Computational mechanisms of human state-action-reward contingency learning under perceptual uncertainty

Dirk Ostwald[1,2] (dirk.ostwald@fu-berlin.de), Rasmus Bruckner[1], Hauke Heekeren[1]
[1]Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany
[2]Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

To successfully interact with an everchanging world imbued with uncertainties, humans often have to learn probabilistic state-action-reward contingencies. Reinforcement learning algorithms have been able to provide a mechanistic picture of the neurocomputational principles that govern such learning and decision processes. However, standard reinforcement learning algorithms assume that the environmental state is fully observable. Humans, on the other hand, often have to learn the expected reward of choice options under considerable perceptual uncertainty. In this project we investigate the computational principles that govern probabilistic state-action-reward learning under perceptual uncertainty. To this end, we designed an integrated perceptual and economic decision making learning task and acquired behavioural data from 52 human participants. To interpret the participants' choice data, we developed a set of artificial agents which describe a range of cognitive-computational strategies. These strategies range from Bayes-optimal exploitative decision making that takes perceptual uncertainty parametrically into account to fully random choice policies. Our behavioural modelling initiative favoured an agent model that suggests that human participants integrate their subjective perceptual uncertainty when learning probabilistic state-action-reward contingencies. They tend, however, to underestimate the degree they should do so from a normative Bayes-optimal perspective.

Keywords: Decisions; learning; perceptual uncertainty

Introduction

Humans often have to learn state-action-reward contingencies under considerable perceptual uncertainty. For example, when learning which varieties of wild berries are edible, perceptual uncertainty about the type of berry can significantly degrade the correct credit assignment between an experienced reward (such as an increase in blood glucose level), an action (choosing to eat a type of berry), and the environmental state (the type of berry). While standard reinforcement learning algorithms, such as Q-learning, have provided a mechanistic picture of state-action-reward contingency learning under full state-observability (Niv, 2009; Niv and Langdon, 2016), it is less clear how such algorithms can be adapted for cases imbued with perceptual uncertainty. Here, we developed a novel computational framework that uses neuroscience-inspired artificial agent models to provide Bayes-optimal solutions to this problem and that can be tested against human choice data. In the following, we first discuss the learning task developed to study state-action-reward contingency learning in humans and artificial agents ("The Gabor-bandit task"), then provide an overview about our computational framework ("Task and agent models"), and finally discuss the results of applying this framework to human behavioural data ("Experimental results").

The Gabor-bandit task

The Gabor-bandit (GB) task is a novel state-action-reward contingency learning task which combines aspects of perceptual and economic decision making. Each trial of this task comprises three stages (Figure 1). In the first stage, two Gabor patches that differ in their contrasts are simultaneously presented to the left and right of a central fixation cross and participants are asked to judge the relative Gabor patch contrasts. Participants report their perceptual judgement (higher/lower contrast on the left or right?) using the left and right cursor buttons of a computer keyboard. In the second stage of each trial, two clearly distinguishable, vertically aligned red and blue fractals are displayed and participants are asked to indicate the fractal which they think is associated with the higher reward probability given the relative locations of the high- and low-contrast Gabor patches. Participants indicate their fractal choice using the up and down cursor buttons on a computer keyboard. Finally, a reward of zero or one point is presented in the third stage of a task trial.

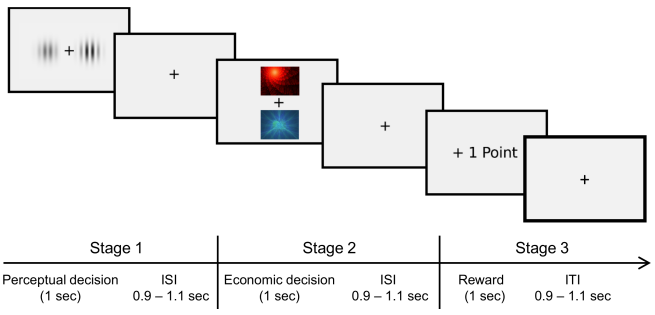

Figure 1: GB task trial structure. The GB task is a novel state-action-reward contingency learning paradigm that combines aspects of perceptual and economic decision making. Participants have to learn state-dependent (Stage 1) associations between actions (Stage 2) and experienced rewards (Stage 3) over a block of 25 trials. A single trial of the task is depicted in the figure.

The central feature of the GB task is the dependency of the fractal choice option reward probabilities on the relative display location of the high-contrast Gabor patch, which is randomly assigned on each trial with equal probabilities for left and right. For example, if on a given trial the high-contrast Gabor patch is displayed on the left side, then the blue fractal choice option is associated with a higher reward probability than the red fractal choice option. In contrast, if the high-contrast Gabor patch is displayed on the right side, then the blue fractal choice option is associated with a lower reward probability than the red fractal choice option.

Crucially, on each block of the GB task, which comprises 25 trials of the type shown in Figure 1, this stimulus-action-reward contingency is unknown to the participant and has to be relearned in order to maximize cumulative rewards. Furthermore, the Gabor patch contrast differences are manipulated on a trial-by-trial basis to induce variable amounts of perceptual uncertainty about the current location of the high-contrast Gabor patch. For each trial, the contrast difference between the two Gabor patches is drawn from a uniform distribution of subtle contrast differences extending over an interval of -0.08 (higher contrast on the left-hand side) to 0.08 (higher contrast on the right-hand side) Michelson contrast. In effect, over the course of each task block, participants face a credit-assignment problem regarding the contingency of the high-contrast Gabor patch location, the fractal choice options, and the received rewards.

Experimental procedures To assess human behaviour on the GB task, 54 participants were recruited from the local participant pool of Freie Universität Berlin and provided informed consent before partaking in the study. The data of two participants were excluded from all analyses due to a malfunction of the data acquisition set-up. The effective study sample thus consisted of 52 participants. All participants completed 12 blocks of 25 trials of the Gabor bandit task. On each task block, the state-action-reward contingency had to be relearned.

Task and agent models

Task model. To render the GB task amenable to computational modelling, we first formulated a mathematical model of the task. In our documentation of this model, we follow the conventions of machine learning literature on probabilistic models, i.e., we do not explicitly distinguish between probability distributions, probability density functions, or probability mass functions. We model a block of the GB task by the tuple

$$\left(T, S, C, R, A, p^{\phi}(s_t), p^{\kappa}(c_t|s_t), p^{a_t,\mu}(r_t|s_t)\right), \qquad (1)$$

where

- $T := 25$ denotes the number of trials per block, which are indexed as $t = 1, 2, ..., T$,
- $S := \{0, 1\}$ is the set of task states governing the action-reward contingencies of the task,

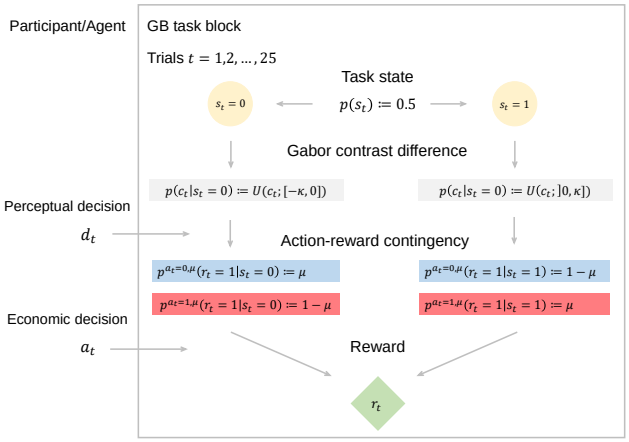

Figure 2: GB task model. Note that from the perspective of the task, the state-action-reward contingency parameter is a non-random entity and that the participants' and agents' perceptual decisions have no direct consequences for the observed rewards.

- $C := [-\kappa, \kappa]$ with $\kappa := 0.08$ is the set of Gabor patch contrast differences,
- $R := \{0, 1\}$ is the set of rewards,
- $A := \{0, 1\}$ is the set of economic decisions, where 0 and 1 represent the selection of the red and blue fractals, respectively,
- $p^{\phi}(s_t) := B(s_t; \phi)$ is the Bernoulli state distribution with parameter $\phi := 0.5$,
- $p^{\kappa}(c_t|s_t)$ is the state-conditional contrast difference uniform distribution defined by

$$p^{\kappa}(c_t|s_t) := U(c_t; [-\kappa, 0])^{1-s_t} U(c_t;]0, \kappa])^{s_t}, \qquad (2)$$

- and $p^{a_t,\mu}(r_t|s_t)$ is the action- and parameter-dependent and state-conditional reward distribution which represents the true state-action-reward contingencies of the GB task. This distribution is defined as

$$p^{a_t,\mu}(r_t|s_t) := \left(B(r_t; \mu)^{1-s_t} B(r_t; 1-\mu)^{s_t}\right)^{1-a_t} \left(B(r_t; 1-\mu)^{1-s_t} B(r_t; \mu)^{s_t}\right)^{a_t} \qquad (3)$$

with contingency parameter $\mu := 0.8$.

As visualized in Figure 2, on each trial, the task (1) samples a state s_t according to $p(s_t)$, (2) records a perceptual decision d_t, (3) samples and displays a contrast difference c_t according to $p^{\kappa}(c_t|s_t)$, (4) records an economic decision a_t, and (5) samples and displays a reward according to $p^{\mu,a_t}(r_t|s_t)$.

Agent models. To formalize the putative cognitive processes of human participants interacting with the GB task, we next developed a set of four neuroscience-inspired agent models. These agents are of similar overall structure, but differ in their precise state-inference and sequential-learning algorithms. All agent models are represented by a tuple

$$\Big(T, M, S, C, D, A, R, p(\mu), p^{\phi}(s_t), p^{\kappa}(c_t|s_t), p^{\sigma^2}(o_t|c_t), p^{a_t}(r_t|s_t,\mu)\Big), \tag{4}$$

where

- $T, S, C, A, R, p^{\phi}(s_t), p^{\kappa}(c_t|s_t), p^{a_t}(r_t|s_t,\mu)$ are as for the task model, with the difference that μ assumes the status of a random variable,
- $M := [0,1]$ is the outcome space of this random variable, which represents the agent's uncertainty about the state-action-reward contingency parameter on a given task block,
- $O \in \mathbb{R}$ is a set of internal agent observations o_t that are assumed to result from the external Gabor patch contrast difference c_t under additive perceptual noise,
- $D := \{0,1\}$ is a set of perceptual decisions, where 0 denotes the perceptual decision indicating that the contrast on the left is larger than on the right, and 1 the opposite,
- $p(\mu)$ is the agent's task block-specific initial uncertainty about μ, which corresponds to a uniform distribution over M, and
- $p^{\sigma^2}(o_t|c_t)$ is the agent's observation likelihood, which we defined by the conditional normal distribution

$$p^{\sigma^2}(o_t|c_t) := N(o_t; c_t, \sigma^2), \tag{5}$$

where $\sigma^2 > 0$ is a perceptual sensitivity parameter.

Crucially, the probability distributions of the agent tuple induce a joint probability distribution, which allows for analytically evaluating an agents' observation-conditional state distribution (modelling human perception) and defining a recursive scheme for the sequential updating of the agents' uncertainty representation about the state-action-reward contingency parameter (modelling human learning and memory). Furthermore, this joint distribution allows for defining perceptual and economic decision policies (modelling human decision making). We next elaborate on these aspects for the four agent models that constitute our model space.

Agent A1 is a Bayes-normative exploitative model: on each trial of a task block, agent A1 makes the economic decision that maximizes its expected reward given its current knowledge about the state-action-reward contingency parameter. To achieve a Bayes-optimal estimation of this parameter, agent A1 sequentially updates its uncertainty about it, by first inferring on its observation-conditional state distribution (belief state) according to

$$\begin{aligned}\pi_s &:= \int p^{\sigma^2,\kappa}(s_t, c_t|o_t)\, dc_t \\ &= \frac{\big(\Phi(0;o_t,\sigma^2) - \Phi(-\kappa;o_t,\sigma^2)\big)^{1-s_t}\,\big(\Phi(\kappa;o_t,\sigma^2) - \Phi(0;o_t,\sigma^2)\big)^{s_t}}{\Phi(-\kappa;o_t,\sigma^2) - \Phi(\kappa;o_t,\sigma^2)},\end{aligned} \tag{6}$$

where Φ denotes the Gaussian cumulative density function and makes its perceptual decisions in accordance with its belief state on each task trial, i.e., it sets

$$p(d_t = 0|o_t) := \pi_0 \text{ and } p(d_t = 1|o_t) := \pi_1. \tag{7}$$

The agent then integrates this trial-wise belief state in a sequential Bayesian updating scheme for its uncertainty about μ. This scheme shares some similarities with standard sequential Beta-Bernoulli Bayesian learning, but instead of a Beta distribution employs probability density functions that are defined in terms of increasing order polynomials in μ. These polynomials are of the general form

$$p(\mu|r_{1:t}, a_{1:t}) := \sum_{k=0}^{t} \rho_{t,k}\mu^{t-k}, \tag{8}$$

where the polynomial coefficients $\rho_{t,k}, k = 0,...,t$ can be evaluated, for each $t = 1,...,T$, in closed form based on the coefficients $\rho_{t-1,k}, k = 0,1,...,t-1$ and the agent's trial-specific belief state (see Djuric and Huang (2000) for similar work). Finally, agent A1 makes its economic decision by choosing that action a_t^* for which

$$a_t^* = \arg\max_A \mathbb{E}_{p^{a_t}(r_t|o_{1:t}, r_{1:t-1})}(r_t), \tag{9}$$

i.e., it chooses that fractal which promises the highest expected reward on a given trial upon integrating over its current uncertainty about μ.

Agent A2 is a behavioural model, which is of similar nature as agent A1, but instead of using the normative belief state (6) for learning and economic decision making defines its trial-by-trial belief state categorically according to a participant's perceptual decision. That is, if d_t denotes the perceptual decision of a participant on a given trial, agent A2 encodes the belief state

$$\pi_0 := \begin{cases} 0 & (d_t = 1) \\ 1 & (d_t = 0) \end{cases}, \quad \pi_1 := \begin{cases} 1 & (d_t = 1) \\ 0 & (d_t = 0) \end{cases} \tag{10}$$

Intuitively, agent A2 thus foregoes a parametric encoding of Bayes-normative state uncertainty when evaluating its state-action-reward contingency and economic choice options based on (8) and (9), respectively.

Agent A3 is a behavioural model that results from a mixture of the A1 and A2 learning and decision making architectures. Specifically, for its trial-wise uncertainty about μ, agent A3 forms the convex combination

$$p_{A_3}(\mu|r_{1:t}, a_{1:t}) := \lambda p_{A_1}(\mu|r_{1:t}, a_{1:t}) + (1-\lambda) p_{A_2}(\mu|r_{1:t}, a_{1:t}), \tag{11}$$

where $\lambda \in [0,1]$, and, similarly, for its economic decision evaluates

$$\mathbb{E}_{p^{a_t}_{A3}(r_t|o_{1:t},r_{1:t-1})}(r_t) := \lambda \mathbb{E}_{p^{a_t}_{A1}(r_t|o_{1:t},r_{1:t-1})}(r_t) + (1-\lambda)\mathbb{E}_{p^{a_t}_{A2}(r_t|o_{1:t},r_{1:t-1})}(r_t). \quad (12)$$

Agent A0, finally, is a random-choice control model that eschews any task state inference and state-action-reward contingency parameter representations. It merely makes perceptual and economic decisions uniformly at random.

Parameter estimation and model evaluation To evaluate agent models A0 to A3 in light of the experimentally acquired human behavioural data, we used a combination of maximum-likelihood estimation and Bayesian-information criterion (BIC)-based model evaluation. Specifically, for each agent model and participant, we first integrated over the agent's internal observation space, then maximized the log probability of the participant's choices using a softmax choice rule, computed the participant- and model-specific BIC scores, and finally evaluated the model-specific group BIC scores using a random-effects Bayesian model selection procedure (Rigoux et al., 2014).

Experimental results

Figure 3 depicts the results of our behavioural modelling initiative. As a measure of the agent models' face validity, we first compared the average cumulative rewards achieved by human participants and simulated task-agent interactions (Figure 3A). To this end, a series of task-agent interactions were simulated under identical conditions as in the experimental study with human participants (e.g., number of GB task blocks, trials, observed stimulus and reward frequencies, estimated perceptual sensitivity and softmax function parameters). The average cumulative reward traces of Figure 3 indicate that the human participants were able to learn the state-action-reward contingency, but, given the high perceptual state uncertainties on a subset of trials, did not achieve the theoretically possible expected reward of 0.8. Moreover, the human participants performed slightly worse than the Bayes-optimal exploitative agent model A1 and slightly better than the heuristic behavioural agent model A2, such that the convex combination of agents A1 and A2 as implemented in A3 achieves the highest performance correspondence. Naturally, the random choice model A0 performs at chance levels.

To formally compare the agent models in light of the participants' choice data, we evaluated the cumulative BIC scores over participants for each agent model (Σ BIC) (Figure 3B). These indicate that model A3 indeed explains the behavioural data best, followed by agent models A1 and A2. Moreover, assessing model plausibility using a random-effects Bayesian model selection procedure confirms this result by allocating a protected model exceedance probability (pEP) of larger than 0.99 to model A3 (Figure 3C).

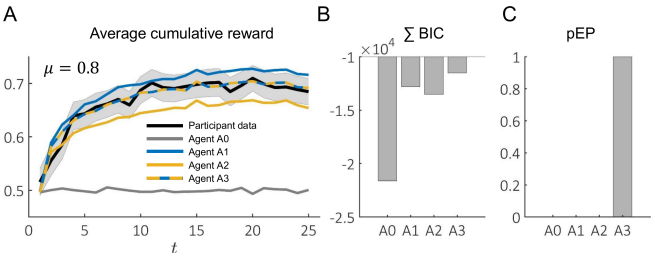

Figure 3: Experimental results. (A). Average cumulative rewards of human participants and agent model simulations. The grey error bars depict the SEM of the human participant data. (B). Cumulative BIC scores for each agent model over participants. (C). Model exceedance probabilities.

Conclusion

In summary, in the current project we investigate the computational mechanisms of state-action-reward contingency learning under perceptual uncertainty. We developed a novel experimental and computational framework that allows for formally evaluating different mechanistic theories on how such learning is algorithmically realized. This framework goes beyond standard reinforcement learning algorithms by explicitly accounting for various uncertainties in a probabilistic manner. Experimentally, we found evidence that human participants may capitalize on the cognitive-computational mechanisms proposed here, and exhibit a mixture of Bayes-optimal and heuristic decision making and learning.

References

Djuric, P. M. and Huang, Y. (2000). Estimation of a bernoulli parameter p from imperfect trials. *IEEE Signal Processing Letters*, 7(6):160–163.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154.

Niv, Y. and Langdon, A. (2016). Reinforcement learning with marr. *Current opinion in behavioral sciences*, 11:67–73.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies revisited. *Neuroimage*, 84:971–985.