**Neural mechanisms underlying**

**verification and falsification in human reasoning**

**Hierarchical Bayesian Model**

**I. Hierarchy**

has two levels.

level 1 target or nontarget?

level 2 what is the rule?

**II. Notation**

n trial number

xn is a 4-tuple (SL,CL,SR,CR) describing the observation at trial n

**xn** is a list (x1 x2 ... xn) with all of the observations up to trial n

tn is a boolean describing the trial type (target/nontarget) at trial n

**tn** is a list (t1 t2 ... tn) with all of the trial types up to time n

r is the rule of the block

**III. Mathematical formula**

**III. 1) Target probability**

For each trial n, we calculate the probability of being target given everything we know about that block (i.e., **xn** and **tn-1**).

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

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

tn is fully determined by r and xn

r is fully determined by **xn-1 , tn-1**

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

**III.2) Conditional target probability**

P(tn | r , xn) is the prediction given by rule r and the observation xn.

This takes a value of 0 or 1 and is calculated using the disjunctive rule.

**III.3) Rule probability**

P(r | **xn-1 , tn-1**) is the probability of rule r being the actual rule, given what we saw before. Using Bayes rule:

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

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

but P(x,t|r) = P(t|x,r)P(x|r) = P(t|x,r)P(x)

and P(x,t) = P(t|x)P(x) = P(t|x)P(x)

thus P(x,t|r)/P(x,t) = P(t|x,r)/ P(t|x)

= 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(r) = P(r | **xn-1**)

= P(**tn-1** | r , **xn-1**) / Σr { P(**tn-1** |r , **xn-1**) } (2)

Given independence of trials, we can say

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

P(ti | r , xi), as already explained in III.2) is sets the prediction of rule r for observation xi.

Σr { P(**tn-1** |r , **xn-1**) } thus corresponds to the total number of different rules that can explain the history of previous trials.

The probability value in equation (2) can be understood as a probabilistic distribution of rules that maximises entropy between the ones that can correctly predict the history of previous trials.

**IV. Implementation of the model**

Given the previous analysis, an HBM is equivalent to the following heuristic:

Keep track of the 36 possible candidate rules (6 features on each side).

In the familiar condition, most of these rules can already be ruled out – so only 9 rules are possible candidates.

On each trial, the probability of being target corresponds to the average response between the candidate predictions. Respond TARGET if this average is over 50%, respond NONTARGET if it's below.

Every trial after receiving feedback, rule out candidates that made wrong prediction.

This simple strategy is equivalent to all of the previous calculations.

**V. Observations**

**V. 1) Policy**

It often happens, and critically in the first trial, that the average response given by all possible candidate rules is exactly 50%. An arbitrary choice has to be made (and performance won't depend on the policy in this case).

In the current implementation, p=50% will give a response TARGET.

That said, this is not the explanation of why the model starts saying "target"at the beginning of each block. Wait for it!

**V. 2) Candidates**

As you may have already guessed, *candidates\_hbm* shows the number of candidate rules by trial. It's nice to see how this number decreases exponentially alike!

**V. 3) Plots**

I've saved and plotted another two variables to the model :

the entropy (log2) about of the rule on each side (left, right). This differs from the number of candidates (in that you may don't know about the final rule but still know it's definitely triangle on the right).

in the plot, both sides overlap.

the actual probability of the response which is not 1 as in the response, but 0.555 both for the familiar and the novel conditions. this is something the task naturally leads to.

i'm not giving to the model, so far, any aprioris about equiprobability of target/nontarget trials. thus, the model doesn't have any reason to expect it. this leads to two big points:

– the model could be improved by adding this information. it would hopefully the best one anyone could think of.

– given that participants weren't told about the equiprobability of target/nontarget, i think it's fine keeping the HBM as it is.

this could stand as an explanation of why the model (and maybe humans) start each block by saying "target".