Dual range adaptation combines experiential scales to learn subjective values through efficient coding

Marwen Belkaid
ETIS UMR 8051
CY Cergy Paris Univ., ENSEA, CNRS
F95000 Cergy, France
ORCID 0000-0003-2358-2560

Uma Prashant Navare
ETIS UMR 8051
CY Cergy Paris Univ., ENSEA, CNRS
F95000 Cergy, France
ORCID 0000-0001-9686-8231

Akram El Wathek Meraghni
ETIS UMR 8051
CY Cergy Paris Univ., ENSEA, CNRS
F95000 Cergy, France
ORCID 0009-0008-1087-0209

***Abstract*—Reward magnitudes can vary substantially from a context to another. Value coding thus requires adapting neural responses to input changes to make use of their full dynamic range. Two models have recently been proposed aiming to provide a theoretical account for context-sensitive choice valuation in the framework of reinforcement learning. One model implements range adaptation relative to the minimum and maximum achievable outcomes, building on the notions of efficient coding and signal normalization. Another model borrows concepts from intrinsic motivation and posits that valuation integrates extrinsic and intrinsic rewards. Both models are able to fit human data well, but the latter was found to perform better under specific experimental conditions. In this paper, we propose that dual range adaptation over locally and globally experienced reward ranges combines the best of the two worlds. Our model benefits from the theoretical foundations and computational advantages of range adaptation. It also displays choice behaviors qualitatively similar to the model integrating intrinsic rewards. Model fitting suggests that data previously favoring the second model are best explained by our model. Our dual range adaptation model thus offers a new mechanistic account as well as testable predictions related to the representation of value for decision-making.**

***Index Terms*—Reinforcement learning, range adaptation, subjective value, efficient coding, computational modeling.**

I. Introduction

Valuation is a fundamental brain mechanism, allowing individuals to make decisions based on the subjective values associated with available options. In the framework of reinforcement learning, option values are typically learned from reward prediction errors, i.e. the difference between the obtained reward and predicted reward. Crucially, values processed in different behavioral conditions can vary substantially. One might choose a quick lunch from the cafeteria's menu, then select a meal at a fancy restaurant for dinner on the same day. We sometimes have to decide between goods worth a few euros, and other times worth hundreds or thousands of euros. How does the valuation system in the brain handle decisions at such variable scales?

In neural networks theories, *efficient coding* provides a compelling framework to address this question. This notion was originally developed in research on sensory processing and perception [1,2]. Simply put, neural systems have a limited physical capacity in terms of number of neurons and firing rates, yet they are required to represent a vast number of possible inputs. Therefore, they must adapt their responses depending on the context to efficiently encode inputs. *Normalization* provides a key mechanism to achieve this, by adjusting the gain of the neural responses so as to use the full available dynamic range and maximize sensitivity to changes in input contexts [3,4]. Divisive normalization, a particular case where responses within a population of neurons are divided by their summed activity, is widely found in the visual cortex and explains light adaptation in the retina [3]. But, normalization is so pervasive in the brain that it is considered as a canonical computation for neural systems [3,4]. In fact, these principles have been also increasingly applied to value coding for economic decisions [4,5]. For instance, neuroimaging studies in human and non-human primates provided evidence of efficient coding of values in the orbitofrontal cortex [6-8].

Substantial effort in the past years have been directed toward a clear theoretical account of context-sensitivity in human subjective valuation. Notably, a series of studies by Palminteri and colleagues combined experimental studies and reinforcement learning models to rigourously characterize the computations underlying this process [5,9,10]. Using experiments where reward contexts were manipulated, they provided strong evidence that human decisions are driven by normalized value representations. Specifically, Bavard and Palminteri showed that human data are better fit by their *range adaptation* model, where the obtained reward is scaled by the minimum and maximum possible outcomes [10]. Thus, while divisive normalization is prevailing in sensory processing, their results suggest that range normalization is a better model for subjective valuation.

Another recent study challenged the normalization account. Molinaro and Collins proposed an alternative model where the subjective value is a combination of extrinsic and intrinsic rewards, respectively represented by the reward obtained from the task and a self-generated reward signaling when the best outcome is achieved [11]. This so-called *intrinsically enhanced* model aimed to describe how internal goals, such as obtaining the highest reward, also shape valuation and

This work was financially supported by the Junior Professor Chair held by MB (ANR-22-CPJ1-0014-01), by CNRS through the MITI interdisciplinary programs, and by CY Initiative, a programme supported by the French government grant "Investissements d'avenir".

decisions [12]. This model successfully described a number of experimental datasets, as well as, or sometimes better than the range adaptation model [11]. It makes a compelling case for people's sensitivity to maximal rewards, which was not sufficiently accounted for by range normalization. Nonetheless, it suffers computational and theoretical weaknesses that have been raised by Bavard and Palminteri in a short commentary paper (preprint) [13]. On the one hand, the authors questioned whether the model actually describes an "intrinsic" reward in the sense that usually intended in intrinsic motivation theories [14,15]. On the other hand, the model is not scale-invariant, in the sense that parameter values for which the model is able to fit data from a certain experimental condition may not generalize to others.

In this paper, we propose to reconcile these two proposals with a new model implementing dual range adaptation. In this model, the same normalization mechanism operates at two experiential scales to adjust the obtained reward with respect to local and global outcome ranges; respectively referring to the immediately available outcomes, and those that are more generally observed within a task or a situation. We compare this model to the range adaptation [10] and intrinsically enhanced [11] models in the experiment designed by Molinaro and Collins [11] where the latter model previously showed better results. Data used in this paper was available at: https://osf.io/sfnc9/. Code to reproduce our results is available at: https://github.com/BelkaidLab/Belkaid_IJCNN2026.

II. Methods

A. Experimental data

Data modeled in this article are from the experiment designed by Molinaro and Collins [11], which was inspired by the paradigm developed by Bavard and Palminteri [9,10]. Fifty participants performed a decision task in which they had to choose between 2 or 3 stimuli presented to them on-screen (hereafter referred to as options) delivering different amounts of points (hereafter also referred to as reward magnitudes). The experiment included three phases: i) a learning phase, during which subjects received complete feedback about the points delivered by the chosen options as well as unchosen ones, ii) a test phase, during which no feedback was given, and iii) an explicit evaluation phase, which is not relevant for this paper.

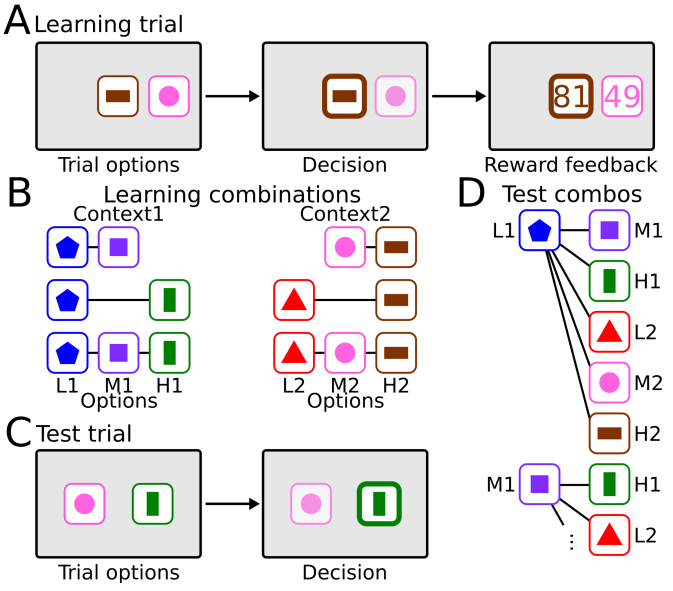

Fig. 1. Task description. **A)** Learning trials provide complete feedback after decisions. **B)** Options are split into 2 learning contexts from which specific combinations are presented, varying exposure to mid-value options. **C)** Test trials involve no feedback. **D)** All pairs of objects across contexts are tested.

This paradigm aims to examine the context-sensitivity of subjective valuation by manipulating subjects' exposure to choice options and their associated outcomes. Specifically, the task comprises 6 options grouped into 2 learning contexts. In the learning phase, options are only presented with other options from the same context. This is what enables manipulating the way participants subjectively learn the value of options based on the rewards observed together. Conversely, all possible combinations of option pairs are presented during the test phase, regardless of learning contexts. Choice rates among options in the test phase are the main measure of interest in this experiment as they implicitly reflect the subjective value that participants assign to each option.

Each learning context includes 3 options: one low-value, one mid-value and one high-value option, respectively noted Lx, Mx, and Hx, where x = 1 or 2 for the context number (Fig. 1). The amount of points delivered in each trial was randomly sampled from a gaussian distribution of standard deviation $\sigma = 2$ and mean $\mu = 14$, 50 and 86 respectively for the low-, mid- and high-value options. In addition to grouping options into learning contexts, the number of options presented per trials in the learning phase was varied: participants had to choose between 2 or 3 options. A key feature of this task is that only a subset of combinations were used such that the mid-value option was sometimes experienced as the best option in the first context and the worst in the second context. With this key feature, the task was specifically designed to distinguish the range adaptation model and the intrinsically enhanced model presented below (see Sections II.C. and II.D.). Overall, participants completed a total of 180 trials: 120 learning trials (20 repetitions of each learning combination) and 60 test trials (4 repetitions of all possible combinations of option pairs). Trials of the learning phase were randomized such that learning contexts were randomly interleaved.

B. General modeling framework

All models simulated in this paper are formulated in a common reinforcement learning framework. In this framework, models make decisions between available options s using the classical softmax rule where the probability of choosing the option s_i depends on its estimated value $V(s_i)$:

$$P(s_i) = \frac{e^{\beta \cdot V(s_i)}}{\sum_k e^{\beta \cdot V(s_k)}} \quad (1)$$

The inverse temperature parameter $\beta > 0$ controls the exploitation/exploration trade-off of the model, i.e. whether the

models mostly selects options that have the highest value or makes decisions more randomly.

Models learn the estimated option values $V(s_i)$ based on observed rewards $\mathcal{R}$ using the Rescorla-Wagner delta rule with a learning rate $\alpha \in [0, 1]$:

$$\Delta V(s_i) = \alpha(\mathcal{R} - V(s_i)) \tag{2}$$

As explained above, this article aims to model data from a behavioral experiment in which, during the learning phase, participants where given complete (including counterfactual) feedback. It is thus assumed that in each trial, models update the value of both chosen and unchosen options. While previous studies used different learning rates for chosen and unchosen options [10,11], the current work uses a single parameter for both for the sake of simplicity.

In a classical reinforcement learning model, $\mathcal{R}$ is assumed to be equal to the *objective* reward R_{obj}, which in the behavioral experiments considered here amounts to the (raw) magnitude of the feedback received at the end of each trial. In contrast, recent studies showed that models that capture the context- or goal-dependent character of value learning better account for human *subjective* valuation [10,11]. The models described below thus propose different formulations for the term $\mathcal{R}$ aimed to allow learning subjective values.

C. The Range Adaption model

The range adaptation model (RA) was proposed by Bavard and Palminteri [10]. The core idea of this model is that rewards are rescaled based on the range of potential rewards in a context-dependent manner, i.e. considering the minimum and maximum reward magnitudes of the present context, noted C_{min} and C_{max} respectively. This normalized reward is then raised to the power $z > 0$:

$$\mathcal{R}_{\text{[RA]}} = \left(\frac{R_{obj} - C_{min}}{C_{max} - C_{min}}\right)^z \tag{3}$$

The z parameter aims to capture non-linearities observed in behavioral experiments [10]. For a normalized reward $\in [0, 1]$, z values below and above 1 respectively produce a concave and a convex function. While there is individual variability, most subjects were fit with $z > 1$ such that mid-value options were typically assigned lower values [10].

Note that it is possible to implement a model that also learns these minimum and maximum value through trial and error [9]. However, previous work suggests that this does not substantially change the performance of the model, while possibly disadvantaging the model due to an additional learning rate parameter. Therefore, for the sake of simplicity , C_{min} and C_{max} are simply hard-coded based on the context [10], assuming that complete feedback allows participants to quickly know these values.

D. The Intrinsically Enhanced model

The intrinsically enhanced model (IE) was developed by Molinaro and Collins [11]. It proposes an alternative, goal-centric explanation for context-sensitivity in human valuation where obtaining the best possible outcome yields an "intrinsic reward" boosting the value of a choice as a consequence of having achieved the goal of maximizing outcomes. To do so, it defines the subjective value as a linear combination of the objective, extrinsic reward R_{obj} and a binary signal B equal to 1 if the highest reward is obtained and 0 otherwise. A weighted average is thus obtained using a weighting parameter ω:

$$\mathcal{R}_{\text{[IE]}} = \omega B + (1 - \omega) R_{obj} \tag{4}$$

Note that the model theoretically combines a binary value with an unconstrained real number, possibly orders of magnitude greater. This implies that ω could also be any real number, making the model scale-dependent as pointed out previously [13]. In practice, the extrinsic reward R_{obj} was divided by a task-dependent factor in order to rescale it [11]. We use the same workaround in this paper when simulating this model.

E. Our Dual Range adaptation model

The dual range adaptation model (DR) proposes to mitigate the disadvantages of the two previous models by considering that range normalization occurs at two different experiential scales. Raw rewards (i.e. delivered by the task) are thus rescaled with respect to the minimum and maximum rewards both locally within the trial, noted L_{min} and L_{max}, and globally for the entire experiment, noted G_{min} and G_{max}. The subjective value is then defined as the weighted average of the two terms using the weighting parameter $\omega \in [0, 1]$:

$$\mathcal{R}_{\text{[DR]}} = \omega \left(\frac{R_{obj} - L_{min}}{L_{max} - L_{min}}\right) + (1 - \omega) \left(\frac{R_{obj} - G_{min}}{G_{max} - G_{min}}\right) \tag{5}$$

For simplicity, the minimum and maximum values are hard-coded like in the RA model; assuming that it would also easily be possible to learn these values from trial and error.

This model shares similarities with the IE model while taking advantage of range normalization. When the best option of the trial is selected, the trial-based normalization operates similarly to the "intrinsic reward" yielding the maximum value of 1. Thereby, the best options are boosted without the need for a hand-crafted binary signal, while also taking into account the rewards associated with other options in the current trial. At the same time, the global normalization places the current reward in the context of all possible outcomes. This term essentially operates like the "extrinsic reward" in the IE model since the latter is rescaled in practice as mentioned in the previous section. Like the IE model, the linear combination of the two normalized rewards is able to account for the non-linear valuation observed in human experimental data. All model features are summarized in Table I.

F. Forward model simulations

All models were first simulated to get model predictions about the choice behaviors in the task. Each model was simulated 100 times with a different set of parameter values. Parameters were randomly sampled from prior distributions based on priors and fitting results from previous studies

TABLE I
SUMMARY OF MODEL FEATURES

Features	RA	IE	DR
sensitivity to reward range	X		X
sensitivity to maximal reward		X	X
reward scale-invariance	X		X
valuation non-linearity	X	X	X

[10,11]. The learning rate α was drawn from a $Beta(1.3, 3.0)$ distribution. The inverse temperature β was drawn from a $Gamma(1.2, 5.0)$ distribution. The non-linearity parameter z of the RA model was drawn from a $Gamma(3.0, 1.0)$ distribution. Last, the weighting parameters of the IE and DR models were drawn from a $Beta(1.3, 1.3)$ distribution. Note that all models have the same number of parameters $= 3$.

G. Model fitting and comparison

Data were fit for both the learning and test phases. Similar to Molinaro and Collins [11], we used the Hierarchical Bayesian Inference (HBI) method proposed by Piray and colleagues [16] which performs both parameter estimation and model comparison concurrently. In this paper, we were primarily interested in comparing the three models described above and whether the dual range adaptation model better explains the data. Because it has been suggested to improve the results of the HBI procedure [16], we also included a fourth model implementing a Win-Shift (WS) strategy. This strategy consists in switching behavior, i.e. selecting a different action, after a successful trial. Since the experiment modeled here involved rewards of different magnitudes for which the subject receives full feedback, we translated the WS strategy into a model assigning a value of 0 to the action that had the highest reward and a value of 1 to the others. Because the win-shift strategy focuses on actions, the WS model assigns values to the left, middle and right choice regardless of the options presented, unlike the other models. Values are set to 0 or 1 directly, without using the learning rule (2). Thus, the model has only one single parameter, which is the inverse temperature β of the softmax decision function. The WS was included in the HBI procedure to reduce the effect of outliers [16]. The idea is that this simple model could capture some of the noise to avoid penalizing the target models we aim to differentiate.

The HBI method involves two steps. The first step uses Laplacian approximation to separately fit each model to the data. Laplacian approximation requires a normal prior distribution for all model parameters. Therefore, as recommended by the authors of the method [16], all priors were set to a mean of 0 and a variance of 6.25 which allows covering a wide range of parameters. Then, values for the learning rate α and weighting parameter ω, which need to be in $[0, 1]$, were transformed using a sigmoid function. Values for the inverse temperature β and the non-linearity parameter z, which on the other hand need to be positive, were transformed using an exponential function.

The second step of HBI uses variational Bayesian inference over all models based on the Laplacian estimates. The algorithm iteratively updates parameter estimates at the individual and group levels. It also estimates model responsibilities, which are the probabilities that each model is generating the individual data. Based on this, the method outputs three useful metrics for model comparison: a) the model frequency, which represents the fractions of individual subjects explained by a model, b) the exceedance probability, which is the probability that a model is more commonly expressed in the whole sample that the other models, and c) the protected exceedance probability, which provides a more conservative measure which takes into account that observed differences are due to chance.

Another feature of HBI is to provide a variant of the classical one-sample t-test. As indicated by the authors [16], hierarchical parameter estimation prevents from using classical statistical tests due to the fact that parameter estimates provided by HBI are not independent but rather regularized by the population variance. The HBI t-test overcomes this limitation and allows assessing whether a parameter is statistically different from a specific value.

III. RESULTS

A. Model predictions for choice rates

We simulated the DR, IE and RA models with randomly sampled parameters to observe the behavior that each of them generally predicts. We were specifically interested in choice rates as an index of the relative values of the options for each model. Results are shown in Fig. 2. Overall, predictions for the IE and RA models are in line with simulations reported in previous studies [10,11], though small differences may be found due to a slightly different choice of prior distributions for some parameters. While the whole choice pattern is relevant, the comparison between choice rates of the two mid-value options M1 and M2 is particularly informative. Because in this task, low-, mid- and high-value options had the same average reward magnitudes across learning contexts, the RA model displayed similar choice rates for M1 and M2 whose subjective values were indeed normalized with respect to the

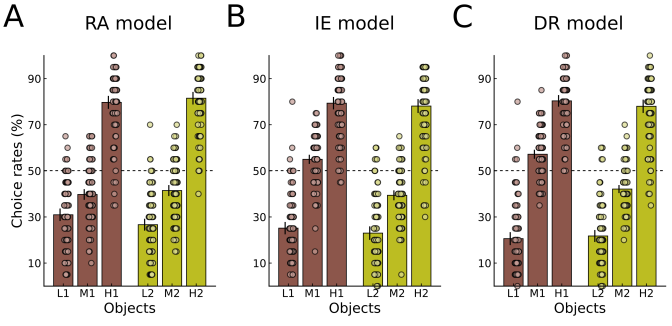

Fig. 2. Choice rates in simulated models. **A)** RA model produces similar choice rates between contexts 1 and 2. **B-C)** IE and DR models both select M1 more often than M2. Dashed lines represent the baseline choice rate for M1 and M2 with classical reinforcement learning (no subjective valuation).

same range (Fig. 2.A). It is worth noting that a classical model using objective reward magnitudes for learning (i.e. where $\mathcal{R} = R_{obj}$) would produce a similar choice pattern; except that the rate for M1 and M2 would be around 50% due to the absence of non-linearity introduced with the parameter z.

On the other hand, the IE model showed a higher choice rate for M1 compared to M2 (Fig. 2.B), due to the fact that the former was more frequently experienced as the best option during the learning phase. This pattern is what allowed IE to better fit the data in Molinaro and Collins's study [11]. Importantly, the DR model was able to exhibit a similar pattern, selecting M1 more frequently than M2 (Fig. 2.C). This demonstrates that our model is able to show sensitivity to maximal rewards (Table 1), a feature that is required in order to account for the experimental data. Thus, DR and IE behave in a qualitatively similar way in this task.

B. Model comparison based on experimental data

Upon verifying that our model was able to behave as expected, we performed hierarchical Bayesian inference to fit and compare all competing models against experimental data at both the individual and group levels. We first looked at model responsibilities, which estimate the probability that each model describes each individual data. Results showed that DR was the best model for all participants with a probability above 60%, followed by the IE model (Fig. 3.A). Since results were rather homogeneous across participants, they are directly reflected in the model frequencies (Fig. 3.B), which essentially aggregate responsibilities at the population level (DR $= 0.65$, IE $= 0.345$, RA $< 10^{-3}$, WS $= 0.005$). We then examined exceedance probabilities, indicating the probability that each model is most commonly expressed in the population. Results strongly favored the DR model with a probability approaching 1 (DR $= 0.985$, IE = 0.015, RA = WS = 0; Fig. 3.C). The protected exceedance probabilities, adjusting these probabilities based on random effects, produce the same results (DR $= 0.985$, IE = 0.015, RA $< 10^{-5}$, WS $< 10^{-5}$; Fig. 3.D), thus suggesting that the advantage of the DR model was not driven by chance.

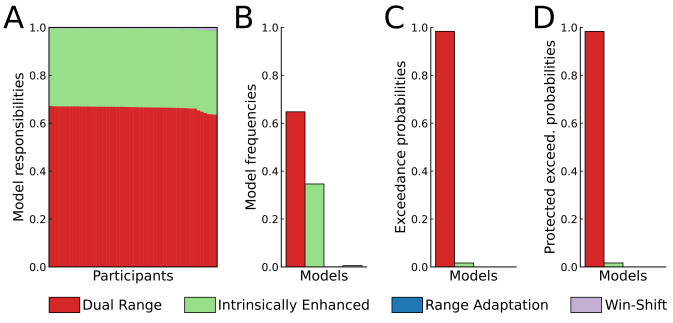

Fig. 3. Model comparison against data. **A)** DR model has the highest probability of describing data from each subject. **B)** Resulting model frequencies on the entire dataset. **C-D)** Simple and protected exceedance probabilities indicate that DR is very likely the best model for the entire dataset.

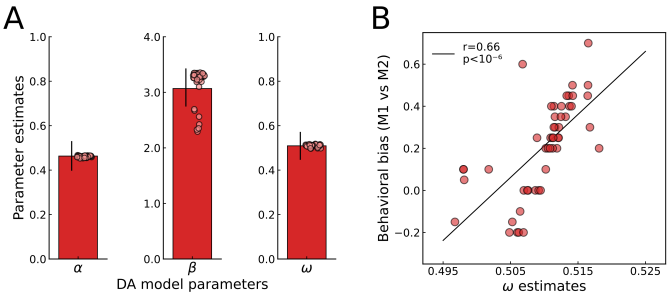

Fig. 4. Estimated parameters for the DR model. **A)** Individual and group-level estimates for α, β and ω. **B)** Individual estimates for ω positively correlate with the difference in choice rates between M1 and M2, which can be used as a behavioral index of valuation subjectivity.

C. Parameter estimates for the Dual range adaptation model

To further assess the performance of the DR model, we looked into the estimated parameter values resulting from the hierarchical Bayesian inference (Fig. 4.A). Average estimates were the following: $\mu_\alpha = 0.46$, $\mu_\beta = 3.06$, and $\mu_\omega = 0.50$. The weighting parameter ω is of particular interest here, as it could reflect a stronger drive from the local or global reward ranges. But with an average estimate at 0.5 and a low variance, results rather seem to qualitatively indicate that valuation was balanced. This was further confirmed quantitatively by the HBI t-test indicating that ω estimates were not significantly different from 0.5 ($t(33.38) = 0.15$, $p = 0.88$).

Beyond its specific value, another way in which the parameter ω can be informative is in how it relates to certain behavioral metrics. Specifically, the difference in choice rates between the M1 and M2 options can be taken as a behavioral index of subjective valuation. In the IE model, it was previously shown that the weight ω correlated positively with this metric [11]. Likewise, we found that despite the low variance in the obtained parameter estimates, the weighting parameter ω of the DR model was significantly correlated with the difference between M1 and M2 choice rates ($r = 0.66$, $p < 10^{-6}$; Fig. 4.B). This suggests that our DR model relies on parameters that can provide useful insights into experimental data and underlying computational mechanisms.

IV. Discussion

Originally developed in research on sensory processing, efficient coding and normalization are useful concepts for understanding value representation in economic decision-making [4,5]. The range adaptation model (RA) has solid theoretical foundations grounded in this framework and provides a compelling high-level account for how the brain manages rewards of variable magnitudes. Reward normalization also offers good computational properties enabling model parameters to be scale-invariant. The RA model successfully captured a series of experimental data [10]. However, as shown by Molinaro and Collins [11], it was challenged under some specific experimental conditions: namely when a mid-value option could often be experienced as the most rewarding one. The intrinsically enhanced model (IE) on the other hand relies

on the idea that achieving the best outcome is intrinsically rewarding. By combining the objective, extrinsic reward with a binary self-generated reward signal, the IE model exhibits both sensitivity to maximal rewards and choice non-linearity, allowing it to fit the data better than the RA model [11].

Despite fitting experimental data well, the IE model also has its pitfalls. While the use of the term *intrinsic reward* was itself challenged by Bavard and Palminteri [13], a more relevant issue for this paper is that the model is not scale-invariant. Specifically, the expression of the subjective values combines a binary signal encoding an intrinsic reward with rewards delivered by the task. Crucially, the magnitude of the latter may greatly vary across experimental conditions, and across real-life situations. As a result, the weighting parameter used in the model is highly sensitive to reward magnitudes [13]. In practice, instead of using the actual (raw) obtained reward in their model simulations, Molinaro and Collins rescaled it such that the extrinsic reward also had a value between 0 and 1 [11]. Without this workaround, it would not be possible to consistently compare with models using normalized rewards or, perhaps more importantly, across datasets involving rewards of different magnitude scales. The authors emphasize that this is done simply for convenience and that one could alternatively rescale the intrinsic reward and the inverse temperature parameter [11]. Critically, this in itself is an argument in favor of the implementation of some form of reward normalization.

In this paper, we proposed a novel model alleviating the issues raised above. Our dual range adaptation (DR) model combines features from the two previous models, by essentially reformulating the key idea of a "bonus" reward from IE in the framework of range normalization. Our results thus showed that it was able to exhibit qualitatively similar behavior to IE on the task where it previously performed better than RA [11], but also to better fit experimental data from this task than the other models. In addition, we found that the weighting parameter ω provided a model correlate of a behavioral index of valuation subjectivity measured from the relative choice rates of mid-value options. Overall, DR thus relies on solid theoretical concepts, enables consistent comparisons with other models and across experiments, and offers potential mechanistic explanations for the cognitive and computational processes underlying economic decisions.

The DR model makes a series of useful predictions for future research. First, instead of evaluating choice options within independent range contexts, the model postulates that rewards are adjusted with respect to "locally" and "globally" experienced ranges. In more realistic implementations where minimum and maximum values would have to be learned [9], this could be achieved with different learning rates for example. Future work could use a task involving more than two learning contexts to experimentally test whether choices reflect context-specific valuation or the predicted local versus global range adaptation. Second, the model posits that subjective valuation integrates reward information at two, local versus global, experiential/time scales. One could thus predict that some situations might favor one over the other. For instance, this could be tested with experimental conditions designed to hinder global integration through cognitive overloading or volatile reward ranges. Last, the model assumes that range normalization can be neurally implemented. However, unlike divisive normalization which has been extensively studied, range adaptation does not have an obvious neural counterpart. Thus, future research should seek to develop a neural implementation of the DR model to test its biological plausibility.

Author contributions

MB and UPN conceptualized the model. AEM ran initial simulations. MB performed the final simulations and analyses, and wrote the paper. All authors reviewed the manuscript.

References

[1] F. Attneave, "Some informational aspects of visual perception." *Psychological Review*, vol. 61, no. 3, pp. 183–193, 1954.

[2] H. B. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," in *Sensory Communication*, W. A. Rosenblith, Ed. The MIT Press, 2012, pp. 216–234.

[3] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.

[4] K. Louie and P. W. Glimcher, "Efficient coding and the neural representation of value," *Annals of the New York Academy of Sciences*, vol. 1251, no. 1, pp. 13–32, 2012.

[5] S. Bavard, M. Lebreton, M. Khamassi, G. Coricelli, and S. Palminteri, "Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences," *Nature Communications*, vol. 9, no. 1, p. 4503, 2018.

[6] C. Padoa-Schioppa, "Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex," *The Journal of Neuroscience*, vol. 29, no. 44, pp. 14 004–14 014, 2009.

[7] K. M. Cox and J. W. Kable, "BOLD Subjective Value Signals Exhibit Robust Range Adaptation," *The Journal of Neuroscience*, vol. 34, no. 49, pp. 16 533–16 543, 2014.

[8] H. Yamada, K. Louie, A. Tymula, and P. W. Glimcher, "Free choice shapes normalized value signals in medial orbitofrontal cortex," *Nature Communications*, vol. 9, no. 1, p. 162, 2018.

[9] S. Bavard, A. Rustichini, and S. Palminteri, "Two sides of the same coin: Beneficial and detrimental consequences of range adaptation in human reinforcement learning," *Science Advances*, vol. 7, no. 14, p. eabe0340, 2021.

[10] S. Bavard and S. Palminteri, "The functional form of value normalization in human reinforcement learning," *eLife*, vol. 12, p. e83891, 2023.

[11] G. Molinaro and A. G. E. Collins, "Intrinsic rewards explain context-sensitive valuation in reinforcement learning," *PLOS Biology*, vol. 21, no. 7, p. e3002201, 2023.

[12] G. Molinaro and A. G. Collins, "A goal-centric outlook on learning," *Trends in Cognitive Sciences*, vol. 27, no. 12, pp. 1150–1164, 2023.

[13] S. Bavard and S. Palminteri, "Does the notion of intrinsic reward explain context-dependent valuation in reinforcement learning?", OSF *(preprint)*, 2024.

[14] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? A typology of computational approaches," *Frontiers in Neurorobotics*, vol. 1, 2007.

[15] G. Baldassarre, T. Stafford, M. Mirolli, P. Redgrave, R. M. Ryan, and A. Barto, "Intrinsic motivations and open-ended development in animals, humans, and robots: an overview," *Frontiers in Psychology*, vol. 5, 2014.

[16] P. Piray, A. Dezfouli, T. Heskes, M. J. Frank, and N. D. Daw, "Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies," *PLOS Computational Biology*, vol. 15, no. 6, p. e1007043, 2019.