 [cited 2017 Jun 8]. Available from: http://gen.lib.rus.ec/book/index.php?md5=C5BC41EB7F60C05D37473B4A9AC77A2A

47. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Class Cond II Curr Res Theory. 1972;2:64–99.

48. Hergueux J, Jacquemet N. Social preferences in the online laboratory: a randomized experiment. Exp Econ. 2015 Jun 1;18(2):251–83.

49. Kahneman D. Maps of Bounded Rationality: Psychology for Behavioral Economics. Am Econ Rev. 2003 Dec;93(5):1449–75.

50. Shavit T, Sonsino D, Benzion U. A comparative study of lotteries-evaluation in class and on the Web. J Econ Psychol. 2001 Aug 1;22(4):483–91.

51. Schoeffler M, Stöter F-R, Bayerlein H, Edler B, Herre J. An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results. In: ISMIR. 2013.

52. Reinecke K, Gajos KZ. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing [Internet]. Vancouver, BC, Canada: Association for Computing Machinery; 2015 [cited 2020 Jul 27]. p. 1364–1378. (CSCW '15). Available from: https://doi.org/10.1145/2675133.2675246