**Hierarchical bayesian model.**

**Notation**

For simplicity's sake, we will note P(t), P(x) and P(r) the probabilities P(T=1), P(X=x) and P(R=r) respectively, where T is the random binary variable describing if any given trial is target (T=1) or nontarget (T=0); R is a random variable associated to the block rule; and X is a variable associated with the stimulus set presented in a trial.

A subscript P(tn), P(xn) will be used to refer to this probability for trial n.

We will use bold notation **xn**=(x1...xn) and **tn**=(t1...tn) to refer to the history of previous trials (stimulus and target respectively).

The same notation will be used to refer to conditional probabilities.

**Assumptions**

Our hierarchical bayesian model (HBM) is built under very few assumptions.

As human subjects did, it knows the space of possible rules and the output (target or nontarget) that each rule predicts for each pair of stimuli presented during a trial. The HBM doesn't take into account (as humans weren't told, either) the fact that each block had 50% target trials and 50% nontarget trials. The HBM was simulated independently on each different block. Thus, the history of rules in previous blocks wasn't exploited to predict the current rule.

Another prior was that the underlying rule r and any observation xi were uniformly distributed across their respective spaces.

The probability for the current trial of being target was thus dependent on the probability of the underlying rule and the history of trials within the block, but not the type of block (i.e., which were the relevant dimensions).

In familiar blocks, rules associated with the irrelevant dimensions were discarded (see Implementation).

**Structure**

Our hierarchical bayesian model uses two layers: the target layer and the rule layer.

The first «target» layer estimates the probability for a certain trial of being target: P(tn|**xn,tn-1**). This is calculated as the marginal probability across all the possible rules:

P(tn|**xn,tn-1**) = Σr { P(tn|r,xn) P(r|**xn-1,tn-1**) } (eq. SE1)

where Σr is the sum across all possible rules (i.e., the 6x6 combinations of features between left and right sides). Note that the target probability P(tn|r,xn) was only dependent on the underlying rule and the current observation, while the probability of a certain rule P(r|**xn-1,tn-1**) depended on the history of previous trials.

The second «rule layer» estimates, from the history of previous trials, the probability for each rule of being the underlying one: P(rn|**xn-1**,**tn-1**). For instance, it can be shown that

P(r|**xn-1,tn-1**)

= P(**tn-1**|r,**xn-1**) / Σr { P(**tn-1** |r , **xn-1**) }

= P(r) P(**xn-1 , tn-1** | r) / P(**xn-1 , tn-1**)

= P(r) P(**tn-1** | r , **xn-1**) / P(**tn-1** | **xn-1**)

= P(r) P(**tn-1** | r , **xn-1**) / Σr { P(**tn-1** |r , **xn-1**) P(r | **xn-1**) }

= P(**tn-1** | r , **xn-1**) / Σr { P(**tn-1** |r , **xn-1**) }

= Πi=1..n-1 { P(ti | r , xi) } / Σr { P(**tn-1** |r , **xn-1**) } (eq. SE2)

where Πi=1..n-1 is the product operator, and under the additional assumptions that **xn** and r are independent, and xi is independent across different trials i.

This probability is 0 if the rule is inconsistent with the history of previous trials, and is equally distributed across all remaining (consistent) rules. For example, if four possible rules were consistent with the history of trials, the probability for each of these would be of 0.25, while 0 for any other rule.

This model is shown to be optimal based on the previous priors.

**Implementation**

We created an instance of this model by keeping track of all possible rules in a 6\*6 binary matrix – called the *candidates matrix* **C**. This matrix allowed us to keep track of the rules that were consistent in previous trials. Each cell (i,j) in this matrix was associated with each possible rule on each side, where i,j ∈ {1...6} corresponded to the feature on the left and right sides respectively (e.g., red, green, blue, square, triangle, circle).

The *candidates matrix* **C** was used in the following way. If **Cij**=1, then the rule (i,j) was consistent with the history of previous trials – and was a candidate for the underlying rule. If **Cij**=0, then the rule (i,j) was inconsistent with the history of previous trials and could not candidate as the underlying rule for the current block.

The *candidates matrix* corresponds to the «rule layer» in our hierarchical bayesian model.

At the beginning of the block, all cells in the candidates matrix were set to 1, meaning that all rules were possibly the underlying rule for the current block. In familiar blocks, the 27 rules corresponding to the irrelevant dimensions were set to 0, leaving only 9 possible rules left.

On each trial and for each candidate rule (i.e., for each rule r=(i,j) such that **Cij**=1), we calculated the conditional probability of being target: P(tn|r,xn). This probability could be 1 (i.e., the current trial is target) or 0 (trial is nontarget) following the prediction given by each rule. From equation (SE1), the probability of a target trial, P(tn|**xn,tn-1**), was the average across these probabilities. This calculation corresponds to the «target layer» of the hierarchical bayesian model.

We used a greedy policy for the model, where the response was target if P(tn)≥0.5 and nontarget in any other case.

**Discussion**

The hierarchical bayesian model is optimal based on the following assumptions:

– independence of block rules across blocks

– independence of the stimulus presented with the history of previous stimuli

– independence of the stimulus presented with the underlying rule

– knowledge about the space of possible rules

– knowledge about the constraint due to the block cues

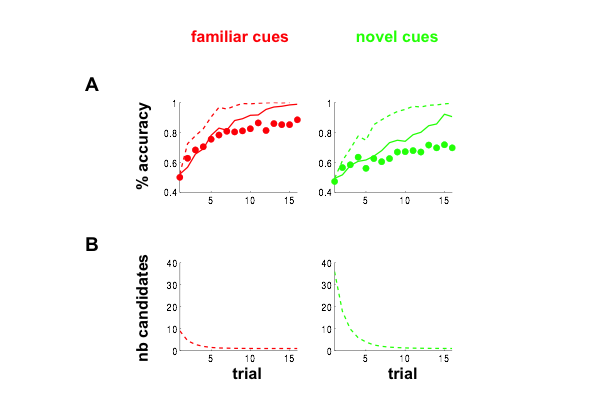
– uniform distribution across rules

– uniform distribution across observations

This model largely outperforms both the reinforcement model presented in equations (1–3) and % accuracy achieved by humans (see Fig SF1-A) by inferring the probability for each trial of being target in a bayesian fashion.

The number of consistent rules explaining the history of previous trials (see Fig SF1-B) decreases from 9 and 36 candidates (familiar and novel blocks, respectively) to 1, when the model can conclude with complete certainty what is the underlying rule for that block.

This model allows us to estimate an upper boundary on the accuracy of responses given by either human behaviour or any other model predictions.

****

**Figure SF1. A.** Behavioural data from 18 human participants (dots) and predictions of the reinforcement model (continued lines) and the hierarchical bayesian model (dotted lines). % accuracy over trials (1-16) is shown for familiar cues (left panels; red) and novel cues (right panels; green). Lines show the same data for the best-peformance parametrisation of the model. **B.** Average number of candidate rules in the hierarchical bayesian model across trials, both for familiar and novel blocks (red; green). At the beginning of the block, 9 and 36 candidate rules were used in familiar and novel bocks respectively. These values decreased to 1 by the end of the block.