

# A Bayesian formulation of behavioral control

Quentin JM Huys<sup>1,2</sup> and Peter Dayan<sup>1</sup>

<sup>1</sup>Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, UK

<sup>2</sup>Center for Theoretical Neuroscience, Columbia University, 1051 Riverside Drive, New York 10025, NY, USA  
qhuys@cantab.net, dayan@gatsby.ucl.ac.uk

September 10, 2008

Supplementary Online Material

# Mathematical formulations of behavioral control

We here give the mathematical details of the various models of control: outcome entropy; fraction of controllably achievable outcomes and fraction of controllably achievable reinforcement.

Briefly, the general setup is the following: Environments are assumed to be characterised by particular levels of control, i.e. the likelihood of observations is parametrised according to some suitably defined control parameter. Organisms collect observations in one (or a few) training environments, and based on this infer a posterior distribution over the setting of the control parameter in the training environments. Organisms are then transferred to a test environment and exposed to a limited number of observations. Organisms combine their prior expectations about the level of control in the test environment (derived in a suitable manner from the posterior distributions over control in the training environments) with the likelihood of the observations in the test environment and arrive at a predictive distribution for future observations in the test environment. Actions in the test environment are chosen according to the predictive probabilities of outcomes.

## 1 Control as conditional entropy / outcome set size

The first and most basic notion of control is that of the entropy of the probability distribution over outcomes, conditioned on an individual action (Maier and Seligman, 1976; Overmier et al., 1980; Gibbon et al., 1974). This is closely related to the effect of outcome set sizes of independent actions, i.e. the number of outcomes that are potentially observable for any one action. The outcome set size is related to the conditional entropy, but is analytically much more convenient. We follow the work of Friedman and Singer (1999); Dearden et al. (1998, 1999) closely. The setup is thus the following: given a number of action-outcome observations, and a prior belief about how many *different* observations are likely to be observed, what is the optimal action choice? The optimal action choice will be derived from the predictions about which outcomes are likely for the action.

Let us first consider a single action, with  $L$  possible outcomes. Let  $X$  be an unordered subset of these outcomes and  $|X|$  be the cardinality of that set, i.e. the number of different elements in the set, e.g. for the subset  $X = \{1, 2, L\}$  (for  $L > 2$ ),  $|X| = 3$ . There are  $\binom{L}{|X|} = L!/(|X|!(L - |X|)!)$  such sets of a given size for a total number of  $L$  outcomes. We will now put a prior distribution  $p(|X|)$  on the size of the outcome set, i.e. on the number of different outcomes expected for a particular action, and assume that all sets of the same cardinality have equal probability. This leads to a prior on sets

$$p(X) = \binom{L}{|X|}^{-1} p(|X|) \quad (1)$$

Let us furthermore parametrise the prior on set size  $p(|X|)$  in equation 1 as a truncated geometric distribution with parameter  $\zeta$ :

$$p(|X||\zeta) = \begin{cases} 1/L & \text{if } \zeta = 1 \\ \zeta^{|X|-1} \frac{1-\zeta}{1-\zeta^L} & \text{else} \end{cases} \quad (2)$$

where as  $\zeta \rightarrow -\infty$  only set size 1 is allowed, and as  $\zeta \rightarrow \infty$  all but set size  $L$  is prohibited. Thus, the parameter  $\zeta$  determines the set size, and is our parametrisation of control for this subsection.

To illustrate the pure effect of a prior on outcome size, we need to integrate out the effect of the actual probability distribution over that set. Let  $\mathbf{c}$  denote the outcome probability vector of an action, i.e. the probability of observing outcome  $i$  is  $c_i$ , and the likelihood of observing outcome  $i$   $n_i$  times is a multinomial

$$p(\mathbf{n}|\mathbf{c}) = \frac{(\sum_i n_i)!}{\prod_i n_i!} \prod_i c_i^{n_i} \quad (3)$$

It is now possible to put a Dirichlet prior, parametrised by the outcome set  $X$ , on the multinomial vector of outcome probabilities  $\mathbf{c}$ :

$$p(\mathbf{c}|X, \alpha) = \frac{\Gamma(|X|\alpha)}{\prod_{i \in X} \Gamma(\alpha)} \prod_{i \in X} c_i^{\alpha-1} \quad (4)$$

$$(5)$$

which puts mass on vectors  $\mathbf{c}$  with  $|X|$  nonzero elements. We let  $\alpha$  be relatively large to ensure that all outcomes  $o \in X$  have a large probability of actually generating data (putting most probability mass on vectors  $\mathbf{c}$  such that  $c_i \approx c_j \forall i, j \in X$ ). The predictive probability that the outcome at the next action  $D + 1$ , given that  $D$  outcomes have already been observed, is a standard multinomial as a Dirichlet prior is conjugate to the multinomial:

$$p(n_{D+1} = j|\mathbf{n}, X, \alpha) = \begin{cases} \frac{\alpha + n_j}{|X|\alpha + N} & \text{if } j \in X \\ 0 & \text{else} \end{cases} \quad (6)$$

Note importantly, that this only applies to outcomes *within* the set  $X$  on which we condition. Given our prior over sets in equation 1, this allows us to derive the probability of observing any outcome by averaging over set sizes. Note however, that sets that do not contain the set of previously observed outcomes (call this set  $Y$ ) have zero likelihood and thus do not contribute to the predictive distribution:

$$p(n_{D+1} = j|\mathbf{n}, \alpha) = \sum_{X \supseteq \{Y, j\}} p(n_{D+1}|\mathbf{n}, X) p(X|\mathbf{n}, \alpha) \quad (7)$$

$$\begin{aligned} p(\mathbf{n}|X, \alpha) &= \int d\mathbf{c} p(\mathbf{n}|\mathbf{c}) p(\mathbf{c}|X, \alpha) \\ &= \frac{N!}{\prod_{i \in X} n_i!} \frac{\Gamma(|X|\alpha)}{\Gamma(|X|\alpha + N)} \prod_{i \in X} \frac{\Gamma(\alpha + n_i)}{\Gamma(\alpha)} \end{aligned} \quad (8)$$

$$p(X|\mathbf{n}, \alpha) = \frac{p(\mathbf{n}|X, \alpha) p(X)}{\sum_X p(\mathbf{n}|X, \alpha) p(X)} = \frac{B(X)}{\sum_{X \supseteq Y} B(X)} \quad (9)$$

$$\begin{aligned} B(X) &= \frac{\Gamma(|X|\alpha)}{\Gamma(|X|\alpha + N)} \prod_{i \in X} \frac{\Gamma(\alpha + n_i)}{\Gamma(\alpha)} \binom{L}{|X|}^{-1} p(|X|) \\ \Rightarrow p(n_{D+1} = j|\mathbf{n}, \alpha) &= \frac{\sum_{X \supseteq \{Y, j\}} \frac{\alpha + n_j}{|X|\alpha + N} B(X)}{\sum_{X \supseteq Y} B(X)} \end{aligned} \quad (10)$$

Equation 8 is a standard Dirichlet integral, equation 9 is Bayes theorem and equation 10 is the predictive distribution given previous observations. As we will here mainly be dealing with problems in which  $L$  is small, say around 6, we can evaluate these sums explicitly. For bigger problems, it is possible to approximate the sumes by sampling from the sets  $X$  with nonzero likelihood.

To ensure that this parametrization does indeed affect environments in a recognisable manner, we perform inference of  $\zeta$  based on a set of observations, drawn from independent actions, via Expectation Maximisation (MacKay, 2003). Figure 1 shows the result of this. For large  $\zeta$ , accurate

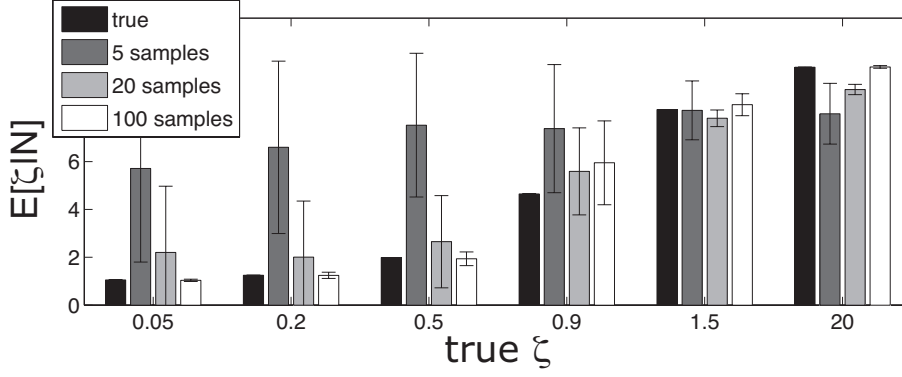


Figure 1: Inferring  $\zeta$  from observations on  $L = 20$  independent actions with  $L = 10$  possible outcomes each, averaging over  $c$  with  $\alpha = 20$ .

inference is possible even when very few samples have been observed, but at low  $\zeta$  the inference is much noisier. At low sample numbers, the likelihood appears to contain two modes, one at low, and one at high  $\zeta$ , to account for the few cases in which 2 or more outcomes are observed for a particular action. The second mode however disappears rapidly with added sampling, or is eliminated by adding in even a weak prior (data not shown).

## 2 Multiple actions with independent outcomes

The second notion of control incorporates both the outcome entropy above, but additionally measures the extent to which all outcomes in an environment are controllably achievable. That is, we here define control in a manner that takes into account whether different actions achieve outcomes reliably, where they achieve different outcomes, and whether these cover the range of outcomes possible in an environment.

**Individual action outcome distribution:** For mathematical convenience we will only deal with a simplified set of outcome distributions. We parametrise the conditional distribution of one action very simply as a mixture of a uniform distribution and a Kronecker delta, i.e. we write the probability of a set of observations  $\mathbf{n}$ ,  $n_i$  being the number of times outcome  $i$  has been observed following the choice of action  $a$

$$P(\mathbf{n}|c, \mathbf{m}) = \frac{(\sum_i n_i)!}{\prod_i n_i!} c^{\mathbf{n}^T \mathbf{m}} \bar{c}^{\mathbf{n}^T (1-\mathbf{m})} \quad \bar{c} = \left( \frac{1-c}{L-1} \right) \quad (11)$$

where  $\mathbf{m} = [0 \dots 0 1 0 \dots 0]^T$  denotes one of the outcomes as the “controllably attainable” one for that particular action. The *scalar* variable  $c$  (not to be confused with the outcome probability vector  $\mathbf{c}$  in the previous subsection) determines the mixing distributions. We will say that it regulates the degree to which the outcome is “controllably achievable”. The outcomes not designated by  $\mathbf{m}$  all have equal probability.  $\mathbf{n}$  is the vector of outcome counts.  $L$  is the number of potential outcomes, and for simplicity we assume that the number of available actions is equally  $L$  (though it is straightforward to relax this).

For  $c \rightarrow 1$ , only one outcome (the one for which  $m_i = 1$  is true) is observed, whereas as  $c \rightarrow 1/L$ ,

any outcome might be observed. The outcome entropy for that action

$$\mathcal{H} = - \sum_i p_i \log p_i = -c \log(c) - (1-c) \log \frac{1-c}{L-1} \quad (12)$$

is a strictly monotonically decreasing function of  $c$  for  $L > 2$ . Thus,  $c$  captures the original notion of control as outcome entropy.

**Multiple actions:** For a set of independent actions, we can write the likelihood of observations (assuming independent observations for different actions):

$$P(\mathbf{N}|c, \mathbf{M}) \propto \prod_a c^{(\mathbf{n}^a)^\top \mathbf{m}^a} \bar{c}^{(\mathbf{n}^a)^\top (1-\mathbf{m}^a)} \propto \prod_{ij} C_{ij}^{N_{ij}} \quad (13)$$

where we have assigned the  $a^{\text{th}}$  column vector  $\mathbf{m}^a$  of the matrix  $\mathbf{M}$  to action  $a$  and the matrix  $C$  is defined below.  $\mathbf{N}$  is a matrix consisting of the column vector observations for each of the actions. The meaning of  $\mathbf{M}$  is important: each column stands for one action, each row for one outcome. A unity entry in a column designates that outcome as the main “controllably achievable” outcome for that action. A goal-directed actor would chose that action in order to maximise the chances of obtaining that outcome. The variable  $c$  determines the probability of actually observing the designated outcome as opposed to any other one.

The second notion of control now becomes apparent, in the relationship between the columns of  $\mathbf{M}$ , i.e. between the controllably achievable outcomes of different actions. Consider the matrices  $\mathbf{M}$  and their associated matrices  $\mathbf{C}$ , whose entry denotes the probability of outcome  $o = i$  given action  $a = j$  was chosen  $C_{ij} = p(o = i|a = j)$

$$\begin{aligned} \mathbf{M}_0 &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & \mathbf{C}_0 &= \begin{bmatrix} \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} \\ c & c & c & c \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} \end{bmatrix} \\ \mathbf{M}_1 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{C}_1 &= \begin{bmatrix} c & \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & c & \frac{1-c}{L-1} & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & c & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & \frac{1-c}{L-1} & c \end{bmatrix} \\ \mathbf{M}_2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{C}_2 &= \begin{bmatrix} c & \frac{1-c}{L-1} & 1/L & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & c & 1/L & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & 1/L & \frac{1-c}{L-1} \\ \frac{1-c}{L-1} & \frac{1-c}{L-1} & 1/L & c \end{bmatrix} \\ \mathbf{M}_3 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} & \mathbf{C}_3 &= \begin{bmatrix} c/2 & c/2 & 1/L & \frac{1-c}{L-1} \\ \frac{1-c}{L-2} & c/2 & 1/L & \frac{1-c}{L-1} \\ c/2 & \frac{1-c}{L-2} & 1/L & \frac{1-c}{L-1} \\ \frac{1-c}{L-2} & \frac{1-c}{L-2} & 1/L & c \end{bmatrix} \end{aligned} \quad (14)$$

These have very different implications. For large  $c$ , these matrices now exemplify various dimensions along which a putative control variable may change.

- $\mathbf{M}_0$ : outcome 2 is attainable, but it is also the only one attainable. For  $c \rightarrow 1$ , all actions deterministically lead to outcome 2.
- $\mathbf{M}_1$ : one action available for each of the outcomes. As  $c \rightarrow 1$ , all actions can deterministically attain their outcomes. In this case, all outcomes would be controllably achievable.

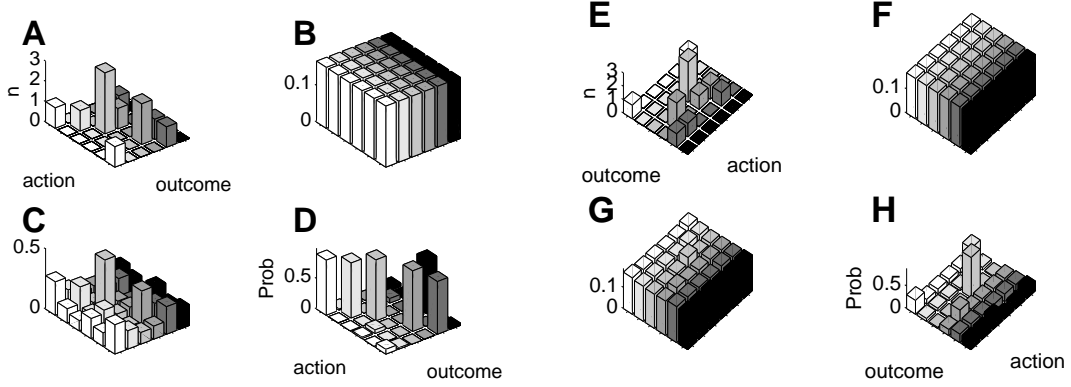


Figure 2: Prior on achievable fraction of outcomes. Effect of entropy ( $c$ ) and assumed achievable fraction on predictive distributions. Panels **B-D** show mean of posterior distribution over matrices  $\mathbf{M}$  given observations  $\mathbf{N}$  in panel **A**, and given different levels of entropy as determined by  $c$ . In these panels, the prior assumed that all outcomes were achievable. Panel **B** shows the posterior mean for  $c = 1/L = 0.16$ , i.e. no control at all due to the outcome entropy; **C** for  $c = 0.25$  and **D** for  $c = 0.9$ . The posterior mean becomes more dominated by a single matrix satisfying the constraints in equation 15 the higher  $c$ . No outcomes have as yet been observed for action 6, but at higher levels of control, its outcome is still inferred with high certainty due to the constraint that all actions lead to a different outcome. The four panels on the right show the effect of relaxing the prior by decreasing  $|\mathbf{M}|$ . **E** shows the data, which is the same as in **A** but rotated for clarity. **F** shows that for  $c = 1/L$  the same predictive distribution is inferred. In comparison to **C**, **G** shows that a small value of  $c$  now leads to much more uncertainty, as outcomes from different actions can no more be used to constrain each other “by elimination”. **H** shows that for  $c = 0.9$ , low-entropy posteriors are only seen for those actions where outcomes have been observed. Note that for action 6, the posterior mean is flat.

- $\mathbf{M}_2$ : actions available for a fraction (here 3/4) of the outcomes.
- $\mathbf{M}_3$ : actions lead to more than a unique outcome, even for high  $c$  this does not lead to full control. We will not consider this setting any further.

as  $c \rightarrow 1/L$ , the observations these matrices generate the same, flat, uncontrollable outcomes. Later, we will also consider the notion that control is “about” some particularly reinforcing outcome.

When more than a single action is considered, we thus need to take the relationship between actions into consideration as illustrated in equation 14. We return to the simple case of equation 13, constraining the matrix  $\mathbf{M}$  to have one unit entry in each column and row. If there are  $L$  actions and  $L$  outcomes, there are  $L!$  such matrices. For small  $L$ , the relevant integrals can be evaluated explicitly. We write the likelihood of observations as in equation 13, and add a prior

$$p(\mathbf{M}) = \frac{1}{L!} \left( \prod_j \delta(1 - \sum_i M_{ij}) \right) \left( \prod_i \delta(1 - \sum_j M_{ij}) \right)$$

to enforce the constraint that each row and column must contain one unit entry. Given a set of observations  $\mathbf{N}$ , this allows us to write the posterior distribution over  $\mathbf{M}$  and the predictive

distributions for action  $a$  as:

$$p(\mathbf{M}|\mathbf{N}, c) = \frac{p(\mathbf{N}|c, \mathbf{M})p(\mathbf{M})}{p(\mathbf{N}|c)} \quad (15)$$

$$p(n_{D+1} = j|\mathbf{N}, c, a) \propto \sum_{\mathbf{M}} c^{(\mathbf{m}^a)^\top \mathbf{d}^j} \bar{c}^{(1-\mathbf{m}^a)^\top \mathbf{d}^j} p(\mathbf{M}|\mathbf{N}, c) \quad (16)$$

where  $d_i^j = \delta_{ij}$ . Figure 2A shows the mean of the posterior distribution  $p(\mathbf{M}|\mathbf{N}, c)$  for three different values of  $c$ . As the  $c$  is shared between actions, and  $\mathbf{M}$  assumes that all outcomes are achievable, this would mean that either, for  $c \rightarrow 1$ , all outcomes are achievable by precisely one action, or, for  $c \rightarrow 1/L$ , no outcome is controllably achievable. Figure 2A-D illustrates the effects of such a constraint.

To relax this assumption, we allow the *number*  $|M|$  of actions with controllably attainable outcomes to vary, i.e. each row and column of the matrix  $\mathbf{M}$  can have either one unity entry, or none, as illustrated by  $\mathbf{M}_2$  in equation 14. Analogous to the previous subsection, we write a prior  $p(|M|)$  over the set size  $|M| = \sum_{ij} M_{ij} \leq L$  of controllably achievable outcomes and then integrate over it, leading to a prior over matrices

$$p(\mathbf{M}) = \sum_{|M|=1}^L p(|M|) \left[ \binom{L}{|M|} \frac{L!}{(L-|M|)!} \right]^{-1} B(\mathbf{M}) \delta \left( \sum_{ij} M_{ij} - |M| \right) \quad (17)$$

$$B(\mathbf{M}) = \left( \prod_j \left[ \delta \left( 1 - \sum_i M_{ij} \right) + \delta \left( \sum_i M_{ij} \right) \right] \right) \times$$

$$\left( \prod_i \left[ \delta \left( 1 - \sum_j M_{ij} \right) + \delta \left( \sum_j M_{ij} \right) \right] \right)$$

where  $B(\mathbf{M})$  ensures that there is at most one unity entry in each row and column of the matrix  $\mathbf{M}$ . In equation 17 we let all matrices with the same number of entries have equal prior probability. For a matrix of size  $L \times L$  with  $|M| = k$ , there are  $\binom{L}{k}$  ways of choosing the columns, and  $L!/(L-k)!$  ways of filling the columns, as we care about the order.

In order to do prediction, we need to find the posterior distribution on the number of controllably achievable outcomes  $|M|$ , given the data. It is also desirable to do inference to ensure that this formulation is an invertible generative model. The posterior is given by:

$$p(|M| = k|\mathbf{n}, c) = \frac{\sum_{\mathbf{M}:|M|=k} p(\mathbf{N}|\mathbf{M}, c)p(\mathbf{M}|k)}{\sum_{|M|} p(|M|) \sum_{\mathbf{M}:|M|=k} p(\mathbf{N}|\mathbf{M}, c)p(\mathbf{M}|k)} \quad (18)$$

Thus if the prior  $p(|M|) = \delta(|M| - L)$ , we return to the previous setting where all outcomes have to be achievable if  $c$  is large enough. For priors that have mass on smaller  $|M|$ , not all outcomes have a dedicated action. Figure 2E-H provide an illustration of the consequence of decreasing  $|M|$ .

As a further check that the parametrisation indeed has effects that are identifiable, we will infer the ML setting of  $c$  and  $|M|$ , conditional on observations. This is again straightforward doing Expectation Maximisation. Figure 3A displays the characteristics of inference of  $c$  from data  $\mathbf{N}$ . Inference is very accurate. We will also look at the characteristics of generalisation based on  $|M|$  and would thus like to infer it. Figure 3B and C show the performance of this inference. At small observation numbers, there is naturally little evidence for the low-control settings, and  $|M|$  is overestimated.

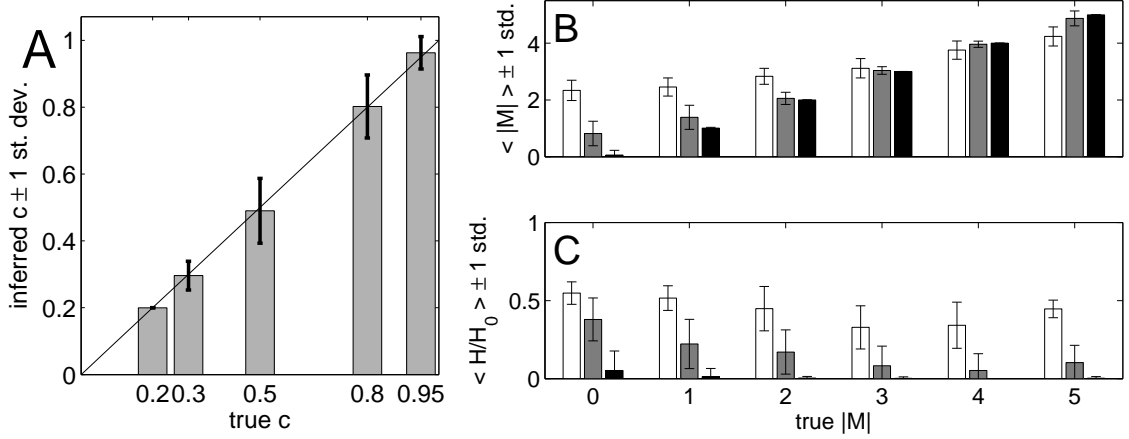


Figure 3: **A**: Inference of  $c$  from outcome data by averaging over  $p(\mathbf{M})$  as defined in equation 17 using EM. For each inference, a total of 20 observations were obtained from a randomly chosen matrix  $\mathbf{M}$  with the true underlying  $c$ , i.e. approx. 4 observations on each of  $L = 5$  actions. **B**: Inferring  $|M|$  from outcome data using EM. White bars are for a total of 20 observations, grey bars for 50 and black bars for 100 observations. The black bars are very near the true values.  $L = 5$  and  $c = 0.9$ . For small numbers of observations, the number of controllably achievable outcomes  $|M|$  is overestimated, but with little confidence. **C**: Ratio of entropy of  $p(k|\mathbf{N}, c)$  and a flat distribution with entropy  $\mathcal{H} = -\log(1/N) \approx 1.6$

## 2.1 Exploration, incentive contrast and average reward with multiple actions

To illustrate both the commonalities and the differences with the previous setting that neglected relations between actions, we again apply it to the vending machine with  $|A| = 5$  buttons,  $L = 5$  possible outcomes and in which we are allowed to press  $D = 4$  buttons. Pressing button  $o$  preferentially leads to outcome  $o$ . Only one button (button 1) has ever been taken before (4 times) and it has always yielded outcome 1 with reward 0. Let the rewards for the outcomes be  $R = [0 \ 0.2 \ 0.23 \ 0.27 \ 0.3]$  which has the property that the expected value of unexplored actions ( $1/L \sum_o R_o = 0.196$ ) is just smaller than the reward associated with the second action. The four other buttons  $a_i$  result in one particular outcome  $o = i$  with highest probability. Button 5 is therefore, unbeknownst to us, the best action. For illustration, let us force exploration to proceed in an ordered manner, from action 1 to 5, i.e. if we decide to try a new button we have to try the next one in the sequence — we can't just jump ahead and try button 5 (there is also no reason why we should want to, given that we know nothing about either of the buttons 2-5). Then, the exploration depth — the action at which exploration ceases — is a measure of the degree of exploration “drive”. Figure 4A shows the consequences of different priors on the exploration depth. Priors are hard and only allow predictions exactly consistent with  $\mathbf{M}$  matrices of a particular  $|M|$ .

- $|M| = 0$ : We believe that no button will reliably lead to any outcome. Even after observing the first outcome 4 times, the predictive distribution is flat for all actions, *including* button 1. Button 1 looks as good as all other buttons about which no information has been gathered. Figure 4A shows that all four draws for a prior that enforces  $|M| = 0$  result in the choice of button 1.
- $|M| = 1$ : We believe that one button will reliably lead to one of the outcomes. The ML

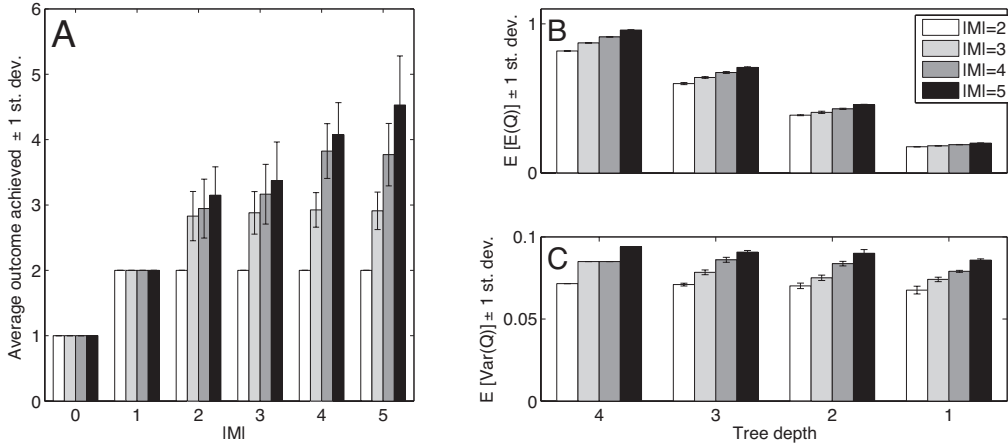


Figure 4: Effect of prior belief about fraction of controllable outcomes on exploration and expected rewards. **A:** Exploration.  $D = 4$  remaining actions, with  $A = L = 5$  actions/outcomes. The bars show which action is taken on the first (white,  $D = 4$ ), second (light grey ( $D = 3$ )), third (dark grey ( $D = 2$ )) and fourth (black ( $D = 1$ )) trials on average over many trials. The bar groups for priors putting exclusive mass on  $|M| = \{0, 1 \dots 5\}$  controllable outcomes. As more outcomes are assumed to be controllably achievable, exploration proceeds further. Action 5 was reached 40% of the time when the prior assumes that all actions are controllably achievable ( $|M| = 5$ ), and only 5% of the time when  $|M| = 4$ . However, for  $|M| = 5$  the variance of the third and fourth trials are large as well because a spurious high reward on one of the other actions (which here occurred in 2/10 trials) leads to exploitation of that action. **B:**  $Q$  values for priors peaked on  $|M| = \{2 \dots 5\}$ . Bars show the mean of the average  $Q$  values across all states, over all trials. The error bars indicate the standard deviation over trials. In all cases, a prior that assumes larger fraction of controllably achievable outcomes on average leads to higher expected rewards. **C:** Variance of the  $Q$  values across states. Bars indicate mean variance, error bars indicate standard deviation of the  $Q$ -value variance over trials. A high control prior leads to larger differences between the value of actions—a larger incentive contrast between actions.

estimate of  $\mathbf{M}$  has its only nonzero entry on button 1 and outcome 1, all other buttons are assumed to generate any of the  $L$  outcomes randomly. Due to our choice of  $R$ , button 2 is advantageous over button 1. Thereafter, button 2 will be chosen, as its outcomes (outcome 2 with  $R_2 = 0.2$ ) are marginally larger than those from the unknown actions. Figure 4A shows that all four actions for a prior that enforces  $|M| = 1$  result in the choice of button 2.

- $|M| = 2$ : We believe that two buttons will reliably lead to two different outcomes (one outcome each). Again, button 2 looks better than button 1 for the first action choice. Thereafter, however, there is a chance that the second nonzero entry is assigned to button 3 / outcome 3. The predictive distribution for button 3 will not be flat, and thus the expected outcome for that button will be greater than the expected reward for button 2. However, exploration will mostly stop at button 3, as shown by the set of columns in figure 4A for  $|M| = 2$ .
- As the matrix  $\mathbf{M}$  is constrained to contain more nonzero entries in different columns, i.e. as we believe more and more of the outcomes are achievable through some button, exploration proceeds until all buttons have been explored.

These exploration effects are due to a graded analogue of the effects shown in figure 2. Figure 4B and C also show that, similarly to the previous setting, the average  $Q$  values increase as the priors put more mass on larger  $|M|$ , and that larger  $|M|$  mean actions differ more in their expected rewards.

### 3 Control over desirable outcomes

We now proceed to define the third notion of control, which is written as a function of the fraction of total available *positive reinforcement* (or equivalently safety), that is controllably achievable. Let  $\mathbf{R}$  be the vector of reinforcements for each of the outcomes for all actions. To ensure the present definition holds for both punishments and rewards, define  $\tilde{\mathbf{R}} = \mathbf{R} - \min_i R_i$ , and then the fractional positive reinforcement for each outcome as  $\mathbf{r} = \tilde{\mathbf{R}} / \sum_j \tilde{R}_j$ . It would also be possible to divide by the maximal reinforcement. A matrix  $\mathbf{M}$  then allows control over a fraction  $\mathbf{r}^T \mathbf{M}$  of the reinforcers, and given the data, the average fraction of reinforcers that is controllably achievable can be written as:

$$\chi_M = \sum_i r_i \sum_j [\mathbb{E}[\mathbf{M}|\mathbf{N}, c]]_{ij} \quad (19)$$

If the expectation is over the set of matrices that have at most one unity entry per column, then  $0 \leq \chi_M \leq 1$ . This definition has a strange relation to  $c$ , which is used in the construction of the posterior  $p(\mathbf{M}|\mathbf{N}, c)$ , but neglected thereafter. For example, for  $c$  close to  $1/L$ , it may still be that the posterior mean is dominated by a full-rank  $\mathbf{M}$ , which would imply high fraction of controllably achievable outcomes although each of the actions has very little preference for a particular outcome. The above metric is readily corrected by a linear mapping:

$$\chi = \frac{Lc - 1}{L - 1} \chi_M \quad (20)$$

which now incorporates  $c$  fully. A value of  $\chi$ , given a known reward vector  $\mathbf{R}$ , can be incorporated as an additional linear constraint on  $\mathbf{M}$ . We write, instead of equation 17, the following:

$$p^*(c, \mathbf{M}|\chi) = \exp\left(-\frac{(\chi - \frac{Lc-1}{L-1} \mathbf{r}^T \mathbf{M} \mathbf{1})^2}{2\sigma^2}\right) \quad (21)$$

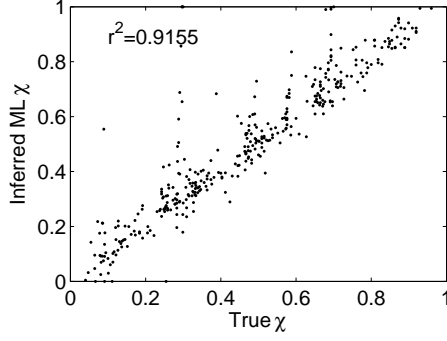


Figure 5: Inference of controllable reinforcement parameter  $\chi$  from outcome observations  $\mathbf{N}$  via EM. The parameter  $\chi$  is recovered accurately throughout the parameter’s range.

where  $\mathbb{1}$  stands for a column vector of ones and the superscript  $*$  indicates that it is an unnormalised quantity. The value of  $\sigma$  defines how strictly  $\chi$  is imposed. We will evaluate the integral over  $c$  by importance sampling, and generally let  $\sigma$  be small.

The prediction of new events  $n_{D+1}$  given  $D$  observations now has the following shape:

$$p(n_{D+1}|\mathbf{N}, \chi, \mathbf{r}) = \sum_{\mathbf{M}} \int dc p(n_{D+1}|\mathbf{M}, c) \frac{p(\mathbf{N}|\mathbf{M}, c)p(c, \mathbf{M}|\chi)}{\sum_{\mathbf{M}} \int dc p(\mathbf{N}|\mathbf{M}, c)p(c, \mathbf{M}|\chi)} \quad (22)$$

Figure 5 shows that  $\chi$  again has identifiable effects on observations.

For the application to learned helplessness (Figure 7 in the main text), we will instead evaluate the posterior probability at a number of points and construct an approximation to the posterior distribution:

$$p(\chi|\mathbf{N}) \approx \sum_i w_i \delta(\chi - \chi_i) \quad (23)$$

$$w_i^* = \frac{p(\chi_i)}{p(\mathbf{N})} \sum_{\mathbf{M}} \int dc p(\mathbf{N}|\mathbf{M}, c)p(c, \mathbf{M}|\chi_i) \quad (24)$$

where  $w_i = w_i^* / \sum_j w_j^*$ , where the posterior is represented as a sum of delta functions, and where the integral is evaluated by importance sampling.

## References

- Dearden, R., Friedman, N., and Andre, D. (1999). Model-based Bayesian exploration. In *Proceedings of the fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–9, Stockholm. [2](#)
- Dearden, R., Friedman, N., and Russell, S. (1998). Bayesian Q-learning. In *Proceedings of the fifteenth National Conference on Artificial Intelligence*, pages 761–8. [2](#)
- Friedman, N. and Singer, Y. (1999). Efficient Bayesian Parameter Estimation in Large Discrete Domains. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press. [2](#)
- Gibbon, J., Berryman, R., and Thompson, R. L. (1974). Contingency spaces and measures in classical and instrumental conditioning. *J Exp Anal Behav*, 21(3):585–605. [2](#)

- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press, Cambridge, UK. 3
- Maier, S. and Seligman, M. (1976). Learned Helplessness: Theory and Evidence. *Journal of Experimental Psychology: General*, 105(1):3–46. 2
- Overmier, J. B., Patterson, J., and Wielkiewicz, R. M. (1980). Environmental contingencies as sources of stress in animals. In Levine, S. and Ursin, H., editors, *Coping and Health*. Plenum Press. 2