Confirmation Bias is explained by Descending Loops in the Cortical Hierarchy

Vincent Bouttier (vincent.bouttier@ens.fr)
LNC^2, Département d'Etudes Cognitives, Ecole Normale Supérieure, 75005 Paris, France
SCALab, Université de Lille, 59000 Lille, France

Sophie Deneve (sophie.deneve@ens.fr)
LNC^2, Département d'Etudes Cognitives, Ecole Normale Supérieure, 75005 Paris, France

Renaud Jardri (renaud.jardri@chru-lille.fr)
LNC^2, Département d'Etudes Cognitives, Ecole Normale Supérieure, 75005 Paris, France
SCALab, Université de Lille, 59000 Lille, France

Abstract

In order to carry decision-making based on several pieces of evidence, one must integrate information over time. An optimal Bayesian observer would simply use Bayes' rule to combine the past knowledge with the new evidence. In this work, we tackle malfunctioning of inference, motivated by biological considerations and the recurrent structure of the brain. Allowing for loops of information when excitation and inhibition are unbalanced, we derive a functional Bayesian model of suboptimal inference, where the likelihood is corrupted by the prior knowledge. We show that, depending on the level of reverberation of the prior information, this "circular inference" model can explain cognitive biases often observed experimentally as the recency effect, the primacy effect, and the confirmation bias. The model is able to fit behavioural data on a task where healthy subjects were injected low doses of ketamine, a hallucinogenic drug thought to modify the E/I balance in favor of excitation. This work could allow to relate the microscale anomalies (E/I imbalance), the mesoscale anomalies (anomalies in frequency bands) and the macroscale anomalies (behavioural suboptimality and cognitive biases) observed in the psychotic state and under hallucinogenic drugs.

Keywords: ketamine; circular inference; computational psychiatry; belief propagation

Background

The "Bayesian brain" hypothesis has gained speed in the last decade. Current Bayesian models do not only fit behaviour, they become more and more interpretable biologically. The brain is undoubtedly unable to carry perfect inference, and that is why several approximate inference algorithms have been proposed. These algorithms include variational inference, sampling, and (loopy) belief propagation.

These recent advances in the field of Bayesian modelling are hugely beneficial to computational psychiatry. This new field of research uses models describing healthy brains, and potentially modifies these models in order to account for malfunctioning behaviors. This approach is particularly promising to explain the positive symptoms of schizophrenia (hallucinations, delusions) or involving psychedelic drugs.

Circular Inference is such a example, trying to explain inference flaws in the psychotic state. It relates the amount of malfunctioning to the level of imbalance between excitation and inhibition in a non-healthy brain.

Here we present the first attempt to fit ketamine behavioural data with the model. There are two reasons to think why Circular Inference is a good model for ketamine data. The first one is that biologically, ketamine acts on NMDA receptors and is thought to modify the excitation-inhibition ratio in favor of excitation. The second reason is that behaviourally, it has been reported that subjects have more confirmation bias under ketamine.

Computational model

Initial idea

The initial idea of Circular Inference was introduced by Jardri and Denève (2013). They make the hypothesis that the brain does inference by propagating probabilistic messages in the cortical hierarchy (in practice by using the Belief Propagation algorithm, an approximate inference algorithm). They assume additionally that unbalancing excitation and inhibition in favor of excitation modifies the algorithm by introducing information loops (the algorithmic equivalent of positive feedback loops). The probabilistic message sent between two nodes of the graphical model is corrupted by the message going in the opposite direction (see Figure 1):

$$M_{ij} = F_{ij}(\sum_{k \neq j} M_{ki} + \alpha M_{ji}) \tag{1}$$

where $\alpha = 0$ in the perfect balance case (true belief propagation algorithm in binary graphical models), and $\alpha > 0$ in the unbalanced case. α can be seen as the level of impairment of the inhibition (or amount of over-excitation in the network). Prior loops are distinguished from sensory loops by using parameter α_p (respectively α_s) if node i is under (resp. above) j in the hierarchy.

The messages are probabilistic information brought to nodes of the graphical model, such that

$$B_i = \sum_k M_{ki}$$

where B_i is the log-odds of the binary variable x_i. Mathematically $B_i = \log(p(x_i = 1)/p(x_i = 0))$.

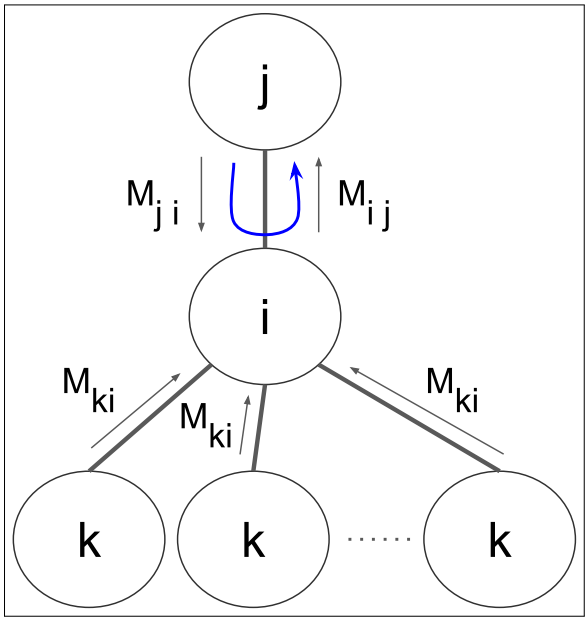

Figure 1: Concept of Circular Inference. Because probabilistic messages are exchanged in both directions, the message M_{ji} may be partially reverberated. If i is below j in the hierarchy, the new evidence (carried by nodes k down in the hierarchy) is corrupted by the prior coming from node j. See Equation 1.

F_{ij} is a sigmodal function with a parameter $w_{ij} \in [0,1]$ representing the strength of the connection between node i and node j:

$$F_{ij}(X) = F(x, w_{ij}) = \log\left(\frac{w_{ij}e^x + (1-w_{ij})}{(1-w_{ij})e^x + w_{ij}}\right)$$

Model of categorization task

Here we deal with a particular problem: a visual categorization task. In this case the generative model is modeled by a graphical model with 3 binary nodes: the evidence node e at the bottom, the representation node at the middle, and the category at the top (also used by Lange, Chattoraj, Beck, Yates, and Haefner (2019)).

Approximations (Taylor expansions at first order in α) lead to the following non-linear equation describing the update of the log-odds of the (binary) category variable:

$$B_{t+1} = F[B_p, w_p] + F[B_s + \alpha_p F(F(B_p, w_p), w_s), w_s] \quad (2)$$

Link to other computational models

Bayes' rule is a particular case of Equation 2, retrieved for $\alpha_p = 0$ and $w_s = w_p = 1$.

Equation 2 can be linearized into a simpler equation:

$$B_{t+1} \approx k_p B_p + k_s B_s \quad (3)$$

where $k_p = [1 + \alpha_p(2w_s - 1)^2](2w_p - 1)$ controls the recency-primacy bias. If prior loops (α_p) are too low then $k_p \in [0,1]$ represents an integration leak. This explains the recency effect often observed in experiments. Instead, if α_p is high enough then $k_p > 1$ (amplification), allowing to account for the primacy effect sometimes observed in experiments. Equation 3 is the model used by Baker, Konova, Daw, and Horga (2019) to fit behavioural data of patients with schizophrenia.

Equation 2 is also close to previous work by the lab (Jardri, Duverne, Litvinova, & Denève, 2017), except from it was derived mathematically from the initial definition of Circular Inference (instead of reflecting its concept qualitatively).

In conclusion, the Circular Inference model proposed here is a generalization of many functional models, but is motivated biologically, and in particular interprets amplification of information (prior loops α_p) as a shift of the E/I ratio.

The model accounts for confirmation bias

As explained in the previous paragraph, the model accounts both for primacy and recency effects, depending on the values of parameters. The first term of Equation 2 represents the natural leak of information, which could be interpreted as finite working memory. This leak is present whatever the new stimulus is. We thus interest ourselves to $B_{t+1} - F(B_p, w_p)$, seen as function $f(B_p, B_s)$ (see Figure 2). Having confirmation bias means that $f(B_p, B_s)$ is stronger if the likelihood goes in the direction of the prior ($B_p B_s > 0$) than if the likelihood contradicts the prior ($B_p B_s < 0$). We showed mathematically that having prior loops is indeed necessary and sufficient for confirmation bias.

Fitting the Circular Inference model to data

Behavioural data

18 healthy human subjects performed a task designed by Valentin Wyart and run by Alexandre Salvador and Raphael Gaillard. This task was a categorical task, variant of the "weather task". It involves accumulating probabilistic information over 4, 8 or 12 independent samples, and reporting the most probable category (see Figure 3). There were around 360 trials per session. Each subjects carried the task twice, by being injected low-doses of ketamine or a placebo. The design is cross-over, double blind.

Predictions

We predicted that the amount of prior loops (α_p) was higher in the ketamine session compared to the placebo session. There are two reasons for that. The first one is that biologically, ketamine acts on NMDA receptors and is thought to modify the excitation-inhibition ratio in favor of excitation. The second reason is that behaviourally, subjects have more confirmation bias under ketamine, as reported in experiments.

Preliminary results

The model introduced above was fitted to behavioural data. The optimisation was done using Adam, a first-order gradient-based method, and gradients were computed through back-propagation. The likelihood of the model was optimized using

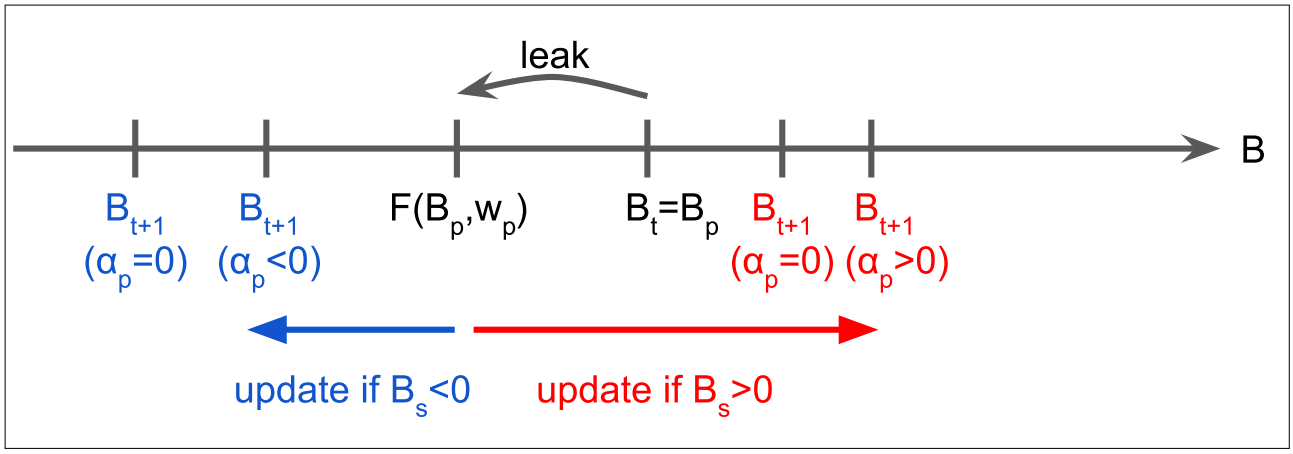

Figure 2: The model accounts for confirmation bias. Thanks to prior loops α_p, the update at sample $t+1$ of the belief B (log-odds of the C, the binary variable representing the category) is bigger if the prior B_p and the likelihood B_s have the same sign (in the example here, B_p is positive). On the contrary, the updates have the same amplitude without prior loops ($\alpha_p = 0$).

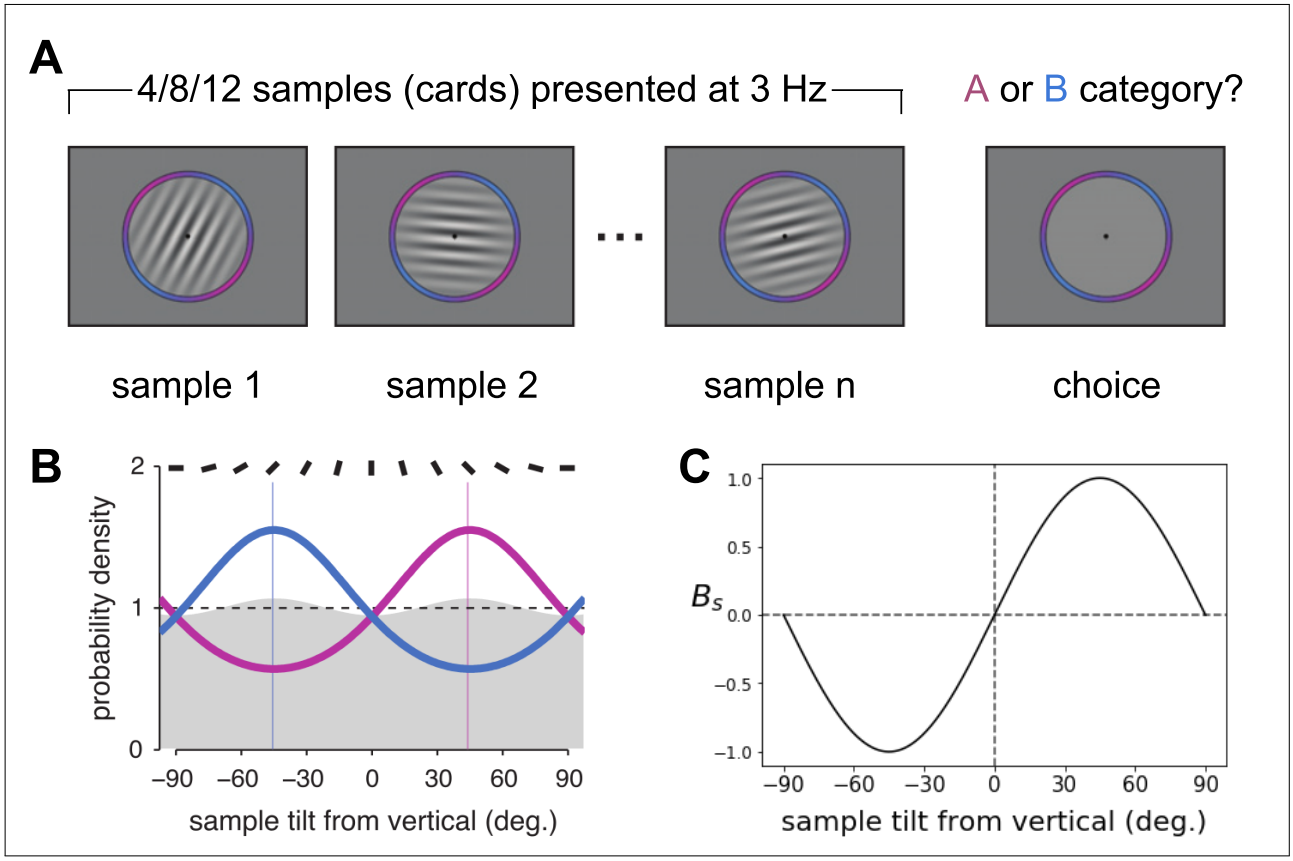

Figure 3: Description of the task. Figures A and B come from Drugowitsch, Wyart, Devauchelle, and Koechlin (2016) (A) Description of a trial. Each trial consists in a sequence of cues, after which participants take a decision based on them. (B) Samples are drawn from a generative probability distribution favoring one of the two colors. At the end of the sequence, subjects are asked to report from which distribution the samples are coming. (C) Probabilistic information brought by sensory cues depending on their orientation. $B_s > 0$ favors the pink distribution (category A).

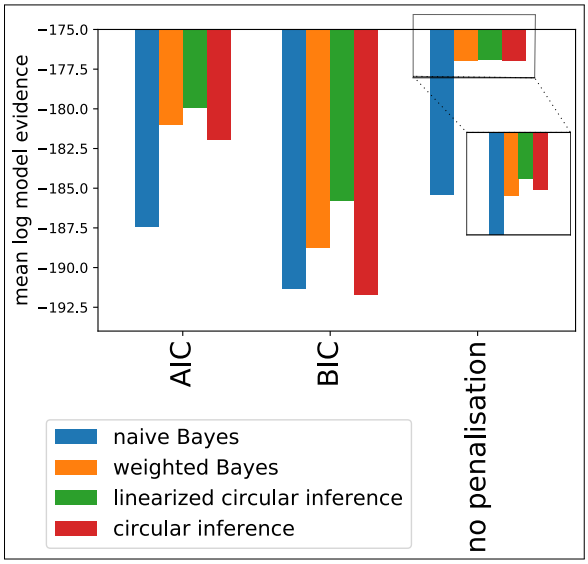

Figure 4: Model comparison with model evidences; a high model evidence means that the corresponding model fits the data well. Naive Bayes (2 parameters) is a special case of Weigthed Bayes (4 parameters), which is itself a special case of Circular Inference (5 parameters). We compare these three models between themselves, and with the linear approximation of Circular Inference.

5 free parameters: w_s, w_p, α_p and two parameters of the decision criterion (softmax), β and a bias term.

We compared models of different complexity (see Figure 4). Weighted Bayes ($\alpha_p = 0$) beats Naive Bayes ($\alpha_p = 0$, $w_s = w_p = 1$) with all 3 measures evaluating model performance. Circular Inference beats Weighted Bayes only for one participant, and is equivalent for the others, explaining that it only slightly beats it. Circular Inference gets surprisingly beaten by the equivalent linear model, not only using the penalized scores but also the (non-penalized) model evidence. This surprising result will be adressed in the future.

However, during the optimisation process, even though the likelihood converged every time, the parameters α_p and w_p kept increasing or decreasing, not oscillating. Because of that, the fitting procedure was not reproducible. Knowing that under ketamine, the working memory (equivalent of w_p) is probably impaired, it was impossible to fix w_p among the sessions (for a given subject). As a consequence, we could not compare the amounts of prior loops (α_p) between the placebo and the ketamine session. However, the fitted metaparameter k_p was reproducible, confirming that α_p and w_p could not be fitted without fixing the other parameter, and that the subjects' behaviour was close to linear. A solution would be to assess during the task the working memory, in order to find w_p and fit only α_p.

Overall, k_p is lower in the ketamine session, but it does not rule out the prediction that α_p is higher under ketamine, because the working memory could be a lot worse under ketamine, decreasing w_p.

Conclusion

We derived a model from belief propagation algorithm in a binary graphical model, to which a change was made in order to allow for reverberation of information. Many models proposed in the literature can be seen as an linear approximation and/or a particular case of the resulting non-linear model. Mathematically, the model can only explain recency effects without prior loops ($\alpha_p = 0$). On the contrary, allowing prior loops allows to account for the recency effect, the primacy effect and the confirmation bias observed experimentally. Importantly, this addition of loops to the Weigthed Bayes model allowed to explain better ketamine data. This work is a step toward explaining some behavioural and perception malfunctions as a consequence of uncompensated positive feedback because of an imbalance between excitation and inhibition in the brain.

Acknowledgments

We thank Valentin Wyart for his advice throughout the project, Marc Szafraniec and Adrian Valente for helping with PyTorch, and Alexandre Salvador and Philippe Domenech for collecting the ketamine data.

Vincent Bouttier was supported by a ANR-16-CE37-0015 PhD fellowship, led by Renaud Jardri.

References

Baker, S. C., Konova, A. B., Daw, N. D., & Horga, G. (2019). A distinct inferential mechanism for delusions in schizophrenia. *Brain*, *142*(6), 1797–1812.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1398–1411.

Jardri, R., & Denève, S. (2013). Circular inferences in schizophrenia. *Brain*, *136*(11), 3227–3241.

Jardri, R., Duverne, S., Litvinova, A. S., & Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nature Communications*, *8*, 14218.

Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L., & Haefner, R. M. (2019). A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *bioRxiv*, 440321.