Why so gloomy? A Bayesian explanation of human pessimism bias in the multi-armed bandit task

Anonymous Author(s)
Affiliation
Address
email

Abstract

How humans make repeated choices among options with imperfectly known reward outcomes is an important problem in psychology and neuroscience. This is often studied using multi-armed bandits, which are also frequently studied in machine learning. We present data from a human stationary bandit experiment, in which we vary the abundance and variability of reward availability (mean and variance of reward rate distributions). Surprisingly, subjects have significantly underestimated prior mean of reward rates – elicited at the end of a bandit game, when they are asked to estimate the reward rates of arms never chosen during the game. Previously, human learning in the bandit task was found to be well captured by a Bayesian ideal learning model, the Dynamic Belief Model (DBM), albeit under an incorrect generative assumption of the temporal structure – humans assume reward rates can change over time even when they are actually fixed. We find that the “pessimism bias” in the bandit task is well captured by the prior mean of DBM when fitted to human choices; but it is poorly captured by the prior mean of the Fixed Belief Model (FBM), an alternative Bayesian model that (correctly) assumes reward rates to be constant. This pessimism bias is also incompletely captured by a simple reinforcement learning model (RL) commonly used in neuroscience and psychology, in terms of fitted initial option values. While it seems highly sub-optimal, and thus mysterious, that humans have an underestimated prior reward expectation, our simulations show that an underestimated prior mean helps to maximize long-term gain, if the observer assumes volatility when reward rates are actually stable and uses a softmax decision policy instead of the optimal one (obtainable by dynamic programming). This raises the intriguing possibility that the brain underestimates reward rates to compensate for the incorrect non-stationarity assumption in the generative model and a suboptimal decision policy.

1 Introduction

Humans and animals frequently have to make choices among options with imperfectly known outcomes. This is often studied using the multi-armed bandit task [1, 2, 3], in which the subject repeatedly chooses among bandit arms with fixed but unknown reward probabilities. The observer learns how rewarding an arm is by choosing it and observing whether it produces a reward – thus each choice pits exploitation against exploration since it affects not only the immediate reward outcome but also the longer-term information gain. Previously, it has been shown that human learning in the bandit task is well captured by a Bayesian ideal learning model [4], the Dynamic Belief Model (DBM) [5], which assumes the reward distribution to undergo occasional, unsigned changes – this occurs despite the reward rates’ actually being fixed during a game. While this finding was consistent with the default (and incorrect) non-stationarity assumption humans make in a variety of other psychological tasks (2-alternative forced choice [6], inhibitory control [6, 7], multi-armed

It appears that this lake has high abundance, low variance of fish.

**Figure 1:** Behavioral experiment. (A) A screenshot of the experimental interface. The total number of attempts, total reward, and the cumulative reward of each option are shown to the subjects on each trial during the experiment. The four panels correspond to the four options (arms). A green circle represents a success (1 point), and a red circle represents a failure (0 point). (B) An example of fishing report in Beta(30, 15). 20 integers are shown, representing the total number of fish caught out of 10 attempts at 20 locations. Subjects are also told whether the environment is high or low abundance, and high or low variance.

bandit [4], and visual search [8]), it has remained nevertheless rather mysterious why humans would persist making this assumptions despite inconsistent environmental statistics.

In this work, we present and model human behavioral data in a variant of the bandit task, in which we vary reward abundance and variability in different environments. Previous multi-armed bandit study has typically been based on a neutral environment [3, 4, 10, 11], i.e. the average of the reward rates of the options is 0.5. Here, we manipulate the true generative prior of reward rates in four different environments: high/low abundance, high/low variance. This setting is different from some contextual bandit problems, in that provided information about the reward environment is not associated with any particular option; instead, the information is about all the options within the same environment. We aim to examine how humans adapt their decision-making to the different reward environments. Specifically, we focus on whether human subjects have veridical prior beliefs about reward rates. To gain greater computational insight into human learning and decision making, we compare the ability of DBM and a number of alternative models in their ability to capture the human data.

Specifically, we consider two Bayesian learning models, DBM and Fixed Belief Model (FBM) [5], coupled with a softmax decision policy. In contrast to DBM, FBM assumes that the environment is stationary, consistent with the actual experimental design of the task we study here. While FBM gives less and less weight to new observations, DBM continuously updates the posterior reward rate distribution by exponentially forgetting past observations, as well as continually injecting a fixed prior belief about the environment. Besides Bayesian learning, we also include a simple reinforcement learning rule (RL), the delta rule [12], which has been widely used in the neuroscience literature. In particular, dopamine has been suggested to encode the prediction error used in the RL model [13, 14]. DBM is related to RL in that the stability parameter in DBM also controls the exponential weights as the learning rate in RL does, but they are not mathematically equivalent. For the decision policy, we employ the softmax policy, which is popular in psychology and neuroscience, and has been frequently used to model human behavior in the bandit task [13, 16].

In the rest of the paper, we first describe the experiment and the data (Sec. 2), then present the model and the analyses (Sec. 3), and finally discuss the implications of our work and potential future directions (Sec. 4).

## 2 Experiment

We recruited 107 UCSD students to participate in a four-armed, binary bandit task (see Fig. 1A for the experimental interface), whereby the reward rates in four environments were identically and independently sampled from four Beta distributions: Beta(4, 2), Beta(2, 4), Beta(30, 15) and Beta(15, 30). The reward rates for the 50 games (15 trials each) were pre-sampled, and randomized for each
Figure 2: Human behavior and model comparison. (A) Reported reward rate estimates by human subjects (orange), and fitted prior mean of DBM (blue), FBM (purple), and RL (green). (+M, -V) denotes high mean (abundance), low variance, and so on. Error bars denote s.e.m. across subjects. The dotted lines indicate the true generative prior mean (0.67 for high abundance environments, 0.33 for low abundance environments). (B) Reward rates obtained by human subjects (blue), and expected reward rate of the optimal model (green) and a random choice policy (purple). (C) The per-trial likelihood of 10-fold cross validation. Error bars denote s.e.m. across subjects. The dotted line indicates chance level (0.25). (D) Fitted softmax $b$ and DBM $\gamma$ parameters in the four environments. Error bars: s.e.m. across subjects.

subject. The cover story was that is an ice fishing contest, where the four arms represent four fishing holes. Participants are informed that the different camps they fish from reside on four different lakes that vary in (a) overall abundance of fish, and (b) variability of fish abundance across locations. At the outset of each environment, we tell them the lake’s fishing conditions (high/low abundance, high/low variance) and provide samples from the distribution (a fishing report showing the number of fish caught out of 10 attempts at 20 random locations in the lake (Fig. 1B)). A subset of 32 subjects were required to report the reward rate of the never-chosen arms at the end of each game.

The reported reward rates are shown in Fig. 2A. Human subjects reported estimates of reward rate significantly lower than the true generative prior mean ($p < 10^{-3}$), except in low abundance and low variance environment ($p = 0.2973$). The average reported estimates across the four reward environments are not significantly different ($F(3, 91) = 1.78, p = 0.157$, see Fig. 2A), indicating that humans do not alter their prior belief about the reward rates even when provided with both explicit (verbal) and implicit (sampled) information about the reward statistics of the current environment. In spite of systematically underestimating expected rewards, our subjects appear to do well in the task. The actual total reward accrued by the subjects are only slightly lower than the optimal algorithm utilizing correct Bayesian inference and the dynamic-programming-derived decision policy (Fig. 2B); humans also perform significantly better than the chance level attained by a random policy ($p < .001$, see Fig. 2B), which is equal to the generative prior mean of the reward rates. Thus, subjects are actually experiencing sample reward rates that are higher than the generative prior mean (since they perform much better than the random policy); nevertheless, they significantly underestimate the mean reward rate.

3 Models

How do humans achieve relatively good performance with an “irrationally” low expectation of reward rates? We attempt to gain some insight into human learning and decision making by computational modeling of this phenomenon. Here, we consider three learning models, DBM, FBM, and RL, coupled with the decision policy of softmax rule. In the following, we first formally describe the models (Sec. 3.1), then compare their ability to explain the data (Sec. 3.2), and finally present simulation results to gain additional insight into what the results imply about human cognition (Sec. 3.4).

3.1 Model description

We denote the reward rate of arm $k$ at time $t$ as $\theta^t_k, k \in \{1, 2, 3, 4\}, 1 \leq t \leq 15$, and $\theta^t = [\theta^t_1, \theta^t_2, \theta^t_3, \theta^t_4]$. We denote the reward outcome at time $t$ as $R^t \in \{0, 1\}$, and $R^t = [R^t_1, R^t_2, \ldots, R^t_{15}]$. We denote the decision at time $t$ as $D^t, D^t \in \{1, 2, 3, 4\}$, and $D^t = [D^t_1, D^t_2, \ldots, D^t_{15}]$. 

3
**Dynamic belief model (DBM).** As with the actual experimental design, DBM assumes that the binary reward (1: win, 0: lose) distribution of the chosen arm is Bernoulli. Unlike the actual experimental design, where reward rates of a game are fixed, DBM assumes that the reward rate for each arm undergoes discrete, un-signalied changes independently with a per-trial probability of 1 − γ, 0 ≤ γ ≤ 1. The reward rate at time t remains the same with probability γ, and is re-sampled from the prior with probability 1 − γ. The observed data has the distribution, \( P(R_t = 1|T^t, D_t = k) = \theta_k^t \).

The generative dynamics of DBM is
\[
p(\theta_k^t = \theta|\theta_k^{t-1}) = \gamma(\delta(\theta_k^{t-1} - \theta) + (1 - \gamma)p^0(\theta)), \tag{1}
\]
where \( \delta(x) \) is the Dirac delta function, and \( p^0(\theta) \) is the assumed prior distribution.

The posterior reward rate distribution given the reward outcomes up to time t can be computed iteratively via Bayes’ rule as
\[
p(\theta_k^t | R^t, D^t) \propto p(R_t | \theta_k^t)p(\theta_k^t | R^{t-1}, D^{t-1}), \text{ if } D_t = k \tag{2}
\]
Only the posterior distribution of the chosen arm is updated with the new observation, while the posterior distribution of the other arms are the same as the predictive reward rate distribution, i.e.
\[
p(\theta_k^t | R^t, D^t) = p(\theta_k^t | R^{t-1}, D^{t-1}), \text{ if } D_t \neq k \text{ (see below)}. \]

The predictive reward rate distribution at time t given the reward outcomes up to time t − 1 is a weighted sum of the posterior probability and the prior probability:
\[
p(\theta_k^t = \theta | R^{t-1}, D^{t-1}) = \gamma p(\theta_k^{t-1} = \theta | R^{t-1}, D^{t-1}) + (1 - \gamma)p^0(\theta). \tag{3}
\]

The expected (mean predicted) reward rate of arm k at trial t is \( \hat{\theta}_k^t = \mathbb{E}[\theta_k^t | R^{t-1}, D^{t-1}] \). DBM can be well approximated by an exponential filter \( [5] \), thus \( \gamma \) is also related to the length of the integration window as well as the exponential decay rate.

**Fixed belief model (FBM).** FBM assumes that the statistical environment remains fixed throughout, e.g. the reward outcomes are Bernoulli samples generated from a fixed rate parameter \( \theta \). It can be viewed as a special case of DBM, with \( \gamma = 1 \).

**Reinforcement Learning (RL).** The update rule of a simple and commonly used reinforcement learning model is
\[
\hat{\theta}_k^t = \hat{\theta}_k^{t-1} + \epsilon (R_t - \hat{\theta}_k^{t-1}), \tag{4}
\]
with a initial value \( \hat{\theta}_k^0 = \theta^0 \), and 0 ≤ \( \epsilon \) ≤ 1. \( \epsilon \) is the learning parameter that control the coefficients of exponential decay of the previous observation sequence. In the multi-armed bandit task, only the chosen arm is updated, while other arms remain the same.

**Softmax decision policy.** We use a version of the softmax decision policy that assumes the choice probabilities among the options to be normalized polynomial functions of the estimated expected reward rates, with a parameter \( b \):
\[
p(D_t = k) = \frac{\left(\hat{\theta}_k^t\right)^b}{\sum_i K \left(\hat{\theta}_k^t\right)_i^b}, \tag{5}
\]
where \( K \) is the total number of options, \( b \geq 0 \). When \( b = 1 \), it is probability matching. When \( b = 0 \), the choice is at random, i.e. \( p(D_t = k) = 1/K \) for all options. When \( b \rightarrow \infty \), the most rewarding option is always chosen. By varying the \( b \) parameter, the softmax policy is able to capture more or less “noisy” choice behavior. However, we note that softmax ascribes unexplained variance in choice behavior entirely to “noise”, when subject may indeed employ a much more strategic policy whose learning and decision components are poorly captured by the model. Thus, smaller fitted \( b \) does not imply subjects are necessarily more noisy or care about rewards less; it may simply mean that the model is less good at capturing subjects’ internal processes.

**Optimal policy.** The multi-armed bandit problem can be viewed as Markov Decision Process, where the state variable is the posterior belief state after making each observation. The optimal solution to the problem considered here can be computed numerically via dynamic programming \( [4, 17] \), where the optimal learning model is FBM with the correct prior distribution. Previously, it has been shown that human behavior does not follow the optimal policy \( [4] \); nevertheless, it is a useful model to consider here in order to assess the performance of human subjects and the various other models in terms of maximal expected total reward.
We use per-trial likelihood as the objective function, calculated as \( \exp(\log \mathcal{L}/N) \), where \( \mathcal{L} \) is the maximum likelihood of the data, and \( N \) is the total data points. We fit prior weight \((\alpha + \beta, \text{related to precision})\) at the group level. We fit prior mean \((\alpha/(\alpha + \beta))\), DBM \( \gamma \), and softmax \( b \) parameters at the individual level, and separately for four reward environments. Since the pre-trial likelihood can also be interpreted as the trial-by-trial predictive accuracy (i.e. on average, how likely is it that the model will choose the same arm the subject chose), we will also refer to this measurement as predictive accuracy.

Fig. 2C shows the held-out per-trial likelihood for DBM, FBM, and RL, averaged across 10 runs of cross-validation. DBM achieves significantly higher per-trial likelihood than FBM \((p < .001)\) and RL \((p < .001)\) based on paired t-test, i.e., predicting human behavior better than the other two models. The predictive accuracy of three fitted models on the whole dataset are 0.4182 (DBM), 0.4038 (RL), and 0.3856 (FBM). DBM also achieves lower BIC and AIC values than the RL or FBM, in spite of incurring a penalty for the additional parameter. To summarize, DBM can predict and explain the human trial-by-trial choice behavior better than the other two models. This result further supports the previous finding \([4]\) that human assumes non-stationarity by default even in a stationary environment.

Next, we examine how well the three learning models can recover the underestimation effect observed in human subjects. The reported estimation is on the arm(s) that they never chose at the end of each game, which is their belief of the mean reward rate before any observation, i.e., mathematically equivalent to the prior mean (DBM & FBM) or initial values (RL). For simplicity, we will refer to them all as prior mean. Fig. 2A shows the average fitted parameter of the models. FBM recovers prior means that are well correlated with the true generative prior means \((r = .96, p < .005)\), and significantly different in the four environments \((F(3, 424) = 13.47, p < .001)\). The recovered prior means for RL are also significantly different in the four environments \((F(3, 424) = 4.21, p < .01)\). In contrast, the recovered prior means for DBM are not significantly different in the four environments \((F(3, 424) = 0.91, p = 0.435)\), just like human estimates (Fig. 2A). DBM also recovers prior means in low abundance and high variance environment slightly lower than in other environments, similar to human estimates. Therefore, compared to human data, DBM allows for better estimates of human internal prior beliefs than the other models.

Taking DBM as the best model for learning, we can then examine the other fitted learning and decision-making parameters. The fitted learning rate of RL is shown in Table 1, and the fitted softmax \( b \) for FBM and RL are shown in Table 2. A higher softmax \( b \) corresponds to a more myopic, less explorative, and less stochastic choice behavior. A lower DBM \( \gamma \) corresponds to a higher change rate and a shorter integration window of the exponential weights \([5]\). The prior weight of DBM is fitted to 6, equivalent to 6 pseudo-observations before the task; it is also the same as the true prior weight in the experimental design for high variance environments. Fig. 3D shows the fitted DBM \( \gamma \) and softmax \( b \) in four reward environments. In high abundance environments, softmax \( b \) is fitted to larger values, while DBM \( \gamma \) is fitted to lower values, than the low abundance environments \((p < .01)\). They do not vary significantly across low/high variance of environments \((p > .05)\). The fitted DBM \( \gamma \) values imply that subjects behave as if they believe the reward rates change on average approximately once every 3 trials in high-abundance environments, and once every 4 trials in low-abundance environment (mean change interval is \(1/(1-\gamma)\)). One possible explanation of the combination of lower DBM \( \gamma \) and higher softmax \( b \) is that, in a more volatile environment, the gain of information is not likely to outweigh the gain of immediate reward, since any information gain will not be useful for future choices after a change.

<table>
<thead>
<tr>
<th>Model</th>
<th>(+M, +V)</th>
<th>(+M, -V)</th>
<th>(-M, +V)</th>
<th>(-M, -V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL</td>
<td>0.35 (SD=0.02)</td>
<td>0.39 (SD=0.02)</td>
<td>0.22 (SD=0.02)</td>
<td>0.25 (SD=0.02)</td>
</tr>
</tbody>
</table>

### Table 1: Inferred RL Learning rates

3.2 Model comparison

Here, we compare the various models to human behavior, in order to identify the best (of those considered) formal description of the underlying psychological processes.

We first evaluate how well the three learning models fit human data. Since they have different numbers of parameters, we perform 10-fold cross-validation to avoid overfitting for comparison. We use per-trial likelihood as the objective function, calculated as \( \exp(\log \mathcal{L}/N) \), where \( \mathcal{L} \) is the maximum likelihood of the data, and \( N \) is the total data points. We fit prior weight \((\alpha + \beta, \text{related to precision})\) at the group level. We fit prior mean \((\alpha/(\alpha + \beta))\), DBM \( \gamma \), and softmax \( b \) parameters at the individual level, and separately for four reward environments. Since the pre-trial likelihood can also be interpreted as the trial-by-trial predictive accuracy (i.e. on average, how likely is it that the model will choose the same arm the subject chose), we will also refer to this measurement as predictive accuracy.

Fig. 2C shows the held-out per-trial likelihood for DBM, FBM, and RL, averaged across 10 runs of cross-validation. DBM achieves significantly higher per-trial likelihood than FBM \((p < .001)\) and RL \((p < .001)\) based on paired t-test, i.e., predicting human behavior better than the other two models. The predictive accuracy of three fitted models on the whole dataset are 0.4182 (DBM), 0.4038 (RL), and 0.3856 (FBM). DBM also achieves lower BIC and AIC values than the RL or FBM, in spite of incurring a penalty for the additional parameter. To summarize, DBM can predict and explain the human trial-by-trial choice behavior better than the other two models. This result further supports the previous finding \([4]\) that human assumes non-stationarity by default even in a stationary environment.

Next, we examine how well the three learning models can recover the underestimation effect observed in human subjects. The reported estimation is on the arm(s) that they never chose at the end of each game, which is their belief of the mean reward rate before any observation, i.e., mathematically equivalent to the prior mean (DBM & FBM) or initial values (RL). For simplicity, we will refer to them all as prior mean. Fig. 2A shows the average fitted parameter of the models. FBM recovers prior means that are well correlated with the true generative prior means \((r = .96, p < .005)\), and significantly different in the four environments \((F(3, 424) = 13.47, p < .001)\). The recovered prior means for RL are also significantly different in the four environments \((F(3, 424) = 4.21, p < .01)\). In contrast, the recovered prior means for DBM are not significantly different in the four environments \((F(3, 424) = 0.91, p = 0.435)\), just like human estimates (Fig. 2A). DBM also recovers prior means in low abundance and high variance environment slightly lower than in other environments, similar to human estimates. Therefore, compared to human data, DBM allows for better estimates of human internal prior beliefs than the other models.

Taking DBM as the best model for learning, we can then examine the other fitted learning and decision-making parameters. The fitted learning rate of RL is shown in Table 1, and the fitted softmax \( b \) for FBM and RL are shown in Table 2. A higher softmax \( b \) corresponds to a more myopic, less explorative, and less stochastic choice behavior. A lower DBM \( \gamma \) corresponds to a higher change rate and a shorter integration window of the exponential weights \([5]\). The prior weight of DBM is fitted to 6, equivalent to 6 pseudo-observations before the task; it is also the same as the true prior weight in the experimental design for high variance environments. Fig. 3D shows the fitted DBM \( \gamma \) and softmax \( b \) in four reward environments. In high abundance environments, softmax \( b \) is fitted to larger values, while DBM \( \gamma \) is fitted to lower values, than the low abundance environments \((p < .01)\). They do not vary significantly across low/high variance of environments \((p > .05)\). The fitted DBM \( \gamma \) values imply that subjects behave as if they believe the reward rates change on average approximately once every 3 trials in high-abundance environments, and once every 4 trials in low-abundance environment (mean change interval is \(1/(1-\gamma)\)). One possible explanation of the combination of lower DBM \( \gamma \) and higher softmax \( b \) is that, in a more volatile environment, the gain of information is not likely to outweigh the gain of immediate reward, since any information gain will not be useful for future choices after a change.
Table 2: Inferred softmax b

<table>
<thead>
<tr>
<th>Model</th>
<th>(+M, +V)</th>
<th>(+M, -V)</th>
<th>(-M, +V)</th>
<th>(-M, -V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL</td>
<td>6.94 (SD=0.54)</td>
<td>5.41 (SD=0.48)</td>
<td>5.72 (SD=0.51)</td>
<td>5.07 (SD=0.51)</td>
</tr>
<tr>
<td>FBM</td>
<td>12.28 (SD=0.58)</td>
<td>10.52 (SD=0.54)</td>
<td>6.25 (SD=0.42)</td>
<td>5.68 (SD=0.37)</td>
</tr>
</tbody>
</table>

3.3 Simulation results

To gain some insight into just how DBM behaves differently than FBM and RL, and thus what it implies about human psychological processes, we consider the empirical/simulated probability of the subject/model switching away from an arm after a “winning” arm suddenly produces a loss (Fig. 3A).

Since DBM assumes reward rates can change any time, a string of wins followed by a loss indicates a high probability of the arm switching to a low reward rate, especially in low reward rate environments where a switch is likely to result in a new reward rate that is low. On the other hand, since FBM assumes reward rate stability, it uses long-term statistics to estimate an arm’s reward rate; especially late in the game, many wins followed by a loss should still induce a high probability of sticking with the same arm. RL can adjust its reward estimate according to unexpected observations, but is much slower than DBM in doing so, since it has a constant learning rate and cannot increase it when there is high posterior probability of a recent change point (as when a string of wins is followed by a loss); it also persistently cannot encode prior information about overall reward abundance in the environment when a change occurs (e.g. in a low-abundance environment, the new reward rate after a change point is likely to be low). We would thus expect RL to behave intermediately between DBM and FBM. Fig. 3A shows that the simulated shift rate of the three models (probability of an algorithm to shift away from the previously chosen arm) late in the game (after at least 8 trials) exactly follow the pattern of behavior described above. Human subjects’ shift rates are closest to what DBM predicts, which is what we would expect from the fact that overall DBM has already been found to fit human data best.

3.4 Simulation results: understanding human reward underestimation

Finally, we try to understand why humans might exhibit a “pessimistic bias” in their reward rate expectation. Fig. 3A shows the simulated average earned reward per trial, of the various model as a function of the assumed prior mean (parameter of DBM, FBM, and RL are set to the average values of those best fit to human data); here we focus on the high variance environments, since model performance is quite insensitive to the assumed prior mean in low variance environments (not shown).

Firstly, human subjects’ actual average per-trial earned reward is very close to DBM prediction at the fitted prior mean; it is quite far from model predictions from FBM and RL at the best fitting prior mean values. This result provides additional evidence that DBM can predict and recover human performance better than the other two models.

More interestingly, while the optimal policy achieves the highest earned reward when the assumed prior is correct (as expected), FBM achieves its maximum at a prior mean much lower than the true generative mean. Given that FBM is the correct generative model, this implies that one way to compensate for using the suboptimal softmax policy, instead of the optimal (dynamic programming-derived) policy, is to somewhat underestimate the prior mean. In addition, DBM achieves maximal earned reward with an assumed prior mean even lower than FBM, implying that even more prior reward rate underestimation is needed to compensate for assuming environmental volatility (when the environment is actually stable).

4 Discussion

Our results show that humans underestimate the expected reward rates (a pessimism bias), and this underestimation is only recoverable by the prior mean of DBM. DBM is also found to be better than FBM or RL at predicting human behavior in terms of trial-by-trial choice and the reward rates earned. Our results also provides further evidence that humans underestimate the stability of the environment, i.e. assuming the environment to be non-stationary when the real setting is stationary. This default non-stationarity belief might be beneficial in long-run real world scenarios, where the environment is often changing [5]. It is worth noting that the best earned per-trial reward rates achievable by DBM
Figure 3: (A) Reward rates achieved in high variance environment by three models and optimal policy in low abundance environments and (B) high abundance environment. The four colors correspond to DBM (blue), FBM (purple), RL (green) and optimal policy (brown). DBM, FBM, RL are simulated with the average parameters that fitted to human data, except the prior mean is varied. The optimal policy is computed with different prior means and correct prior variance. The vertical dotted lines indicate the true generative prior mean. The markers with different colors represent the fitted prior mean of three models in the x-value, and human reward performance in the y-value. (C) Probability of shifting to other arms after a failure preceded by three consecutive successes on the same arm. The models are simulated with averaged fitted parameters. Error bar shows the s.e.m. across subjects.

and FBM are actually quite similar. In other words, as long as a softmax policy is being used, there is no disadvantage to incorrectly assuming environmental volatility as long as the assumed prior mean parameter is appropriately tuned. However, we see that subjects’ actual assumed prior mean (if DBM is correct) is not at the mode of the simulated performance curve, which underestimates prior mean even more than they do, but rather at a location that is a compromise between the mode and the true generative mean. This may reflect a tension to accurately internalize environmental statistics and assume a statistical prior that achieves better outcomes.

One of the limitations of this study is that the human reports might be unreliable, or biased by the experimental setting. For example, one might object to our “pessimism bias” interpretation of lower expected means for unchosen alternatives, by attributing it to the choice-supportive bias. The objector would say that subjects are reporting lower expected reward rates due to a tendency to retroactively view discarded options as more negative. However, this explanation fails to account for the higher pessimism bias (i.e., relatively lower expected reward rates compared to the true generative rates) observed in high abundance environments as compared to low abundance environments. In particular, for low abundance, low variance environments, we do not observe a significant underestimation of reward rates. This effect is predicted by the DBM, but it is not explained by the choice-supportive bias. Therefore, we believe the reward rate under-estimation effect is real.

Humans often have mis-specified beliefs, even though most theories predict optimal behavior when environmentally statistics are correctly internalized. For example, humans have often been found to overestimate their own abilities and underestimate probabilities of the negative events from the environment [18]. Our result might seems to contradict these earlier findings. However, having a lower prior expectation is not necessarily in conflict with humans’ optimism bias. Human subjects earn relatively high reward rate while reporting low expectation on the unseen option(s). It is possible that they are optimistic about their ability to succeed in a overall hostile environment (as they believe). Moreover, by having a low prior mean, one would overestimate the probability that continuing with the same option (which on average has the higher, true prior mean) is better than the other options (which have the lower, assumed prior mean).

Another interesting finding is that even though FBM is the correct generative model (in terms of stationarity), it still achieves the most earned reward by somewhat under-estimating the environmental reward rates when coupled with the softmax policy. This suggests that one way to compensate for using the suboptimal softmax policy, instead of the optimal (dynamic programming-derived) policy, is to underestimate the prior mean.

Another limitation of the current study is that we only include softmax as the decision policy. A previous study has found Knowledge Gradient (KG), which approximates the relative action values...
of exploiting and exploring to deterministically choose the best option on each trial, to be the best model among several decision policies \[1\] (not including softmax). However, in a later study, softmax was found to explain human data better than KG \[10\]. This is not entirely surprising, because even though KG is simpler than the optimal algorithm, it is still significantly more complicated than softmax. Moreover, even if humans use something like KG, they may not give the optimal weight to the exploratory knowledge gain on each trial, or use a deterministically best decision policy. We therefore also conceive a “soft-KG” policy, that adds a weighted knowledge gain to the immediate reward rates in the softmax policy. Based on 10-fold cross validation, the held-out likelihood of softmax and soft-KG, in explaining human data, are not significantly different \((p = 0.0693)\), and the likelihoods are on average 0.4248 and 0.4250. 44 subjects have higher, 44 subjects have lower and 19 subjects have equal held-out likelihood with Soft-KG model than softmax. For simplicity, we only included the softmax policy in the main text. Future studies with greater statistical power might be able to discriminate between softmax and KG, but this is beyond the scope and intention of this work.

While this study is primarily focused on modeling human behavior in the bandit task, it may have interesting implications for the study of bandit problems in machine learning as well. For example, our study suggests that human learning and decision making are sub-optimal in various ways: assuming an incorrect prior, assuming environmental statistics to be non-stationary when they are stable, and utilizing a simplistic decision policy (softmax instead of dynamic programming solution). However, our analyses also show that all these sub-optimalities may actually be combined to achieve much better than the expected sub-optimal performance, as they somehow compensate for each other. Given that these sub-optimal assumptions have certain computational advantages, e.g. softmax is computationally much simpler than optimal policy, DBM can handle a much broader range of problems than FBM, understanding how these algorithms fit together in humans may, in the future, yield better algorithms for machine learning applications as well.

References


