Tracking what matters: A decision-variable account of human behavior in bandit tasks

Vishwajeet Agrawal (vishwa.grawal@gmail.com)
Computer Science Department, IIT Delhi

Pradeep Shenoy (shenoypradeep@google.com)
Google Research India

Abstract

We study human learning & decision-making in tasks with probabilistic rewards. Recent studies in a 2-armed bandit task find that a modification of classical *Q*-learning algorithms, with outcome-dependent learning rates, better explains behavior compared to constant learning rates. We propose a simple alternative: humans directly track the decision variable underlying choice in the task. Under this *policy learning* perspective, asymmetric learning can be reinterpreted as increasing confidence in the preferred choice. We provide specific update rules for incorporating partial feedback (outcomes on chosen arms) and complete feedback (outcome on chosen & unchosen arms), and show that our model consistently outperforms previously proposed models on a range of datasets. Our model and update rules also add nuance to previous findings of perseverative behavior in bandit tasks; we show evidence of *outcome-dependent choice perseveration*, i.e., that humans persevere in their choices unless contradictory evidence is presented.

Keywords: 2-armed bandits; reinforcement learning; decision-making; optimism bias; confirmation bias

Introduction

How do humans and other animals learn about actions & rewards from probabilistic outcomes? A popular experimental paradigm for studying this question is the 2-armed bandit task (Figure 1), involving a repeated choice between two actions associated with some predetermined (but unknown) probability of rewards. Classical reinforcement-learning algorithms for this task such as the Rescorla-Wagner model (Rescorla, 1972) track running estimates of average rewards associated with actions, updated using prediction errors with respect to observed outcomes. However, empirical data and model fits to behavior (Palminteri et al., 2017) suggest that RW models do not fully capture human behavior; instead, learning appears to be *biased towards* positive outcomes (Sharot et al., 2011), and towards reinforcing current decisions (Palminteri et al., 2017).

In this paper, we propose a simple reframing of the learning goal in 2-armed bandit tasks: Instead of maintaining estimates of average rewardability of the two arms, we directly maintain a *decision variable* encoding the better choice, in a manner similar to a large body of work in perceptual decision-making (see e.g., Platt & Glimcher (1999)). This choice of representation is both more task-relevant and easier for the brain to maintain and update over time; indeed, recent work argues that the brain may represent *action policy* instead of value (Hayden & Niv, 2021) over a broad range of contexts.

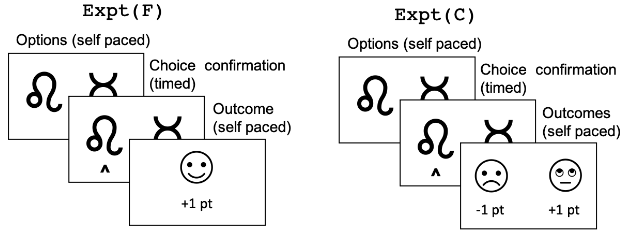

Figure 1: Experiment protocol: Subjects choose between two abstract options, and receive a higher or a lower outcome with predetermined probabilities associated with each option. In `Expt(P)`, they see an outcome on chosen arm, and in `Expt(C)`, they see outcomes on both chosen and unchosen options. Figure adapted from Palminteri et al. (2017)

We provide update rules for incorporating outcomes into the decision variable, and compare against previous models for this task on a range of datasets involving feedback on both factual (chosen action) and counterfactual (unchosen action) outcomes. Our model consistently outperforms previously proposed learning rules across all datasets and experiments.

Our proposal for counterfactual learning separates out *ambiguous* scenarios (e.g., where both arms get same reward), from *unambiguous* scenarios where rewards on the arms are different. While unambiguous feedback moves the decision-variable in the relevant direction, we show that in equal reward situations, the current choice is reinforced. We therefore propose a notion of *outcome-dependent* perseveration of choice, in contrast to previous proposals that suggest a bias term tracking choice autocorrelation over time, independent of outcome (see e.g., Akaishi et al. (2014)).

Tasks and previous models

We summarize key results in the literature on human reward-based learning in highly controlled laboratory experiments involving probabilistic rewards, specifically 2-armed bandits involving two possible outcomes (a higher rewarding/ less punishing outcome and a lower rewarding/ more punishing outcome) (Figure 1). We cover experiments with partial feedback (hereafter referred to as `Expt(P)`), and complete feedback (factual and counterfactual both, abbreviated as `Expt(C)`). Human behavior in this task has been analyzed using the following reinforcement learning methods:

Rescorla-Wagner (RW): This model (Rescorla, 1972) tracks rewards associated with each arm using two Q-values, which are updated using a delta rule every time an outcome associated with an arm has been observed. Concretely, $q_c \leftarrow q_c + \alpha(r - q_c)$, where arm c has been observed to yield reward r, and α is a learning rate parameter. The two Q-values are expected to converge to average reward of the arms. The following modifications to this classical Q-learning approach have been shown to better fit human behaviour.

Valence-driven learning (VM): Here, separate learning rates are associated with positive and negative prediction errors (equivalently higher and lower outcomes respectively). The update rule in VM is:

$$q_c \leftarrow q_c + (r - q_c) * \begin{cases} \alpha_p & \text{if } (r - q_c) > 0 \\ \alpha_n & \text{if } (r - q_c) < 0 \end{cases} \tag{1}$$

Model fits confirm that VM better explains behavior in the bandit task compared to RW (Lefebvre et al., 2017). The asymmetry in learning ($\alpha_p > \alpha_n$) has also been interpreted as "optimism" (see e.g., Sharot et al. (2011)). Importantly, the Q-values no longer converge to true reward probabilities of the arms.

Confirmation-disconfirmation (CD): Palminteri et al. (2017) argue that VM is only a partial explanation, due to a limitation of experimental design where only chosen options' outcomes are shown. When *counterfactual evidence* (Fig 1, Expt(C)) is presented on each trial, there is an interaction between chosen-unchosen arms, and higher-lower outcome (equivalently, prediction error). The update equation for CD is:

$$\begin{aligned} q_c &\leftarrow q_c + (r_c - q_c) * \begin{cases} \alpha_{con} & \text{if } (r_c - q_c) > 0 \\ \alpha_{dis} & \text{if } (r_c - q_c) < 0 \end{cases} \\ q_u &\leftarrow q_u + (r_u - q_u) * \begin{cases} \alpha_{dis} & \text{if } (r_u - q_u) > 0 \\ \alpha_{con} & \text{if } (r_u - q_u) < 0 \end{cases} \end{aligned} \tag{2}$$

Specifically, they find α_{con} significantly higher than α_{dis}, i.e. learning rates are higher on *higher outcome* for chosen arms ($r_c - q_c > 0$) but *lower outcome* for unchosen arms ($r_u - q_u < 0$). This is interpreted by the authors as a *confirmatory* update to the Q-values: outcomes confirming the validity of the choice (higher outcome on chosen or lower outcome on unchosen arm) show a higher learning rate than disconfirmatory outcomes.

Estimation under nonstationarity (DBM): In recent work, Zhou et al. (2020) propose a dynamic belief updating model where reward probabilities are estimated under an assumption that at any instant, with fixed probability, the reward probability associated with each arm may be "reset", i.e., redrawn from a fixed prior distribution over reward probabilities. They only model Expt(P), i.e., partial feedback experiments.

Let $p(\theta_i)$, be the probability density function (pdf) for the random variable θ_i denoting the probability of arm i giving the higher outcome in some particular arm-pairing context. Suppose an arm c is chosen in a particular trial, and its outcome is observed. The pdf for *unseen arms* (unchosen arm in the current trial, as well as arms not included in the present trial) are "reset" as,

$$p(\theta_i) \leftarrow (1 - \alpha p_0(\theta)) + \alpha p(\theta_i) \; \forall i \neq c \tag{3}$$

where $p_0(\theta)$ is the fixed prior distribution to which arms are reset. The pdf for seen (or chosen) arm is updated based on Bayes rule, incorporating the seen outcome (r_c) as,

$$p(\theta_c) \leftarrow \lambda p(r_c|\theta_c) p(\theta_c) \tag{4}$$

where λ is the normalizing constant. The final Q-values for making a choice are estimated as the expected value of each pdf, i.e. $q_i = E[p(\theta_i)]$. For the prior $p_0(\theta)$, the authors use a beta distribution, $Beta(s \cdot p, s \cdot (1 - p))$ with a fixed scale parameter s for all participants, and a free parameter p (mean of the beta distribution) for each participant.

The authors show that this model is a better fit than VM for one experiment from Lefebvre et al. (2017) where factual evidence alone is presented; they also remark that in the specific experiment modeled, this model mimics *choice perseveration* (a tendency to continue choosing previously chosen options (Akaishi et al., 2014)). We note in passing that this formulation and finding are closely related to previous proposals of *forgetting rates* in Q-value estimation (see e.g., Barraclough et al. (2004); Ito & Doya (2009)). They do not analyze other open source datasets with similar experimental conditions, or experiments where counterfactual evidence is presented.

Decision Variable Tracking

In all experiments studied here, arms have binary outcomes: a higher reward and a lower reward. In our models, we represent these binary outcomes as 1 and -1, respectively.[1]

Instead of maintaining an estimate of average rewards of the two arms, we propose to track a decision variable, $d \in [-1, 1]$ encoding the identity of the higher rewarding arm, with $d > 0$ (or < 0) signaling that arm 1 (resp., arm 2) is more rewarding than the other arm. To preserve symmetry in the mathematical form of the update equations, we show update equations for d in terms of $d_1, d_2 \in [-1, 1]$, where d_1, d_2, and d are related as follows: $d \overset{\text{def}}{=} d_1 \overset{\text{def}}{=} -d_2$. d_2 is simply the reverse of d_1, i.e. positive when arm 2 is believed to be better. The choice on the next trial depends on the softmax of d_1, d_2. The magnitude of d can also be interpreted as the degree of confidence that a specific arm is more rewarding than the other.

Partial feedback experiment

We start with the update equation for incorporating an observed outcome r_c after having chosen arm $c \in \{1, 2\}$. Evidence for the chosen arm being better (d_c) is updated towards

[1] This representation of relative valuation instead of absolute values of outcomes does not affect quality-of-fit of previous models (RW, VM, CD and DBM) in the experiments modeled here.

-1 on seeing an outcome of -1, and towards 1 on an outcome of 1. The update equation for the model, which we call DV1, are:

$$d_c \leftarrow d_c + \begin{cases} \alpha_p(1-d_c) & : r_c = 1 \\ \alpha_n(-1-d_c) & : r_c = -1 \end{cases} \tag{5}$$

We model separate rates for positive and negative outcomes. When $\alpha_p > \alpha_n$, there is a *tendency towards certainty* in the current choice, i.e., for d to converge towards one of the choices. Indeed, in asymptote, it is desirable for a decision variable such as d to converge towards a particular choice regarding which arm is better. If α_p were set equal to α_n, the decision variable would converge to $E(r_1 - r_2)$, i.e., the same as the mean value of $(q_1 - q_2)$ from a basic Q-learning (RW) model; this would gain lesser reward in asymptote under a stochastic policy (also see Lefebvre et al. (2020)). We find in our experiments that fitted $\alpha_p > \alpha_n$ for most subjects (see Figure 3).

Differences with VM: DV1 is different from VM, since it only represents one quantity at any given time. In contrast, the choice in VM depends on $\Delta q = q_c - q_u$, which depends on both q_c and q_u, and cannot be represented by a single equivalent update. One possible interpretation of DV1 in terms of prior work is that it is mathematically equivalent to CD update rule with a *hallucinated* outcome on unchosen arm that is perfectly anti-correlated with chosen arm outcome (i.e., $r_u = -r_c$).

Complete feedback experiment

In Expt(C), there are 4 possible outcomes for the paired outcomes on chosen and unchosen arms (r_c, r_u). When the two arms have opposite outcomes, the decision variable should naturally reinforce towards the arm with higher one. In these two cases, the update is similar to DV1; towards 1 with a learning rate α_p, or towards -1 with a learning rate α_n, depending on which arm had higher outcome. But what should the update be, when both the outcomes are same? We consider two hypotheses – update d_c either towards $t = 1$ or -1, i.e. "increase" or "decrease" confidence with respect to current choice:

$$d_c \leftarrow d_c + \alpha_0(t - d_c) : r_c = r_u, t \in \{-1, 1\} \tag{6}$$

Model comparison between the two version showed 100% support for $t = 1$ (for each participant, $t = 1$ version had lower NLL compared to $t = -1$), a finding we term as *outcome-dependent choice perseveration*: increased confidence in current choice in the absence of unambiguous reward signals from the environment. Additionally, we find $\alpha_p > \alpha_n, \alpha_0$ for each participant, again reflecting *tendency towards certainty*. Finally, there was a significant correlation between the three parameters indicating degeneracy, with mean value of α_0 very close to mean value of $(\alpha_p - \alpha_n)$. So, we choose $\alpha_0 \equiv \alpha_p - \alpha_n$, and $t = 1$, and refer to the reduced model as DVc.

$$d_c \leftarrow d_c + \begin{cases} \alpha_p(1-d_c) & r_c > r_u \\ \alpha_n(-1-d_c) & r_c < r_u \\ (\alpha_p - \alpha_n)(1-d_c) & r_c = r_u \end{cases} \tag{7}$$

Differences with CD: We note that CD and DVc differ primarily on same-outcome trials; if outcomes were perfectly anti-correlated, DVc CD models would become exactly equal. However, on same-outcome trials, the CD updates for (q_c, q_u) cannot be reduced to an update of Δq. In fact, unlike DVc, the quantity Δq in the CD model may reduce, or increase, after same-outcome trials depending on the actual values of (q_u, q_c) before that trial.

Partial feedback: an alternate model

Finally, we consider another model for Expt(P) which we refer to as DV2,

$$d_c \leftarrow d_c + \alpha_1(r_c - d_c) + \alpha_2(1 - d_c) \tag{8}$$

(Also, $\alpha_1 + \alpha_2 \leq 1$ because $|d_c| \leq 1$).
This model can be understood as a simplification of the DVc model proposed for Expt(C). If the unchosen arm was revealed, and its outcome were opposite to the factual outcome r_c, the update in DVc is towards r_c. These cases ($r_c \in \{1, -1\}, r_u \neq r_c$) are reflected in the above update equation as the first term, $\alpha_1(r_c - d_c)$. In case the two outcomes are equal, the update should be towards 1, and is captured by the second term above, $\alpha_2(1 - d_c)$. As in DVc, this second term reflects a *tendency towards certainty*, since for $\alpha_2 > 0$, the decision is reinforced towards the current choice, and the mean value of d_c moves closer to 1. The DV2 model combines both of these update terms into a single update equation shown above.

Results

Datasets

We used open-source data from previous studies (Palminteri et al., 2017; Chambon et al., 2020; Lefebvre et al., 2017; Lebreton et al., 2019)[2]. The datasets covered a range of different experimental conditions where a number of factors were varied, including: 1) the probability of reward associated with each arm, and their pairing, e.g., both high, both low, or contrasting, 2) the nature of reinforcement – reward, punishment or both, 3) presence of a reversal condition (arm reward probabilities swapped in the middle of the block), etc. The experiments are summarized in Table 1. We do not break down our analyses or customize models to each specific condition; the details are presented for completeness, and to demonstrate the wide range of scenarios under which model comparisons show consistent gains.

[2] In some studies (Chambon et al., 2020; Lebreton et al., 2019), trials related to Expt(P) and Expt(C) paradigms were interleaved (with different arm pairs & visual stimuli for each), so we separated trials out into Expt(P) and Expt(C) datasets. Also, from Chambon et al. (2020) we ignore "forced-choice" blocks that involved implementing a computer chosen choice.

Table 1: **Summary of datasets.** The two possible outcomes for any arm are ≥ 0 in Pos and ≤ 0 in Neg. In Mix, the higher outcome is > 0 and the lower is < 0. Reward probabilities (contingencies): one arm > 0.5 and other < 0.5 in C, both arms > 0.5 in H, equal to 0.5 in S, equal but > 0.5 in sH, equal but < 0.5 in sL. Contingencies of the two arms are contrasting but are reversed in the middle of the block in R. "Length" is # trials per pair of arms in a single learning sequence (or block).

Dataset	#subjects	#trials	Length	Outcomes	Contingencies	Reference
`Palm-E1`	20	192	24	Pos	C, S, R	Expt 1, (Palminteri et al., 2017)
`Cham-E1`	24	240	40	Pos	H, L	Expt 1, (Chambon et al., 2020)
`Cham-E3`	24	120	20	Pos	H, L	Expt 3, (Chambon et al., 2020)
`Lefe-E1`	50	96	24	Pos	sH, sL, C	Expt 1, (Lefebvre et al., 2017)
`Lefe-E2`	35	96	24	Mix	sH, sL, C	Expt 2, (Lefebvre et al., 2017)
`Lebr-E1a`	18	144	24	Pos, Neg	C	Expt 1, (Lebreton et al., 2019)
`Lebr-E2a`	18	144	24	Pos, Neg	C	Expt 2, (Lebreton et al., 2019)
`Lebr-E3`	48	360	30	Pos, Neg	C, R	Expt 3, (Lebreton et al., 2019)
Total `Expt(P)`	237	43,104				
`Palm-E2`	20	192	34	Pos	C, S, R	Expt 2, (Palminteri et al., 2017)
`Lebr-E1b`	18	144	24	Pos, Neg	C	Expt 1, (Lebreton et al., 2019)
`Lebr-E2b`	18	144	24	Pos, Neg	C	Expt 2, (Lebreton et al., 2019)
Total `Expt(C)`	56	9,024				

Model fits

We compare the previously proposed models (`VM`, `CD`, `DBM`) and our proposed models (`DV1`, `DV2`, `DVc`) on the datasets & experiments described above. For each model, the free parameters were estimated to minimize Negative Log Likelihood (NLL) under a softmax decision policy, using `matlab`'s `fmincon` function, in a manner similar to that described in Palminteri et al. (2017)[3]. The softmax policy computes the probability for a model selecting the same arm (c) as the participant as,

$$P(a=c) = 1/(1+\exp(-\beta\Delta q))$$

where $\Delta q = (q_c - q_u)$ for value-estimation models `VM`, `CD`, and `DBM`, and $(d_c - d_u) \equiv 2 \cdot d_c$ for decision-variable models `DV1`, `DV2`, and `DVc`. Log-likelihood of the data (sequence of choices of arm $\{a_i\}$ taken by a participant) given a model is computed as

$$\sum_{i=1}^{N} \log P(a_i = c_i)$$

Since all the models being compared have the same number of free parameters, i.e. 3 (two parameters in the update equations, and one parameter (β) in the softmax policy), any penalty for #parameters such as BIC/AIC would yield the exact same ordering of models in fit quality; we therefore do not report BIC scores. We present group comparison of models (Table 5), which is estimated using log likelihood of data computed from best fitted parameters.

We do not include a random prediction baseline, or the basic `RW` model in our results, as `RW` already performs much better than random, and `VM` & `CD` beat `RW` in `Expt(P)`, and `Expt(C)` respectively as shown in Palminteri et al. (2017)[4].

Partial Feedback Experiments

Model fits: As shown in Table 2, `DV2` shows consistently better fits to data across datasets, having clearly better log likelihood measure without resort to additional parameters. `DV1`, the algorithm structurally most related to `VM`, outperforms it in all datasets, and is at par with `DBM`, whereas `DV2`, which is inspired by `DVc` as an admixture of its update rules, provides a better fit than both `VM` and `DBM`. Table 5 shows a group analysis of the models with data pooled across all datasets, showing that exceedance probabilties (XP in the table; the probability that a given model best fits a majority of subjects) as well as estimated model frequencies clearly favor our models. Figure 2 shows model comparison across subjects pooled over all `Expt(P)` datasets. We see that `DV2` has better average NLL than `DBM` and `VM` for a significant majority of subjects.

Table 6 show mean values of fitted learning rates for our models. We find mean α_p considerably greater than mean α_n in `DV1` and α_2 greater than 0 in `DV2`, both reflecting a *tendency towards certainty* as described earlier. Figure 3 compares (α_p, α_n) for individual subjects pooled across all datasets, and shows that $\alpha_p > \alpha_n$, to a significant degree, for most subjects.

[3] We used code shared by the authors (Palminteri et al., 2017) in reproducing the performance of competing models `VM` and `CD` and are able to replicate the exact numbers reported in Palminteri et al. (2017), as a sanity check on the model fitting process.

[4] For reference, a rough calculation for random prediction model on `Palm-E1` gives NLL = -192*log(1/2) = 133, much worse than all models.

Table 2: **Expt(P) model fit**: NLL for best-fit parameters, averaged across participants. Our models consistently outperform previous models on a range of datasets. DV1 is significantly better than VM, while DV2 performs best overall.

Dataset	VM	DBM	DV1	DV2
Palm-E1	83.94	80.53	80.16	**79.23**
Cham-E1	92.10	91.01	91.22	**88.57**
Cham-E3	52.54	53.99	52.48	**51.74**
Lefe-E1	40.25	39.44	39.39	**38.90**
Lefe-E2	40.00	39.40	39.07	**38.39**
Lebr-E1a	55.94	55.98	55.46	**53.15**
Lebr-E2a	57.62	57.48	55.67	**54.13**
Lebr-E3	171.26	164.81	162.28	**156.79**

Generalization on holdout: Table 3 shows the evaluation of models on held-out blocks from each dataset, with parameters fitted using half of the blocks–this is a measure of *generalization on held-out data* commonly used in the machine learning literature. Not only is DV2 better on all datasets, but the percentage difference between the models is significantly magnified, suggesting that the difference between the models is in fact larger than apparent in Table 2.

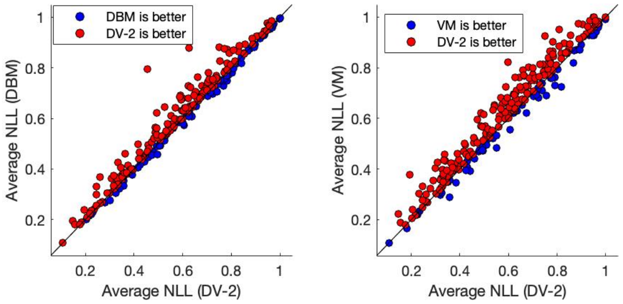

Figure 2: Comparison of Negative Log Likelihood (NLL) averaged over number of trials (N) $(1/N \cdot \log_2(P(data|model)))$ on our model DV2 and previous models (VM, DBM). Participants labelled red perform better on DV2.

Table 3: **Expt(P) generalization**: Train on $\lfloor \frac{n+1}{2} \rfloor$ blocks and test on the remaining blocks. Lefe-E1 and Lefe-E2 dropped from comparison since they only have one block.

Dataset	VM	DBM	DV1	DV2
Palm-E1	43.81	40.45	40.11	**39.07**
Cham-E1	33.98	32.09	32.90	**30.95**
Cham-E3	38.85	34.91	35.43	**34.53**
Lebr-E1a	19.57	18.75	19.43	**17.75**
Lebr-E2a	21.16	18.60	18.03	**17.71**
Lebr-E3	55.80	52.26	51.28	**49.61**

Complete Feedback Experiments

Model fits: Table 4 compares various models on the counterfactual experiment Expt(C). Our model DVc shows a better fit to data than DBM and VM using the same number of learned parameters. Interestingly, DBM shows a fairly poor fit to the data, significantly worse than CD. In factual experiments, as described in Zhou et al. (2020), the primary advantage of DBM was the decay of reward estimates for unchosen arms in the absence of feedback, resulting in a choice-perseveration-like behavior. However, in the counterfactual evidence scenario, unchosen outcomes are explicitly available, and need to be incorporated into the belief. We found that DBM-1, a version we implemented that ignores counterfactual evidence entirely, shows better fit than DBM, supporting the hypothesis that the gains from DBM come from a tendency towards choice perseveration that is removed in the face of explicit counterfactual evidence. Finally, from Table 5, we see that estimated model frequency (MF) for DVc compared to CD is 0.98, suggesting that for almost all participants DVc better fits their behavioral data than CD.

Table 4: **Expt(C) model fit**: DVc outperforms CD on all three datasets. DBM as noted earlier is unable to generalize to Expt(C), and performs poorly.

Dataset	DBM	DBM-1	CD	DVc
Palm-E2	82.50	74.91	69.60	**68.82**
Lebr-E1b	59.54	52.37	50.63	**49.31**
Lebr-E2b	55.49	45.26	43.99	**43.59**

Table 5: **Pairwise group comparison** of our models (DV1, DV2, DVc) with previous (VM, DBM, CD) on random effects Bayesian analysis; XP, exceedence probability; PP, posterior probability; MF, model frequency. Null Hypothesis (difference due to chance) was not rejected only in DV1 v/s DBM.

	DV1	DV2	DV1	DV2	DVc
	v/s VM	v/s VM	v/s	v/s	v/s CD
			DBM	DBM	
XP	1.00	1.00	1.00	1.00	1.00
PP	0.87	0.87	0.58	0.84	0.98
MF	0.87	0.87	0.58	0.84	0.97

Table 6: **Fitted Parameters:** Learning rates averaged across all participants (with standard mean error) for our models.

	DV1	DVc		DV2
α_p	0.42 ±0.02	0.42 ±0.02	α_1	0.24 ±0.01
α_n	0.13 ±0.01	0.11 ±0.01	α_2	0.17 ±0.01

Outcome-dependent choice reinforcement: As remarked in Eq 6, for the specific situations where both factual and counterfactual outcomes are equal, the choice of t controls whether current choice is reinforced ($t = 1$) or has reduction in confidence ($t = -1$). We found that $t = 1$ model was a better fit in terms of NLL for every single subject in our data pool.

Further, as discussed earlier, only these equal outcome trials causes DVc and CD to differ–if there were no trials with equal outcomes on both arms, DVc and CD would behave exactly the same, with the Q-values in CD having the relationship $q_c \equiv -q_u$. To investigate this difference further, we compared the two models on trials split into two bins: those that

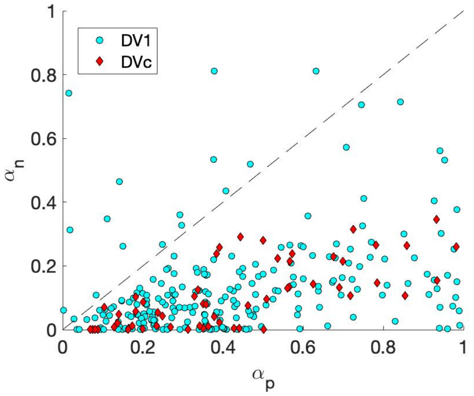

Figure 3: Fitted Learning rates (α_p vs α_n) of our models, DV1 and DVc on Expt(P) and Expt(C) respectively.

follow an equal-rewards trial, and those that follow an unequal rewards trial (Table 7). The relative NLL gains of DVc over CD are much larger on trials following equal-rewards trials, thereby validating our choice of update rule in Equation 7. This choice reinforcement is very different from unbiased Q-value estimation (RW model) in which decision confidence in current choice often reduces after same-outcome trials. It is also different from classically studied perseverative behavior which typically models autocorrelation in choice independent of outcomes.

The fitted parameters of DVc (Table 6) also find the mean of $\alpha_p - \alpha_n$ considerably greater than 0; this is also seen on a per-subject level in the scatterplot of Figure 3 where for most subjects, the fitted parameters show an asymmetry in value.

Table 7: NLL on Expt(C), split into trials following same (different) outcome trials.

		On trials following		
Dataset		any	$r_c = r_u$	$r_c \neq r_u$
Palm-E2	NLL CD	69.6	24.35	39.7
	NLL DVc	68.8	23.77	39.5
	Δ NLL	0.78	0.58	0.2
Lebr-E1b+	NLL CD	60.12	20.24	35.03
Lebr-E2b	NLL DVc	59.06	19.60	34.61
	ΔNLL	1.06	0.64	0.42

Discussion & Future Work

We presented a decision-variable account of human behavior in bandit tasks with factual and counterfactual feedback. Our model consistently outperforms previous proposals across a range of datasets and experimental conditions. Some key insights from our model and experiments are that decision variables (more generally, *action policies*, see e.g., Hayden & Niv (2021)) are a parsimonious representation in the task, both conceptually in terms of maintenance complexity and cost, and empirically in terms of fit to human behavior. We also showed strong evidence for *outcome-dependent perseveration*–the tendency to reinforce current choice in the absence of directional evidence. We discuss below some implications of our work and its connection to the broader literature on decision making & neuroscience.

Biased learning in humans: Asymmetric-update models such as VM and CD have been interpreted as *optimism bias* (Sharot et al., 2011), or *confirmation bias* (Palminteri et al., 2017) as they appear to reinforce positive or confirmatory interpretations of evidence. From an RL perspective, one could view these algorithms as learning biased *Q*-estimates; however, other work has argued that such estimates can be reward-maximizing, in comparison to unbiased estimates. For instance, Lefebvre et al. (2020) suggest that the CD update rules sharpen the gap between *Q*-values, and allow the learner to overcome "decision noise" inherent in the brain (e.g., instantiated as the temperature parameter of a softmax decision function). Similarly, Cazé & van der Meer (2013) suggest that "optimism" (i.e., higher learning rates for positive outcomes) increases reward in certain environments, whereas the reverse would be beneficial in others. Our work suggests a more straightforward interpretation of the apparent asymmetry as a tendency to *update decision variables towards certainty* (see also Hayden & Niv (2021)), as opposed to tracking biased individual values, providing robustness against stochasticity in reward distribution. In particular, this explains the finding (Palminteri et al., 2017) that humans to converge upon a specific choice when repeatedly offered two equally rewarding options (Palminteri et al., 2017), in a manner inconsistent with unbiased *Q*-estimation. A natural question for future work is the analytical and empirical evaluation of our models in terms of reward maximization, under a range of environmental conditions.

Representation of value: Taken at face value, our model suggests that only the decision variable, which measures a relative valuation between the two options, needs to be maintained. However, there is substantial prior work suggesting value representation and update in brain & behavior. In the 2-armed bandit task we study here, subjects can be probed, after the experiment, to estimate the actual likelihood of reward of each arm (Lebreton et al., 2019), and appear to have close-to-veridical representations of these probabilities. Further, experiments tracking neural representation of value suggest that individual arm values are estimated and tracked over trials–see e.g., Pischedda et al. (2020), who also, intriguingly, suggest that value representations switch from absolute to relative when comparing factual & counterfactual experiments. More broadly speaking, there is significant recent debate about whether the brain represent *value* or *action policies* (Hayden & Niv, 2021)–our proposal of a decision variable is directly in line with the hypothesis of policy tracking. We plan to explore in future work the computational and ecological value of a hybrid model that explicitly tracks decision variables in addition to maintaining individual (or, indeed, contextual) value estimates.

Relationship to decision-making in general: Perceptual decision-making models have largely focused on decision variables (Platt & Glimcher, 1999; Gold & Shadlen, 2007) which are primarily veridical representations of integrated sensory evidence over time. Interestingly, a growing body of work suggests that a closely related concept termed *decision confidence* (Pouget et al., 2016; Bang & Fleming, 2018)) may be explicitly represented in the brain, separate from probabilistic sensory evidence, for use in subsequent / downstream decision making. Other recent work (Yeon & Rahnev, 2020) suggests, for instance, that in perceptual decision-making with multiple alternatives, sensory evidence is not veridically represented, and instead a demonstrably suboptimal "decision-level" representation is used for choice behavior. Another interesting area of future investigation is whether the kind of decision-variable approach we propose here will apply to perceptual decision-making, and to multi-alternative choice & learning scenarios.

References

Akaishi, R., Umeda, K., Nagase, A., & Sakai, K. (2014). Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, *81*(1), 195–206.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *PNAS*, *115*(23), 6082–6087.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410.

Cazé, R. D., & van der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*(6), 711–719.

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067–1079.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574.

Hayden, B. Y., & Niv, Y. (2021). The case against economic values in the orbitofrontal cortex (or anywhere else in the brain). *PsyArXiv*.

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, *29*(31), 9861–9874.

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, *15*(4), 1-27.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9.

Lefebvre, G., Summerfield, C., & Bogacz, R. (2020). A normative account of confirmatory biases during reinforcement learning. *ArXiv*.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, *13*(8).

Pischedda, D., Palminteri, S., & Coricelli, G. (2020). The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *Journal of Neuroscience*, *40*(16), 3268–3277.

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, 64–99.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience*, *14*(11), 1475–1479.

Yeon, J., & Rahnev, D. (2020). The suboptimality of perceptual decision making with multiple alternatives. *Nature Communications*, *11*(1), 1–12.

Zhou, C. Y., Guo, D., & Yu, A. J. (2020). Devaluation of Unchosen Options: A Bayesian Account of the Provenance and Maintenance of Overly Optimistic Expectations. In *Cogsci 2020* (pp. 1682–1688).