# Dynamical circular inference in the general population and the psychosis spectrum : insights from perceptual decision making

Pantelis Leptourgos

# THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

**Préparée à Ecole Normale Supérieure**

# Dynamical circular inference in the general population and in the psychosis spectrum: Insights from perceptual decision making

**Ecole doctorale n°158**
CERVEAU, COGNITION, COMPORTEMENT

**Spécialité** Neurosciences et Sciences Cognitives

**COMPOSITION DU JURY :**

Mme. SERIES Peggy
University of Edinburgh, Rapporteur

M. CORLETT Philip
Yale University, Rapporteur

M. MAMASSIAN Pascal
Ecole Normale Supérieure, Membre du jury, Président

M. STEPHAN Klaas Enno
ETH, Membre du jury

**Soutenue par Pantelis LEPTOURGOS**
**le 14 Novembre 2018**

Dirigée par **Sophie DENEVE**
et **Renaud JARDRI**

ENS

PSL RESEARCH UNIVERSITY PARIS

# THESE DE DOCTORAT

Specialité: Neurosciences et Sciences Cognitives

En vue de l'obtention du grade de

**DOCTEUR de l'ECOLE NORMALE SUPERIEURE**

Ecole Doctorale: Cerveau Cognition Comportement

Présentée par:

**Pantelis Leptourgos**

Titre :

## Dynamical circular inference in the general population and in the psychosis spectrum: Insights from perceptual decision making

dirigée par: Dr Sophie Denève
Prof. Renaud Jardri

Soutenue le 14 Novembre 2018

devant le jury composé de:

| | |
|---|---|
| Prof. Peggy Seriès | Rapporteur |
| Prof. Philip Corlett | Rapporteur |
| Dr Pascal Mamassian | Examinateur (Président) |
| Prof. Klaas Enno Stephan | Examinateur |
| Dr Sophie Denève | Directrice de thèse |
| Prof. Renaud Jardri | Co-directeur de thèse |

*« …le monde n'est qu'un amas de taches confuses, jetées sur le vide par un peintre insensé, sans cesse effacées par nos larmes… »*

*Marguerite Yourcenar*

# Acknowledgements

After 4 years, this journey comes to an end. In the lines that follow, I would like to acknowledge all the people who were part of this amazing experience, making it unique.

First of all, I would like to deeply thank my advisors, Sophie Denève and Renaud Jardri for their scientific guidance and moral support. Also, for offering me the opportunity to work on an amazing, "hot" topic and giving me the freedom to put forward my own ideas. At the same time, for keeping my feet on the ground every time I got overenthusiastic!

I am grateful to all the members of the PhD committee (but also the Comité de suivi de thèse) for having accepted to read and discuss this work.

I would especially like to thank all my colleagues and collaborators: Charles-Edouard Notredame, Marion Eck, Vincent Bouttier, Maxime Tiberghien and all the members of the GNT in Paris and the ScaLab / Cure in Lille (former and present). Working and discussing with all of you during those 4 years, was a real pleasure.

I would like to acknowledge my doctorate school (ED3C) for accepting my PhD proposal, but also PSL and ENS for hosting me, and for funding my thesis.

I would like to express my most sincere gratitude to my family; my parents, my sister and Jean-Edouard for all their love and support. Having you by my side makes me feel invincible. In particular I would like to mention my grandmother and my grandfather, whose memories I always keep in my heart.

During my whole life I have had the chance to be surrounded by an "extended family". To all my friends, wherever you might be (Greece, France, Netherlands, UK, US, …), thank you so much for everything. Knowing that you are out there always makes me feel at home, no matter where I am.

Last but not least, I would like to express my utter love and gratitude to Sofia, the person with whom I shared all the good and bad moments during the past 4 years. Thank you for your patience. This thesis is dedicated to you.

# Dynamical circular inference in the general population and in the psychosis spectrum: insights from perceptual decision making


## Abstract

We live in an uncertain world, yet our survival depends on how quickly and accurately we can make decisions and act upon them. To address this problem, modern neuroscience reconceptualised perception as an inference process, in which the brain combines sensory inputs and prior expectations to reconstruct a plausible image of the world. In addition to that, influential theories in the emerging field of computational psychiatry suggest that various psychiatric disorders, including schizophrenia, could be the outcome of impaired predictive processing. Among those theories, the circular inference framework suggests that an unconstrained propagation of information in the cortex, underlain by an excitatory to inhibitory imbalance, can generate false percepts and beliefs, similar to those exhibited by schizophrenia patients. In the present thesis, we probed the role of circular inference from normal to pathological brain functioning, gaining insights from perceptual decision making in the presence of high ambiguity.

In the first part of the thesis, we focused on the role of circularity in bistable perception in the general population. Bistability occurs when two mutually exclusive interpretations compete and switch as dominant percepts every few seconds. In a **1st article**, we manipulated sensory evidence and priors in a Necker cube task, asking how the brain combines low-level and high-level information to form perceptual interpretations. We found a significant effect of each manipulation but also an interaction between the two, a finding incompatible with Bayes-optimal integration. Bayesian model comparison further supported this observation, showing that a circular inference model outperformed purely Bayesian models. Having established a link between circular inference and bistable perception, we then put forward a functional theory of bistability, based on circularity (**2nd article**). In particular, we derived the dynamics of a dynamical circular inference model, showing that descending loops (i.e. a form of circularity resulting in aberrant amplification of the priors) transform what is normally a leaky integration of noisy evidence into a bistable attractor with two highly trusted stable states. Importantly, this model can explain both the existence and the phenomenological properties of bistable perception, making a number of testable predictions. Finally, in a **3rd article**, we tested one of the model's predictions, namely the perceptual behaviour when the stimulus is presented discontinuously. We ran two Necker cube experiments using a novel intermittent-presentation

methodology, and we calculated the stabilisation curves (i.e. persistence as a function of blank durations). We found that participants' behaviour was compatible with the model's prediction for a system with descending loops, suggesting that circularity constitutes a general mechanism that shapes the way healthy individuals perceive the world.

In the second part, we studied circular inference in pathological conditions related to psychosis. We notably focused on two varieties of the psychotic experience, namely schizophrenia-related psychosis and drug-induced psychosis. After discussing the links between behaviour, aberrant message-passing and the corresponding neural networks (**4th article**), we used bistable perception to probe the computational mechanisms underlying schizophrenia in a **5th article**. We compared patients with prominent positive symptoms with matched healthy controls in two bistable perception tasks. Our results suggest an enhanced amplification of sensory inputs in patients, combined with an overestimation of the environmental volatility. In the last article (**6th**), we delineated a multiscale account of psychedelics, ultimately linking the macroscale (i.e. phenomenological considerations such as the crossmodal character of the psychedelics experience), the mesoscale (i.e. loops) and the microscale (i.e. neuromodulators and canonical microcircuits).

## Keywords

Bayesian inference, Circular inference, schizophrenia, psychosis, bistable perception, Necker cube, psychedelics, canonical microcircuit, hierarchical, functional, message-passing algorithms, dynamical

# Inférence circulaire dynamique en population générale et dans le spectre psychotique : Apports de la prise de décision perceptive


## Resumé

Nous évoluons dans un monde incertain. De ce fait, notre survie dépend de notre capacité à prendre rapidement des décisions, et ce de manière fiable et adaptative. Il est possible de mieux comprendre cette capacité en considérant la perception comme un processus d'inférence probabiliste au cours duquel les informations sensorielles sont combinées à nos attentes pour produire une interprétation plausible de notre environnement. Les théories récentes de psychiatrie computationnelle suggèrent par ailleurs que la grande variabilité des troubles psychiatriques, au rang desquelles figure la schizophrénie, pourrait résulter d'une altération de ces mêmes processifs prédictifs. L'*Inférence Circulaire* est l'une de ces théories. Ce cadre de pensée stipule qu'une propagation incontrôlée d'information dans la hiérarchie corticale pourrait générer des percepts ou des croyances aberrantes. Afin d'explorer le rôle joué par l'*Inférence Circulaire* en condition normale ou pathologique, ce travail de thèse s'est appuyé sur des tâches de prise de décision en conditions perceptives ambigües.

Dans une première partie, nous nous sommes intéressés au rôle joué par la circularité dans la perception bistable. Le phénomène de bistabilité survient lorsque deux interprétations se succèdent à intervalle régulier pour un même percept. Nous présentons les résultats d'une tâche conduite en population saine où nous avons manipulé les informations sensorielles et à priori utilisées par les participants lors de la visualisation d'un cube de Necker (**article 1**). Nous avons pu montrer un effet propre à chaque manipulation, mais également une interaction entre ces deux sources d'information, incompatible avec une intégration Bayésienne optimale. Résultat confirmé par la comparaison de divers modèles computationnels ajustés aux données, qui a pu mettre en évidence la supériorité de l'*Inférence Circulaire* sur les modèles Bayésiens classiques.

Nous avons ensuite voulu tester un modèle fonctionnel de la bistabilité (**article 2**). Nous avons donc dérivé la dynamique du modèle et montré que la présence de boucles descendantes dans la hiérarchie corticale, transformait ce qui était jusque là un intégrateur imparfait du bruit sensoriel en *modèle à attracteur bistable*. Ce modèle ne reproduit pas seulement le phénomène de bistabilité, mais également l'ensemble de ces caractéristiques phénoménologiques. Dans un **3ème article**, nous avons testé une prédiction, notamment en cas de présentation discontinue

d'un stimulus bistable. Deux expériences complémentaires utilisant un paradigme de présentation intermittente du cube de Necker ont donc été conduites en population générale. Nos resultats etaient compatible avec les prédictions faites par le modèle de *l'Inférence Circulaire Dynamique*, suggérant que la circularité puisse être un mécanisme générique à l'origine de notre façon de voir le monde.

Dans la seconde partie de ce travail, nous avons étudié *l'Inférence Circulaire* en condition pathologique, notamment lors d'expériences psychotiques (schizophrénie, psychédéliques). Nous avons utilisé la perception bistable pour explorer les mécanismes computationnels à l'œuvre dans la schizophrénie (**article 4,5**). Nous avons comparé les performances de patients présentant des symptômes psychotiques à des témoins sains appariés lors d'une tâche de perception bistable. Nous avons pu montrer chez les patients une amplification des informations sensorielles combinée à une surestimation de la volatilité environnementale. Enfin nous terminons ce travail en proposant une approche transversale de l'effet des psychédéliques (**article 6**), sur la base des résultats précédents et de la spécificité clinique de ces expériences sensorielles cross-modales, afin de relier l'échelle macroscopique (i.e., comportement et phénoménologie), mésoscopique (i.e., les boucles inférentielles) et microscopique (i.e., les différents neurotransmetteurs impliqués aboutissant à un microcircuit canonique).

## Mots-clefs

Inférence Bayésienne, Inférence circulaire, schizophrénie, psychose, perception bistable, Cube de Necker, psychédéliques, microcircuit canonique, hierarchique, fonctionnel, dynamique

# Contents

# Chapter 1

## General Introduction

The brain constantly processes a multitude of complex information. As all information processing systems, it receives some input (visual, auditory, proprioceptive etc.) and transforms it into some output (percepts, decisions, actions etc.). What is the goal of the brain's computation? How is information represented by the brain and what are the well-defined steps of this transformation? What is the physical substrate in which it takes place? All these questions have fuelled a prolific scientific debate on brain functions for decades.

In his seminal book about vision, David Marr related those questions to his famous 3 levels of analysis [1]: (i) Computational level; (ii) Algorithmic level and (iii) Implementational level. He argued that any comprehensive analysis of the brain (or any other information processing system) must offer an in-depth understanding of all three dimensions. Importantly, although each level is constrained by the other two, they remain largely independent: many algorithms can achieve the same goal (e.g., summation can be done with binary or decimal variables) while the same algorithm can be implemented by different hardware.

In this thesis, we studied perception (or aberrant perception, in the context of psychosis) using a top-down (functional) approach. Our argumentation starts from the computational level, by considering what is the task undertaken by perceptual systems (make decisions under uncertainty; probabilistic inference) and how it should be accomplished (normative approach). Based on that, we derived a detailed algorithm performing this task (message-passing algorithms; belief propagation; Hidden Markov Model), that is compatible with brain's anatomical and physiological structure (hierarchical structure; long-range reciprocal excitatory connections; neural excitation to inhibition balance). Additionally, we compared participants' behaviour in perceptual tasks with the optimal behaviour for this task (or patients' behaviour with that of matched healthy controls), detecting potential deviations from the normative approach. Finally, we suggest links between the different information processing steps and specific neural substrates (neuromodulators, cortical microcircuits, long-range networks).

In this general introduction, we provide the necessary information and assumptions behind this work, and present an outline of the thesis.

## The Bayesian brain hypothesis

During the 20$^{th}$ century, a dominant scientific view illustrated the brain as a complex filter, extracting valuable information from the different sensations in a passive way [2]. This feature extraction took place at the level of neurons [3] and it was considered a gradual process,

with simple features preceding the more complex ones [4,5]. According to this view, the sensory signal already contains all the necessary information; one needs to look carefully and detect all the relevant cues, from simple orientations to complex 3D structures.

The 21st century witnessed an important paradigm shift [6]. The brain has to make decisions in an uncertain and constantly-evolving environment. In most cases, available information is degraded or simply limited (visual perception in dark places; 3D perception based on the 2D retinal image etc. [7]). Besides, any plausible interpretation of the sensory data is necessarily context-dependent (e.g. the colour of a surface depends on the background colour). From this point of view, perception (but also other brain functions such as decision making or motor-control) is an ill-posed problem, which can only be approached using statistical methods.

Probabilistic inference offers the necessary tools to make decisions under uncertainty. **Figure 1** illustrates an example of a problem with inherent uncertainty (named the beads task; [8]). In such problems, the optimal strategy corresponds to estimating the level of the uncertainty and then use this estimate to make decisions. In our example (**Figure 1a**), this coincides with using Bayes theorem to combine low-level information (i.e. percentages of red and black balls) with high-level prior information (i.e. preference) to estimate the posterior probability that a red ball was drawn from urn 1 (or urn 2 respectively). Then, we can use this information to choose the most probable cause (Urn 1 in our case).

The previous example illustrates the simplest possible causal link between two variables. In reality, we usually face problems where many variables interact with each other, forming complex hierarchical causal structures. A relatively simple hierarchy is illustrated in **Figure 1b**. As before, a ball is drawn from one of the two urns but unlike the previous example, the urns are also not fixed, but picked randomly (and according to a higher-level preference) between two different sets. In order to make optimal decisions, one still has to combine low-level and high-level information, but because of the hierarchical causal structure, this computation is performed at all levels. In particular, every level is constrained from the level below by a likelihood function and from the level above by a prior. This local computation of posterior probabilities gives rise to powerful, biologically plausible algorithms (i.e., message-passing algorithms) which can be used by the brain to perform perceptual and other cognitive tasks (**Figure 1c**) [9]. From this point of view, perception corresponds to the process of combining sensory and prior information in order to find optimal interpretations of the world [10].

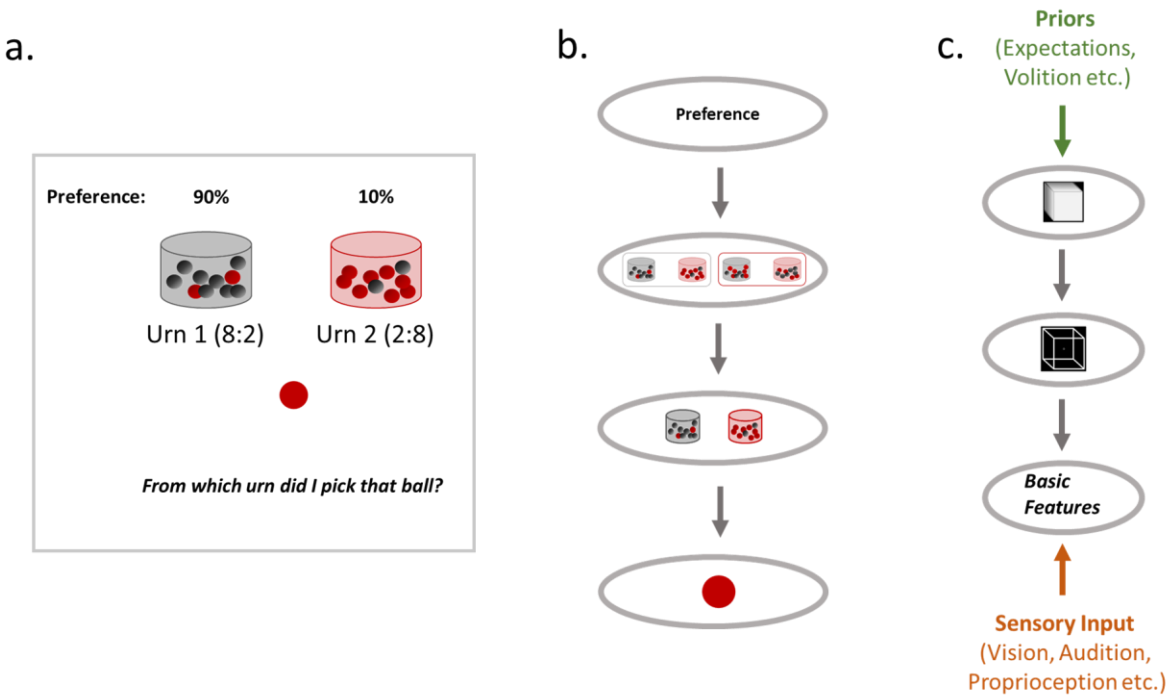


***Figure 1: Probabilistic inference.*** ***(a.)*** *The beads task. A ball is picked randomly from one of the two urns. In order to make an optimal decision, we should combine the likelihood (probability of picking a red ball from urn 1/2) with the prior (preference), according to Bayes theorem.* ***(b.)*** *A hierarchical causal model. When there is a chain of causal links, we can use message-passing algorithms to calculate posterior probabilities.* ***(c.)*** *The Bayesian brain hypothesis. The brain learns the causal structure of the world and combines sensations and prior knowledge in order to construct plausible interpretations of the world*

## Message-passing algorithms

### *Predictive Coding*

An important algorithm can be obtained by plugging Gaussian variables in the hierarchy. In its simplest form (2 variables), predictive coding can be written in the following way [11]:

$$\hat{x}_{new} = \hat{x}_{prior} + k\left(s - \hat{x}_{prior}\right) \quad (1)$$

with $s$ corresponding to the mean of the likelihood, $\hat{x}_{prior}$ to the mean of the prior (prior estimate) and $\hat{x}_{new}$ being the mean of the posterior (updated estimate). The difference $\left(s - \hat{x}_{prior}\right)$ is called prediction error and corresponds to how well the model can predict the state of the world (if prediction error is zero, then the model predicts perfectly the new sensation, consequently it is not updated). Importantly, the prediction error is weighted by $k$,

which is a parameter that depends on the precisions of the likelihood (inverse of variance $\sigma_s^2$) and the prior (same for $\sigma_{prior}^2$):

$$k = \frac{\sigma_{prior}^2}{\sigma_{prior}^2 + \sigma_s^2} \quad (2)$$

As a result, the updating of the model also depends on the reliability of the new information, compared to the old one.

When there is a hierarchy of causes, equation (1) can be written as follows for level i [11] (see also **Figure 2**):

$$\hat{x}_{i,new} = \hat{x}_{i,old} + k_f \varepsilon_i - k_b \varepsilon_{i+1} \quad (3)$$

where $\hat{x}_{i,new}$ is the updated estimate, $\hat{x}_{i,old}$ is the old estimate before the update, $k_f$ and $k_b$ are the feedforward and feedback weights (defined by equations similar to (2)) and $\varepsilon_i$, $\varepsilon_{i+1}$ are the lower-level and upper-level prediction errors:

$$\varepsilon_i = \hat{x}_i - \hat{x}_{i-1} \quad (4)$$

Several versions of this algorithm have been proposed [12–15]. They all suggest that different populations of neurons encode the predictions (feedback) and the prediction errors (feedforward signal), roughly corresponding to the distinction between superficial and deep layers [16]. They also postulate an inhibitory effect of the feedback, which attenuates the activity of prediction error-encoding neurons when the stimulus is predictable [17].

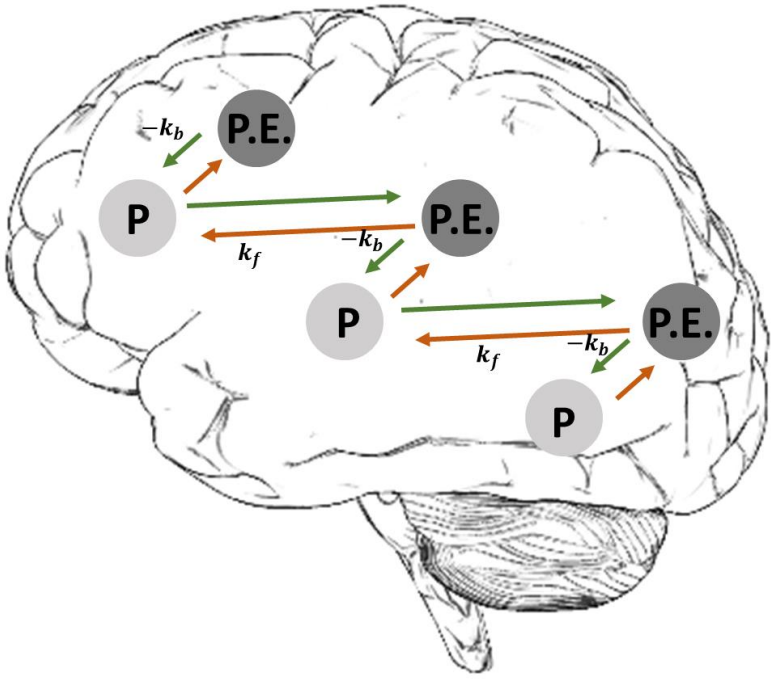


***Figure 2: Predictive coding.*** *Prediction errors ascend the hierarchy, while top-down predictions descend the hierarchy. Predictions and prediction errors are represented by different populations of neurons, in different parts of the cortical column (deep and superficial layers respectively).*

*Belief Propagation*

An alternative way to do hierarchical Bayesian inference is belief propagation. A detailed mathematical description of the algorithm can be found in the **Supplementary Material** of **Chapter 3** of the present thesis (see also **Figure S1** of the same chapter for an illustration).

Contrary to predictive coding, belief propagation can be used with any type of variable (including discrete / binary variables, ideal for modelling decisions in ''2-alternative forced-choice tasks'') [18]. As in Bayes theorem, posteriors are calculated by combining priors and likelihoods (summing them in the log-scale). Consequently, the bottom-up channel transmits the sensation per-se, and not the difference between sensation and prediction [11]. This has an important physiological consequence: feedback is not necessarily suppressive, in agreement with various neurophysiological [19,20] and behavioural studies [21]. Note also that unlike predictive coding, belief propagation calculates and represents distributions [10,22,23], and not just estimates.

## Computational psychiatry and psychosis

During the last decade, a second intellectual revolution occurred in the field of psychiatry. At the end of the 20th and beginning of 21st century (but even today), psychiatry faced important challenges, related to classification, diagnosis, treatment and pathogenesis of mental disorders [24,25]. In particular, the symptom-based nosology fails to capture the underlying mechanisms that trigger the disorders, and has no means to predict the effectiveness of treatments.

Computational psychiatry hopes to fix those shortcomings of classical psychiatric practice, by characterizing psychopathologies in terms of mechanistic and computational dysfunctions over multiple scales [26–28] (note that this is also the aim of the Research Domain Criteria initiative [29]). This characterization can trigger a new type of classification of disorders, based on the objective causes instead of the subjective reports. As a result, it can fill the epistemological, explanatory gap that exists between neurobiology and the clinical level (e.g., "excessive dopamine release results in psychotic symptoms"), also guiding the choice of treatments for different clusters of patients.

Interestingly, computational psychiatry is a very active field of research and plenty of theories have already emerged. For example in schizophrenia, at least two computational theories have been suggested, based on the message-passing algorithms presented in the previous section (please refer to [30] for a comprehensive review). Predictive coding accounts on one hand suggest that an impairment in the precision weighting of the prediction error (that may be due to dopaminergic abnormalities or NMDA hypofunction), causes the system to rely more on priors or sensory evidence (**Figure 3a,b**), resulting in false percepts and delusional thoughts [31–34].

In belief propagation on the other hand, because of the presence of recurrent excitation, it is crucial to set apart old and new information, to avoid aberrant overcounting. Inadequate elimination of redundant information results in *circular inference*, a pathological form of inference in which information is amplified, affecting the calculation of the posteriors and thus, the decision making process (**Figure 3c,d**) [35]. Recent behavioural evidence suggest that information loops exist in patients with schizophrenia but also, in a more moderate form, in healthy individuals [36]. Furthermore, different types of circularity (related to amplification of priors or sensory inputs) might trigger the different dimensions of schizophrenia (positive, negative and cognitive disorganization clusters), offering a compelling explanation of the heterogeneity of the disease.

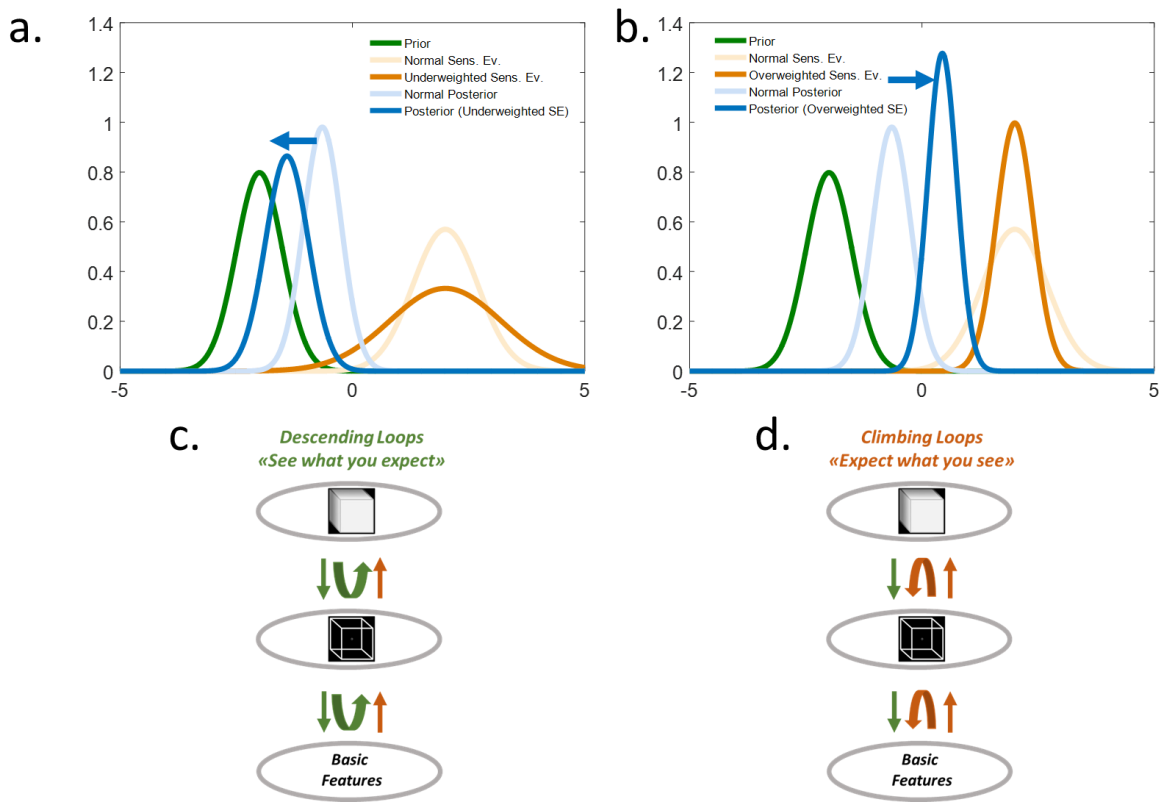


***Figure 3: Impairments in predictive processing could generate psychotic symptoms.*** *(**a.,b.**): Predictive coding. Underweighting **(a.)** and overweighting **(b.)** of the sensory evidence (same for overweighting / underweighting of priors) result in a shifted posterior (towards prior / sensory evidence). (**c.,d.**): Circular inference. The presence of loops results in the amplification of priors (descending loops) or sensory evidence (climbing loops), resulting in a system that "sees what it expects" or "expects what it sees".*

## Outline of this thesis

This thesis is structured on a series of articles. It consists of 6 largely independent chapters / articles (**Chapters 2-7**), divided into 2 sections. **Chapter 8** is a general discussion in which we summarize the main findings of the work and suggest potential future directions.

In section 1 (**Chapters 2-4**), we focus on the problem of bistable perception. Bistable perception's uniqueness resides in the fact that it constitutes one of the few known cases of clear dissociation between stimulation and perception, in which dynamics is also an important component (e.g. compare bistable perception with other known visual illusions [37]). This dissociation offers a unique opportunity to study the computational mechanisms generating conscious (or unconscious) perception. Here we build upon previous evidence suggesting that healthy subjects exhibit mild form of circular inference [36]. We argue that if this is the case,

then the footprints of circularity should be present in participants' behaviour during perceptual tasks. Given that the effects of loops are more pronounced in ambiguous environments [35], bistable perception constitutes the ideal task. In **Chapter 2**, we present the results of an experiment, designed to study how healthy participants combine sensory evidence and priors. An interaction between the two offered evidence in favour of our claim. A Bayesian model comparison confirmed the superiority of circular inference models compared to classical Bayesian models. Having established a first link between bistability and circularity, we set to put forward a complete theoretical account of perception under ambiguity (**Chapter 3**). We derived the dynamics of a *dynamical circular inference* model, showing that loops change a leaky integrator into a bistable attractor. This functional model explains both the existence and the phenomenological properties of bistable perception, making a number of testable predictions. In **Chapter 4**, we tested one of the predictions, namely the perceptual behaviour when the stimulus is presented discontinuously. We found that participants' behaviour was compatible with the model's prediction for a system with descending loops. Overall, in this first part of the thesis we offer theoretical and experimental evidence that circularity constitutes a general mechanism that shapes the way healthy individuals perceive the world.

In the second section of the work (**Chapters 5-7**), we explore the links between circular inference and the psychosis spectrum. In **Chapter 5**, we review the connections between behaviour, aberrant message-passing and their neural substrates, in schizophrenia and in the general population. Then, in **Chapter 6**, we focus on schizophrenia and used bistable perception to probe the computational mechanisms underlying the positive dimension. We compared patients with prominent positive symptoms with matched healthy controls in two bistable perception tasks, already introduced in section 1. Our results suggest an enhanced amplification of sensory inputs in patients (in agreement with previous results), combined with an overestimation of the environmental volatility. Finally, in **Chapter 7**, we study drug-induced psychosis and suggest a multiscale account of psychedelics based on circular inference. We used in-silico simulations to show that descending loops (i.e., amplification of priors) can result in crossmodal, mainly prior-driven aberrant perceptual experiences, which constitute the hallmark of the psychedelic experience. We further propose a link between loops and neuromodulation, namely that dopamine regulates climbing loops while serotonin prevents the reverberation and amplification of priors (descending loops). Finally, we put forward a canonical microcircuit implementing (circular) belief propagation, ultimately linking our understanding at the macroscale, the mesoscale and the microscale.

# Part I

## Bistable perception in the general population

# Chapter 2

## Circular inference in bistable perception

Submitted for publication as:

Leptourgos P., Notredame CE., Eck M., Jardri R., Deneve S. (2018). Circular inference in bistable perception. *Under Review*

## Abstract

When facing ambiguous images, the brain switches between mutually exclusive interpretations, a phenomenon known as bistable perception. Despite years of research, there is no consensus on whether bistability is driven primarily by bottom-up or top-down mechanisms. Here, we adopted a Bayesian approach in an effort to reconcile these two theories. Fifty-five healthy participants were exposed to an adaptation of the Necker cube paradigm, in which we manipulated sensory evidence and prior knowledge. We found that manipulations of both sensory evidence and priors significantly affected the way participants perceived the Necker cube. However, we observed an interaction between the effect of the cue and the effect of the instructions, a finding incompatible with Bayes-optimal integration. In contrast, the data were well predicted by a circular inference model. In this model, ambiguous sensory evidence is systematically biased in the direction of current expectations, ultimately resulting in a bistable percept.

## Introduction

Perception can be defined as the process of combining available information to create valid and useful interpretations of the world. Although our phenomenological experience makes us think that perceptual decisions are trivial, the truth might be very different. An interesting example is visual perception of depth. When we see an object, our brain must reconstruct its 3D shape from a 2D retinal image; in other words, the brain must solve an inference problem [1]. Unfortunately, such problems are ill-posed, as in most cases the 2D retinal projection is compatible with many different 3D objects [2]. To cope with perceptual uncertainty, the brain must combine ambiguous information received by peripheral sensors (e.g., disparity cues and movement cues) with pre-existing information (either hard-wired or learned) concerning properties of the environment or the potential cost of a wrong decision [3,4]. Such combinations can be expressed through Bayes' theorem, in which prior probability distributions and sensory likelihoods are multiplied, resulting in a posterior probability distribution over possible solutions to the perceptual problem. Most of the time, only a single dominant (most probable) interpretation will emerge from these constraints.

However, when the level of ambiguity is too high, finding a single interpretation is not possible. Strikingly, ambiguous figures compatible with more than one plausible interpretation [5,6] lead to *bistable* (or more generally *multistable*) perception [7]. When facing those figures, the perceptual system is unable to commit to a single stable interpretation and instead oscillates between mutually exclusive interpretations every few seconds. A famous figure known to induce bistability is the *Necker cube* (NC) ([5]; **Figure 1a**), in which a 2D collection of lines is automatically interpreted as a 3D cube, which is either "seen from above" (SFA interpretation) or "seen from below" (SFB interpretation). Interestingly, the NC is an asymmetrical stimulus, meaning that it generates an implicit preference for the SFA interpretation (i.e., the general preference of humans to interpret things as if they were below the level of their eyes) [3,8].

While the concept of perception as inference under uncertainty offers a principled way to explain the efficiency of perceptual systems and certain perceptual illusions, it can account for bistable perception less directly. Indeed, if the brain uses explicit representations of uncertainty, e.g., a probability distribution instead of a point estimate [9-12], ambiguous stimuli should be recognized as such and not generate a unique, persistent representation. We note that bistable perception is far from unique in that case. Although many studies have reported that the brain is able to reach Bayes-optimal decisions [13-16], there are numerous tasks in which human behavior deviates significantly from that of a Bayesian observer [17-20].

Deviations from Bayesian optimality could be the consequence of highly non-linear and state-dependent interactions between feedback and feedforward streams of information in brain circuits [21]. Some of these effects can be quantified by the *circular inference* framework [22]. According to this framework, hierarchical processing in the brain is analogous to the propagation of probabilistic messages (beliefs) in a hierarchical model of the world [23]. The combination of feedforward and feedback inputs is equivalent to the product of prior and likelihood in Bayes' theorem. However, because neural circuits are highly recurrent, sensory evidence and prior information can easily reverberate and be artificially amplified through feedforward/feedback loops in the brain, resulting in the corruption of sensory evidence by prior information and *vice versa*. Such reverberation can be avoided if excitation (E) and inhibition (I) are perfectly balanced in cortical circuits [22], a well-known property of the healthy brain [24,25].

Recently, our team hypothesized a link between E/I imbalance in schizophrenia and the occurrence of psychotic symptoms (hallucinations and delusions). This hypothesis was recently reinforced by experimental evidence in a probabilistic reasoning task [26]. Interestingly, we also detected a certain amount of circularity in healthy participants, particularly the corruption of sensory evidence by prior information. If circular inference is a more general mechanism than initially predicted, an interesting question arises: is it possible to find evidence of circularity [27] in the perceptual behavior of healthy subjects in the absence of any psychotic experience? Here, we propose that bistability could be an example of percepts induced by such circularity.

To investigate this theory, we induced bistability in healthy participants using the NC. We asked how different pieces of information, including (a) pre-existing priors (i.e., the SFA preference), (b) newly acquired priors (i.e., instructions), and (c) visual cues, are combined to generate the percept. We compared different Bayesian and circular inference (CI) models for their ability to fit the data. We particularly sought to understand whether circularity and aberrant correlations between priors and sensory evidence significantly contribute to the way we perceive the world.

## Results

To determine the effects of prior knowledge and sensory evidence in an ambiguous perceptual context, 55 participants were exposed to continuous presentation of a NC. The dominant percept was discontinuously sampled according to the procedure presented in **Figure**

2 and was analyzed in terms of *relative predominance* (RP). RP corresponds to the overall probability of perceiving the SFA or SFB interpretation. A value of 1 or 0 would correspond to the SFA or SFB interpretation, respectively, fully dominating perception. A value of 0.5 would characterize a purely chance level wherein the 2 percepts are equiprobable.

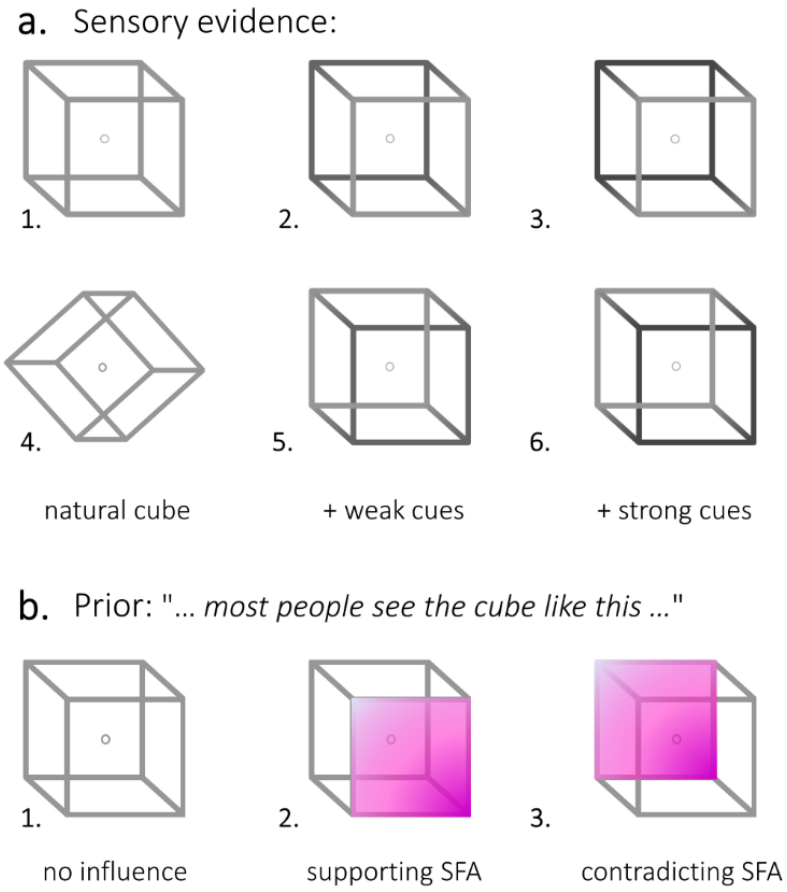


***Figure 1: Stimuli and instructions***. ***(a)*** *Different Necker cubes were used to induce bistable perception, in which the 2D figure is perceived as a 3D cube with either the left or the right side closer to the observer. Even in the case of the completely ambiguous stimulus (**1**), people have an implicit preference to interpret the cube as seen from above (SFA interpretation), which was interpreted an implicit prior. This prior can be annihilated by tilting the stimulus (**4**). Sensory evidence was manipulated by adding visual cues in the form of contrasts (**2-3,5-6**). The contrast could be strong (**3,6**) or weak (**2,5**) and could support (**2,3**) or contradict (**5,6**) the implicit prior.* ***(b)*** *A further manipulation of the prior was achieved by giving correct or wrong information to the participants about which interpretation was generally stronger (explicit prior). Instructions could support or contradict the implicit prior. An additional control group received no particular instructions. Crucially, to avoid additional priming effects, all groups received the same visual instructions (including the stimulus and the 2 possible interpretations), and the differences were only verbal. Note that the color used in the present figure has only been added for illustration purposes; during the experiment, participants were presented with full cubes.*

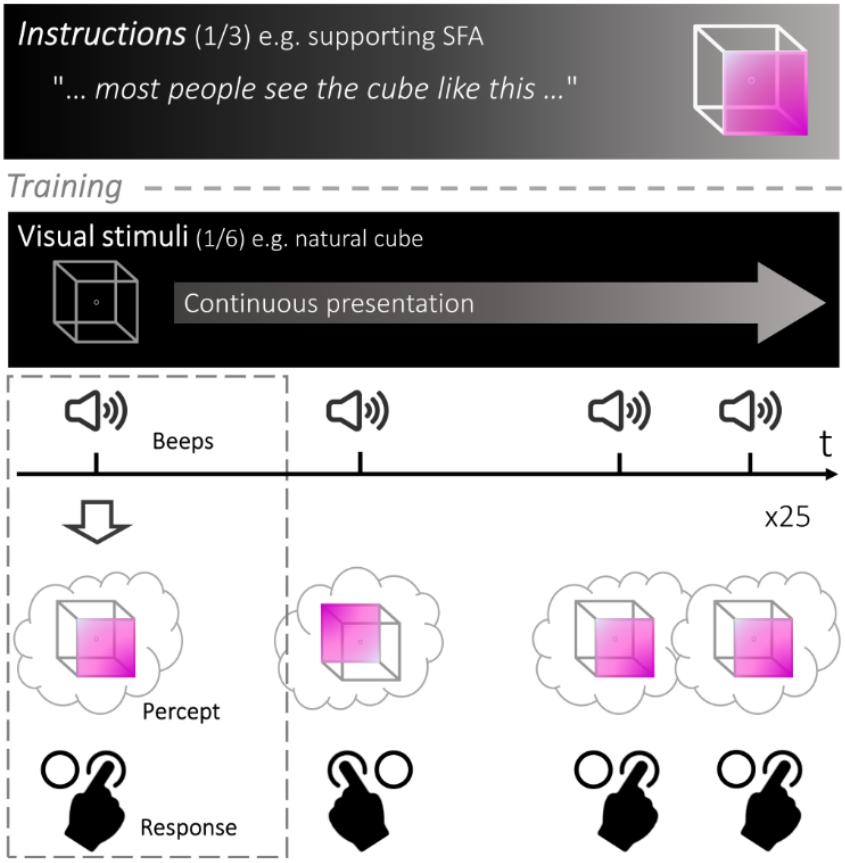


***Figure 2: Experimental design****. The task was inspired by [28]. Instructions were given at the beginning of the experiment (each participant received one set of instructions, creating a between-subjects design) and were followed by a short training phase to familiarize participants with the stimulus and the switches. During each run, one version of the cube was continuously presented to the participants, who were asked to discontinuously report their dominant percept by pressing a button every time a sound was heard. Each run consisted of 25 sound-trials (mean inter-sound-interval = 1.5 s). The main experiment consisted of 30 runs separated into 6 blocks of 5 runs each. In each block, a different variant of the stimulus was used. The first and fourth blocks always contained the ambiguous cube. The four cue conditions were randomly assigned to the four remaining blocks.*

Sensory evidence was manipulated by casting the cube into shadow in such a way that it either contradicted or supported the SFA implicit prior (see stimuli, **Figure 1a (2-3,5-6)**). Visual cues were either strong or weak so that the analysis could reconstitute a cue pseudo-continuum from *strongly contradicting* to *strongly supporting*. In the completely ambiguous condition, no difference existed in the color of the two sides of the cube (see stimuli, **Figure 1a (1)**).

Prior knowledge was manipulated by randomly allocating participants to 4 groups. The first group was exposed to a tilted cube, which was expected to neutralize the SFA implicit bias (**Figure 1a (4)**). The remaining 3 groups viewed a normal cube but received different explicit

instructions that either "supported", "contradicted", or were "neutral" with respect to the SFA bias (**Figure 1b**).

***Table 1: Demographic characteristics of the 4 groups (without outliers)**. The 4 groups did not differ in terms of age, education or sex. ◊: F-test, °: Chi-squared test*

| Variables | Tilted (n = 12) | Instr. Supp. (n = 14) | Instr. Contr. (n = 14) | No Instr. (n = 15) | Comparison | |
|---|---|---|---|---|---|---|
| | | | | | Test | P |
| **Age** mean (sd) | 23.33 (2.77) | 28.64 (7.19) | 28.93 (9.60) | 29.27 (11.73) | 1.31◊ | 0.28 |
| **Education** mean (sd) | 17.25 (2.42) | 19.07 (1.94) | 18.57 (2.17) | 18.00 (1.96) | 1.77◊ | 0.16 |
| **Sex ratio** (m:f) | 3:9 | 7:7 | 8:6 | 9:6 | 3.87° | 0.28 |

## Model-free analysis

The effects of prior knowledge and sensory evidence manipulation are presented in **Figure 3**. RP was not significantly different between the 2 ambiguous blocks (runs 1-5 and 16-20) in any of the groups ($\mathbf{p > 0.1}$), indicating only minor effects of fatigue (at least until the 20th run) and a stable effect of the instructions. Manipulation of sensory evidence significantly impacted bistability, with RP increasing as the visual cue changed from strongly contradicting to strongly supporting ($\mathbf{\beta = 0.415,\ p < 0.001}$). Manipulation of prior knowledge through instructions only affected RP in the case of contradicting instructions, with a significant overall reduction in RP ($\mathbf{\beta = -0.096,\ p < 0.001}$). Tilting the cube in the absence of any instruction resulted in a significant decrease in RP ($\mathbf{\beta = 0.103,\ p < 0.001}$), which substantiated the effect of an implicit prior that naturally biases perception toward SFA dominance. Importantly, we found a significant interaction between the continuous effect of cue and the effect of contradicting instructions (compared to the normal cube with supporting instructions and the tilted cube with no instructions; $\mathbf{\beta = 0.265,\ p = 0.016}$ and $\mathbf{\beta = 0.265,\ p = 0.021}$, respectively). Note that this interaction should not be present for a purely Bayesian observer, since the contribution of sensory evidence and priors (when expressed as the log odds ratio) should be additive.

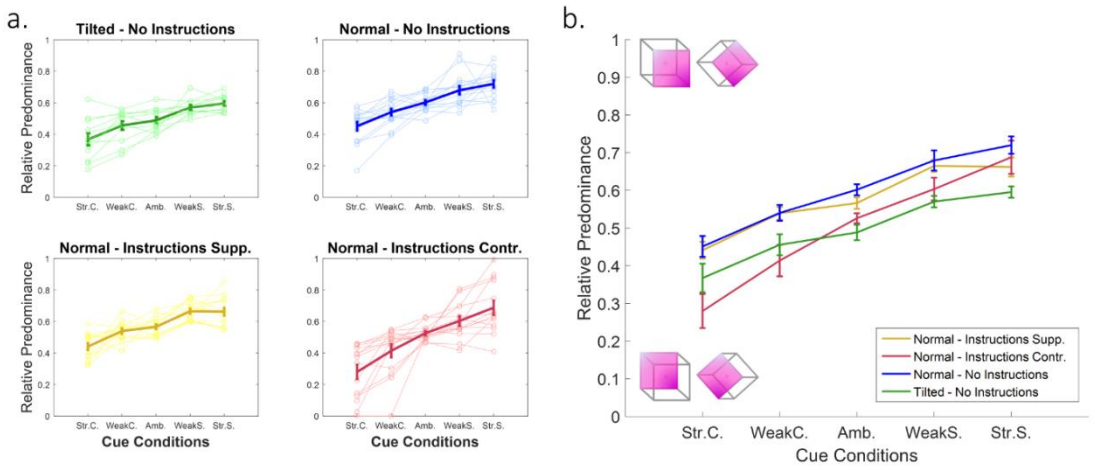


***Figure 3: Relative predominance between conditions***. ***(a)*** *The four subplots illustrate the four different prior conditions: tilted cube (top left, green; N=12) or normal cube with no instruction (top right, blue; N=15), supporting instructions (bottom left, yellow; N=14) or contradictory instructions (bottom right, red; N=14). The x-axis presents the 5 cue conditions, ranging from strong cue supporting the SFB interpretation (left) to strong cue supporting the SFA interpretation (right). Thin lines correspond to the behavior of single participants (outliers are not presented), and thick lines represent the average RP for each group, after removing the outliers (±SE).* ***(b)*** *Between-groups comparison of average RP. A linear mixed-effects model revealed significant effects of sensory evidence (**p < 0.001**) as well as the prior (contradictory instructions, **p < 0.001**) and tilt (**p < 0.001**)) manipulations. We also observed a cue x instruction interaction for the contradictory instructions (red curve) compared to supporting instructions (yellow curve, **p = 0.016**) and the tilted cube (green curve, **p = 0.021**).*

## Model-based analysis

To test our hypothesis that circularity shapes bistable perception, we fitted a CI model to the average data, similar to the one introduced by Jardri and colleagues [26]. This model assumes that participants perform approximate inference due to the reverberation of sensory evidence and priors in the hierarchy as a result of unbalanced inhibitory control (**Figure 4a, right panel**). Furthermore, we compared the performance of our CI model against that of 2 Bayesian models performing exact inference: first, a naïve Bayes (NB) model, which is identical to the multiplicative rule of Bayes' theorem (**Figure 4a, left panel**), and second, a weighted Bayes (WB) model in which different levels of trust (weights) could be assigned to sensory cues and priors. The WB model was equivalent to a NB model in which all the weights were set to 1 and equivalent to a CI model without any reverberating messages (**Figure 4a, middle panel**). The NB, WB and CI models can thus be considered 3 versions of the same model with an

increasing number of parameters being fitted to the data. Predictions for the 3 models are presented in **Figures 4b** and **4c**.

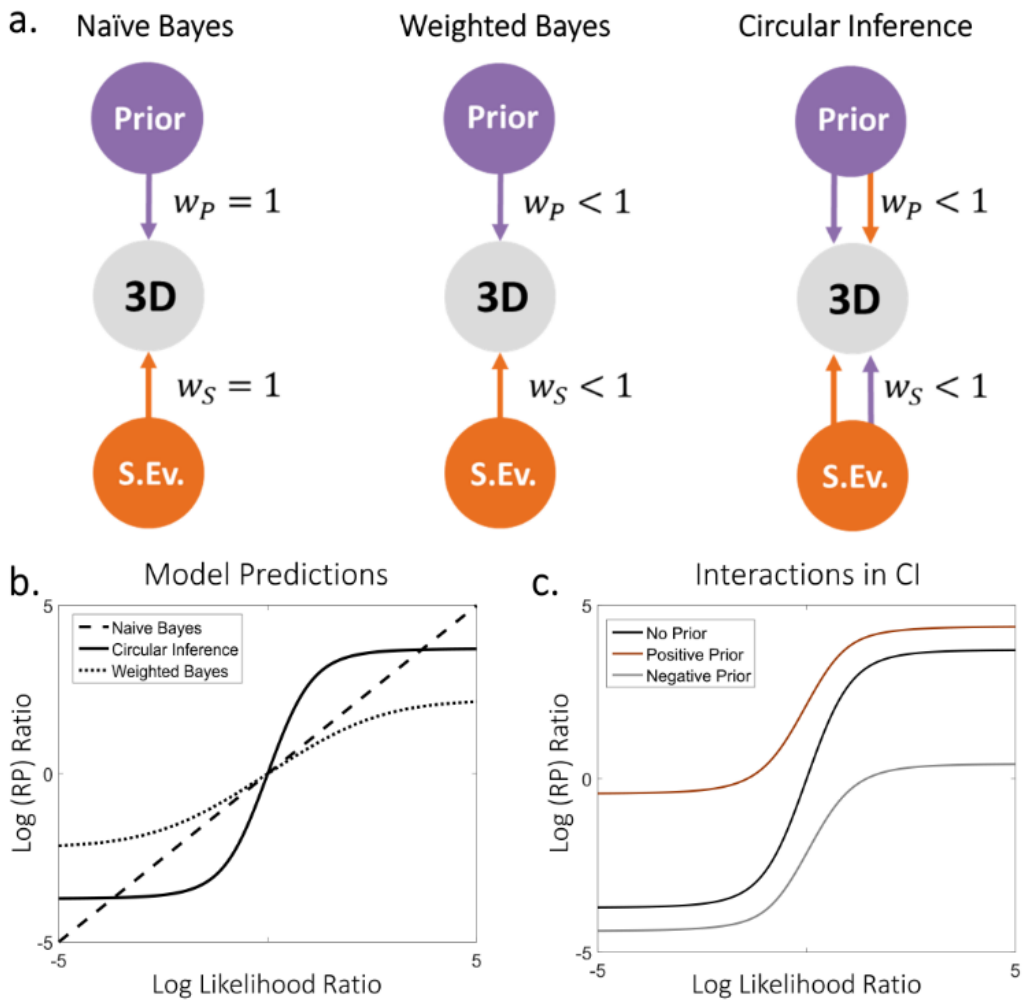


***Figure 4: Illustration of models and model predictions***. ***(a)*** *Three different models were used to fit the data. The simplest model (naïve Bayes (NB),* ***left panel****) consisted of a simple addition of the sensory evidence and prior on the log scale and is equivalent to a three-layer generative model in which all the connections are equal to 1. The weighted Bayes (WB) model (****middle panel****) further assumes that there is only partial trust between the nodes of the generative model. Importantly, both the NB and WB models perform exact inference. Finally, we used a circular inference (CI) model (****right panel****) that further allows reverberation and overcounting of sensory evidence and prior knowledge.* ***(b)*** *Log(RP) ratio predicted by the models as a function of the log-likelihood ratio. The NB model predicts a linear dependence, whereas both the WB and CI models predict sigmoid curves (due to the saturation imposed by the weights). Furthermore, the 3 models make different predictions about the slope of the curves around zero. The NB and WB models predict a slope of 1 and less than 1, respectively, and only the CI model predicts a slope greater than 1.* ***(c)*** *In the CI model, the slope of the log-likelihood/log-posterior curve also depends on the log-prior as a result of the reverberations, indicating an interaction between the two different types of information [27]. Weaker priors are associated with steeper sigmoid curves.*

**Figure 5** illustrates the best-fitting NB (**5a**), WB (**5b**) and CI models (**5c**). **Figure 6** presents the values of the free parameters in the 3 models. The 3 models predict very different values for likelihoods and priors. These differences can be easily explained by the NB model assuming perfect trust in sensory evidence and priors, whereas the other 2 models predict much lower weights ($w_S = 0.77, w_P = 0.59$ for the WB model and $w_S = 0.66, w_P = 0.59$ for the CI model).

The NB model captures most trends in the data qualitatively, with the following exceptions. First, it underestimates RP in the case of the normal cube without instructions (**Figure 5a, blue curve**), and second, it is unable to predict the correct slopes. The latter limitation is especially striking in the case of a normal cube with contradicting instructions, where the slope is larger than predicted (i.e., larger than 1; **Figure 5a, red curve**). The WB model performs better than the NB model in most conditions, but it also underestimates the effect of the cue when the instruction contradicts the SFA preference (see **Figure 5b, red curve**). In contrast, the CI model captures this last trend (see **Figure 5c**), suggesting that the variability of the cue effect (the slope) in different conditions is due to circularity in the inference process.

A quantitative comparison of the 3 models using BIC scores, which penalizes the use of extra free parameters in the WB and CI models, indicated that the CI model significantly outperformed the 2 Bayesian models (BIC scores for NB = -242.65, for WB = -240.77, and for CI = -249.49). A lower BIC score indicates that the model better fits the data, with a difference larger than 2 considered positive and a difference larger than 6 considered strong ($\delta_{BIC}$ = 6.84 for comparison of the CI and NB models and $\delta_{BIC}$ = 8.72 for comparison of the CI and WB models).

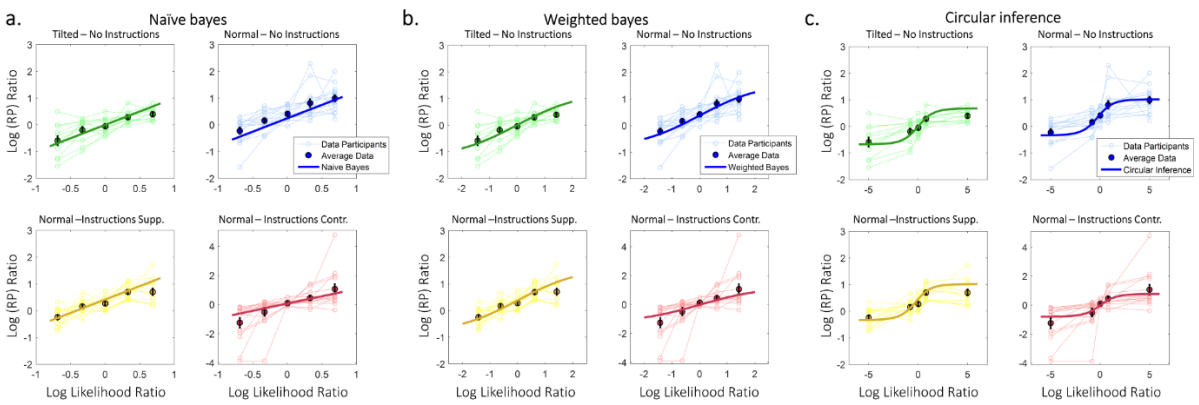


***Figure 5: Observed and predicted log(RP) ratios as a function of the log-likelihood ratio***. *Different colors correspond to different prior conditions. Thin lines represent single participants' data, highlighted*

*points correspond to average RP (±SE), and thick lines illustrate model predictions. The three models are presented separately, since likelihood was itself considered a free parameter [**(a)**: NB, **(b)**: WB, **(c)**: CI]. The models were fitted to aggregated data from all participants by minimizing the mean squared distance between the observed and predicted log(RP) ratios.*

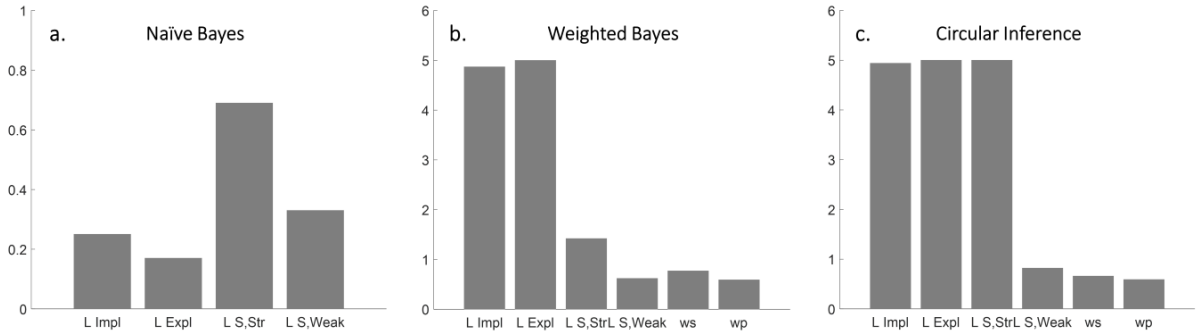


***Figure 6: Optimal values of free parameters for the three models*** *[**(a)**: NB, **(b)**: WB, **(c)**: CI]. The NB model had fewer free parameters than the other 2 models, since the two weights were by definition fixed to 1. We observed important differences in the values of the likelihoods ($L_{S,Str}, L_{S,Weak}$) as well as in the values of the priors ($L_{impl}, L_{expl}$) between the NB model, on one hand, and the WB and CI models, on the other hand. These differences were mainly due to different values for the weights ($w_S, w_P$).*

## Discussion

The goal of the current study was to decipher how priors and sensory evidence are combined to shape bistable perception. We particularly wished to investigate whether such integration is probabilistically optimal or if other principles are at play, contributing to the debate on whether bistable perception is a by-product of perceptual inference (regardless of its neural implementation). Our results suggest an imperfect neural implementation of probabilistic inference, possibly due to an imbalance between excitation and inhibition in neural circuits.

As previously reported, we found an asymmetry in the way participants interpreted the completely ambiguous NC [8]. This finding supports the notion of an implicit preference (implicit prior) to perceive objects in an SFA configuration [3]. More surprisingly, we showed that this preference could be explicitly manipulated by giving information that either confirmed or rejected it (explicit prior). In agreement with previous studies [28-30], adding visual cues also significantly biased perception toward the corresponding interpretation. The qualitative effects of implicit priors, explicit priors and sensory evidence appeared compatible with a probabilistic combination of information, suggesting that Bayesian inference was still at work.

However, we also found a significant interaction between priors and sensory evidence that could not be explained by exact inference. In particular, the effect of sensory cues was

stronger when the prior was more ambiguous (e.g., when the implicit preference for SFA was contradicted by instructions) and weaker in the absence of a prior (e.g., a tilted cube). In contrast, Bayes' theorem predicts that sensory cues are weighted according to their reliability, independently of the prior. Through parametric model comparison, we found that the present data could be better accounted for by a CI model, in which prior beliefs (i.e., the instructions and SFA preferences) corrupt new sensory evidence (i.e., ambiguous cues are misinterpreted as supporting the current belief) and *vice versa*. This corruption could be the result of feedback to sensory areas insufficiently controlled by inhibition [22]. Such feedback could also cause multistable perception (i.e. generate a bistable attractor; see **Supplementary Figure S5**) by temporarily stabilizing the current percept despite the absence of supporting evidence [27].

These findings add new elements to a long-lasting debate in neuroscience that questions whether perception is mostly driven by bottom-up processes, or whether top-down effects are equally important [21]. Multiple studies have investigated how low- or high-level manipulations affect bistability, without offering definitive answers. For the former, authors have used priming or suppressing effects (usually attributed to adaptation) [31-34], changes in retinal location [35], manipulation of the type of presentation (continuous–intermittent) [36,37], and direct manipulation of the properties of the stimulus, like intensity [38] and completeness [39]. In contrast, studies of high-level manipulations have focused on the effects of volition [40,41], expectation and prediction [42], attention [43-45], learning [46], mental imagery [47], knowledge of reversibility [48] and finally the preference for stimuli with a statistical structure similar to that of natural images [8,49,50]. Note however that the present study was not designed to test specific neural mechanisms such as adaptation and noise.

Consistent with the present study, some authors have focused on how these various effects are combined [51-53]. Moreno-Bote et al showed that cue combination in a bistable display can be well explained by a multiplicative law (their predictions are similar to the NB model described here) [54], whereas Zhang and colleagues demonstrated that different types of priors are effectively combined [4]. Here, we have gone a step further and investigated how top-down (prior manipulation) and bottom-up (sensory cues) effects interact. Rather than inducing a prior through learning, as is widely done in the literature [46,47], we directly manipulated participants' expectations. This manipulation assumes that instructions can generate a high-level prior affecting perceptual processing in a way similar to a learned prior (as in [55]).

Despite the amount of available data and the apparent simplicity of the problem, very few studies in the literature have applied normative explanations for bistable perception that

include data-fitting [54]. Although proposing a complete model of bistable perception based on circular inference goes beyond the scope of the paper, our present results suggest that a local message passing algorithm with the addition of information loops could constitute the basic principle of such a normative model. Some alternative normative models have relied on a simplified form of Markov Monte-Carlo sampling. More precisely, they assumed the current percept is based on taking one sample from the posterior distribution and using this sample as a prior for the next time step [56,57]. However, Markov Monte-Carlo sampling requires very long sampling times (because of temporal correlation between samples) to perform accurate inference. A possible argument in favor of circular inference would be that it can reach correct conclusions quickly and accurately in most perceptual tasks, except in particularly ambiguous cases [22], making circular inference a powerful model for perceptual inference in unambiguous cases.

From a methodological point of view, and in contrast to most studies on bistable perception, in which participants continuously report the dominant percept with a sustained button-press [58,59], we asked participants to respond discontinuously, after being exposed to a go-signal [29]. This procedure has two main advantages. First, it minimizes the role of attention. Indeed, it has been shown that attention plays a crucial role in bistable perception, especially for certain bistable stimuli [41,60]. The inability to control for differences in attentional load between participants could be an important source of uncertainty and even partly explain the huge variability usually observed in some publications (see [29]). Second, this procedure is less affected by differences in reaction time, as one could use the time of the sound as a proxy for the time of the decision. As a consequence, discrete sampling not only seems ideal for a rigorous experimental exploration of bistable perception but is also useful for adapting this task to specific clinical populations with well-known attentional and motor problems.

Finally, some limitations need to be acknowledged. First, because of the type of priors used (instructions), we were obligated to use a between-subjects design, which prevented us from comparing the effects of different instructions in the same participant. As a result, there were only 5 conditions per participant, and we could only fit our models to averaged data, ignoring variability between participants (see also [15,54]). Second, all the models under consideration were based on an assumption of temporal independence between the percepts at the time of the sounds. This assumption can be partly justified by the weak autocorrelation of the averaged data (see **Supplementary Figure S6**), although these autocorrelations may be stronger in individual participants [56]. Nevertheless, temporal statistics would not affect the qualitative predictions of the models [54]. In particular, temporal statistics without circular

inference would not provide a valid alternative to the present findings, including the slopes and the cue × instruction interaction. Third, a response bias could partially account for instruction effects (explicit priors). However, a response bias would have a similar effect over responses across different cue conditions while leaving perceptual processing completely unaffected. Although the above is a possible interpretation of the data, it remains highly improbable, given the non-linear interaction found between instructions and visual cues (see also **Supplementary Figure S7** for additional arguments).

Overall, this study confirms that circular inference can be observed to a certain degree in healthy individuals. This unprecedented observation opens a range of crucial questions that suggest opportunities for further research: in what other ways could circularity affect cognition, and what are its neural substrates? Crucially, we must determine under what circumstances circular inference generates aberrant beliefs or percepts, such as those observed in pathological (neurological or psychiatric) contexts.

## Methods

Participants were healthy volunteers meeting the following inclusion criteria: age > 18 years, provision of informed consent, normal or corrected-to-normal near visual acuity, no past or current medical history of neurological or psychiatric disorders, and no current or recent use of psychotropic medication or toxic drugs. Near visual acuity was quantified using the Parinaud score; we considered values less than or equal to 2 to be normal. Of the 65 participants initially recruited, 10 were excluded because of outlying mean RP values (with cutoffs set at $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, where $Q1$, $Q3$ are the lower and upper quartiles respectively and $IQR$ is the interquartile range). We highlight that 7 of the 10 excluded participants also exhibited qualitatively bizarre behavior (such as opposite effects of visual cues), indicating a misunderstanding of the instructions or low attention levels.

### Experimental setting and procedure.

The general procedure (**Figure 2**) was inspired by Mamassian and Goutcher's protocol [29] and consisted of 6 blocks of 5 consecutive runs. During each run, a 200 x 200 pixel NC displayed in the middle of a black screen was continuously presented to the participants. Using a forced-choice method, we asked participants to report their ongoing interpretation as soon as they heard a warning sound, which occurred 25 times in a pseudo-regular manner (mean inter-

sound interval = 1.5 s, uniformly distributed between 1 and 2 s). Each response corresponded to a trial, providing a discontinuous sampling of the task's perceptual dynamics. Runs were separated from one another by a black screen with a duration of 10 s to minimize between-run influences. The experiment was also interspersed with 5 between-block breaks of non-predefined duration. Prior to the experiment, participants were informed that they would be presented with empty cubes, the 2 possible interpretations of which were explicitly mentioned. The basic instruction was to passively view these cubes without trying to constrain perception.

We manipulated sensory evidence either by making the cubes homogeneously gray (i.e., perfectly ambiguous) or cuing them by shadows (**Figure 1a (1-3,5-6)**). This additional depth information was intended to bias perception toward one interpretation or the other. It was specified by two parameters. First, its orientation was defined in relation to the implicit prior. A shadow falling on the top left corner was expected to emphasize the SFA preference and thus was classified as a supporting cue. Conversely, a shadow that fell on the bottom right corner was characterized as a contradictory cue, as it went against implicit bias. Second, the strength of the cue (which can also be conceived in terms of the amount of sensory information) was controlled by the shadowing contrast level. Weak and strong cues corresponded to 20% and 30% contrast, respectively. The 1$^{st}$ and 4$^{th}$ blocks always consisted of presentation of an ambiguous cube. The other blocks were randomly allocated a different type of cue, defined by the 2 x 2 factorial combination of 2 possible orientations (contradicting or supporting) and 2 possible strengths (weak or strong).

Participants were separated into 4 groups (n = 12, 14, 14, and 15) that differed in terms of how we altered their prior knowledge. The first group was presented with a tilted cube, which was expected to neutralize the SFA implicit bias (**Figure 1a (4)**). The remaining 3 groups viewed a normal cube—where the implicit prior is deemed present—but received different types of instructions, which we used to manipulate their implicit prior. In Group 2, instructions explicitly mentioned the presence of the implicit bias:

> *"When looking at the cube, most people tend to see it with its front side on the right. Differently said, there is a natural tendency to see the cube mostly "from above". In the present experiment, we aim to study the characteristics of this spontaneous preference."*

Because the statement was correct, the instructions were considered to support the spontaneous bias (supporting instructions). In Group 3, participants were informed about a natural tendency to see the cube primarily as though it were seen from below. The wording was similar, but the statement was incorrect, thus contradicting the implicit prior (contradictory instructions). In Group 4, the participants received no complementary information. In this case, their prior knowledge could be considered akin to the implicit bias (neutral instructions). Note that, to avoid any additional priming effects, the difference among the 4 groups was only verbal, while all groups received the same visual instructions, including the stimulus and the 2 possible interpretations. As shown in Table 1, the 4 groups did not significantly differ in terms of demographic characteristics.

To neutralize the potential confounding bias of eye-movements, participants were additionally instructed to gaze at a fixation point in the middle of the screen. A training session allowed each participant to familiarize himself/herself with the stimuli and the apparatus.

The experiments were implemented in MATLAB v. 2011b (MathWorks, Natich, MA), using Psychtoolbox v. 3.0.10. Stimuli were displayed on a 17-inch LED screen with a resolution of 1280 x 1024 pixels. Responses were collected using a classical computer keyboard. A chin-cup and forehead bar ensured immobilization of the participant's head at a distance of 60 cm between the eye and the screen.

## Model-free analysis

### *Measured Variable*

RP was calculated by taking the grand mean of responses across trials, runs and participants. It can be interpreted as the general probability to perceive one interpretation or the other on each trial.

### *Statistical analysis*

Because RP is a ranged variable, we performed exclusively non-parametric analyses. The effects of priors, sensory evidence, and their interaction were tested using a linear mixed-effects model comprising the effects of cues and instructions as well as their interaction as fixed effects, together with Gaussian random effects for intercepts and slopes. For significant omnibus effects, we performed post hoc comparisons using either paired or unpaired rank-sum tests to clarify

simple effects in the 2 x 2 design. Finally, one-sample Wilcoxon signed rank tests were performed to compare the mean RP with 0.5, i.e., chance level. All significance tests were performed on the sample of the 55 participants (12, 14, 14 and 15 for each group respectively), they were two-tailed and used an alpha value of 0.05 in the statistical toolbox of Matlab v. 2011b (MathWorks, Natich, MA).

## Model-based analysis

*Models*

We conceptualized perception as an inferential process, in which the brain generates a subjective belief about the possible interpretations of the NC (i.e., a posterior probability) and uses it to make a perceptual decision, particularly whether it is an SFA or SFB cube. Three different models were fitted to the average RPs of the 4 groups. All the models assumed independence between the sequential perceptual decisions within a run. They differed in how the 3 main effects of the experiment (sensory evidence $S$, an implicit prior $P_{impl}$, and an explicit prior $P_{expl}$) were combined to give rise to the posterior probability $P(X|S, P_{impl}, P_{expl})$. In this expression, $X$ is a binary variable that corresponds to the 3D interpretation ($X = 1$ corresponds to SFA, $X = 0$ corresponds to SFB.

The simplest model that was fitted to the data is the NB model, which assumes perfect integration of likelihoods and priors according to the Bayes theorem. Consequently, it's equivalent to a basic multiplicative rule [54] (additive rule in the log scale) (eq. 1; **Figure 4a, left panel**). The WB model extended the NB model by assuming only partial trust to the sensory evidence and prior information (eq. 2; **Figure 4a, middle panel**). Crucially, both models are Bayesian models performing exact inference. Finally, the third model is a *circular inference model* [26] , meaning that information is not only weighted, as in the WB model, but it's also amplified, due to information loops (eq. 3; **Figure 4a, right panel**). As a result, the CI model is doing sub-optimal inference, which renders it qualitatively different from the other 2 models.

The 3 models are quantitatively described by the following equations:

$$L_{RP} = L_S + L_{impl} + L_{expl} \quad (1)$$

$$L_{RP} = F(L_S, w_S) + F(L_{impl} + L_{expl}, w_P) \quad (2)$$

$$L_{RP} = F(L_S + F(L_S, w_S) + F(L_{Pr}, w_P), w_S) + F(L_{Pr} + F(L_S, w_S) + F(L_{Pr}, w_P), w_P) \quad (3)$$

where $F(L, w)$ is a sigmoid function:

$$F(L, w) = log\left(\frac{we^L + 1 - w}{(1-w)e^L + w}\right) \quad (4)$$

and $L_{Pr} = L_{impl} + L_{expl}$. $L_{RP}$ corresponds to the log-ratio of the RP and is taken to be equal to the log-posterior ratio. That assumption is because we assume that perceptual decisions are made using probability matching, a commonly observed strategy in sequential 2AFC tasks [20,54,61]. We note that applying a SoftMax to the log posterior odds (a more appropriate model for perceptual decisions) would only induce a global change in the gain of the former and would not affect any of our conclusions.

$$L_{RP} = log\left(\frac{RP}{1 - RP}\right) \quad (5)$$

The log-likelihood ratio $L_s$, the implicit log-prior ratio $L_{impl}$ and the explicit log-prior ratio $L_{expl}$ are given by the following equations:

$$L_s = log\left(\frac{S}{1 - S}\right) \quad (6)$$

$$L_{impl} = log\left(\frac{P_{impl}}{1 - P_{impl}}\right) \quad (7)$$

$$L_{expl} = log\left(\frac{P_{expl}}{1 - P_{expl}}\right) \quad (8)$$

Because none of these variables was known, they were all treated as free parameters. To reduce as much as possible the total number of free parameters that needed to be optimized, we further

considered symmetry both in the effects of the cues and the instructions, resulting in 4 free parameters ($L_{s,strong}, L_{s,weak}, L_{impl}, L_{expl}$).

Finally, $w_S$and $w_P$ (appearing only in the WB and CI models) correspond to participants' trust (or weight) in the sensory evidence and priors, respectively, and constituted the 2 additional free parameters of those models:

$$w_S = P(X = 1|S = 1) = P(X = 0|S = 0) \quad (9)$$

$$w_P = P(X = 1|P = 1) = P(X = 0|P = 0) \quad (10)$$

Importantly, since the SFA prior was completely uninformative in the case of the titled cube, we considered the following:

$$w_P > 0.5 \; if \; Normal \; Cube, w_P = 0.5 \; if \; Tilted \; Cube \quad (11)$$

As a control, we also considered the case in which $w_P$ has the same value in all conditions (see **Supplementary Figure 1**).

An illustration of the different models is presented in **Figure 4a** The CI model (**Figure 4a, right panel**) hypothesizes that the perceptual system performs approximate inference due to unbalanced inhibitory control. Those impairments lead to a failure to remove efficiently redundant messages: a reverberating prior, which is misinterpreted as sensory evidence, re-ascends the hierarchy and corrupts the likelihood term and redundant sensory evidence, which descends the hierarchy and corrupts the prior term. Additionally, as in [26], a cross-term is added to each component, rendering likelihood and prior information completely inseparable. Because of those extra terms, the sensory evidence and prior components become aberrantly correlated, and consequently they generate an interaction (**Figure 3c**; [27]). Note that the WB model (**Figure 4a, middle panel**) can be derived from the CI model by removing the reverberated terms, while the NB model (**Figure 4a, left panel**) by further assuming: $w_S = w_P = 1$.

The CI model used here was similar to the model used by Jardri and colleagues to explain participants' behavior (both those suffering from schizophrenia and healthy participants) in a

probabilistic reasoning task [26]. Nevertheless, an important difference needs to be highlighted. In the present study, the redundant messages corrupted the original messages only once (there was still overcounting of information, but the amount of amplification stayed constrained), which is equivalent to setting $a_S$ and $a_P$ (the parameters in the original model that represented the number of times the redundant messages were taken into account) equal to 1. The reason was twofold. First, fixing the number of loops did not change the results qualitatively. Indeed, the resulting model predicted both a slope larger than 1 and an interaction between sensory evidence and priors, the two characteristic features of circular inference observed in the data. Second, the additional complexity (2 more free parameters) did not further improve the fit (see **Supplementary Figure S2**).

**Figure 4b** illustrates the predictions of the 3 models. Contrary to the linear NB model, both the WB model and the CI model are non-linear models, due to the saturation of the posterior that is caused by the weights. Importantly, the 3 models make different predictions about the slope of the log-likelihood/log-posterior curve around 0: the NB model and WB model predict a slope equal to and smaller than one, respectively. Interestingly, only the CI model can generate a slope that is larger than one, due to overcounting of the prior and of sensory evidence. Moreover, it predicts interaction between the prior and sensory evidence, such that the slope differs depending on prior strength and weight (**Figure 4c**).

Finally, in eq. 1-3, we assumed that instructions act as an additional prior term, essentially changing the strength of the implicit preference independently of the presence of a visual cue. As a result, any interaction between the effect of the cue and the effect of the instructions is forbidden under Bayesian formalisms and can only be explained by non-Bayesian mechanisms such as the presence of circular inference. It is worth noting though that alternative interpretations of the instructions (which are even more complex) might also generate such an interaction, notably likelihood-dependent instructions, or instructions that directly affect the reliability of the sensory evidence. Those additional models were also considered and compared to the CI model (see Supplementary Figures S3 and S4).

*Model fitting*

All the models were fitted to the data by minimizing the mean squared distance between the log(RP) ratio for the different conditions and the predictions of the models. Instead of simply considering the means, we used data points from each participant, making full use of the

available information but assuming that the parameters did not vary between participants. The optimal values for parameters were obtained using a non-linear programming method (sequential quadratic programming; a built-in MATLAB function), appropriate for non-linear constrained multivariable functions. To avoid local minima, the optimization process was repeated 100 times for each model, with initial values chosen each time randomly from the parameter space.

*Model comparison*

We compared the quality of the fits for the 3 models using BIC scores. We approximated the likelihoods of all the models as normally distributed. The BIC score can then be calculated by the following equation:

$$BIC = nlog(\sigma^2) + klog(n) \quad (12)$$

where $n$ is the total number of data points (5 points per participant), $\sigma^2$ is the mean squared error, and $k$ is the number of free parameters (4 for NB, 6 for the other models).

## Supplementary Material

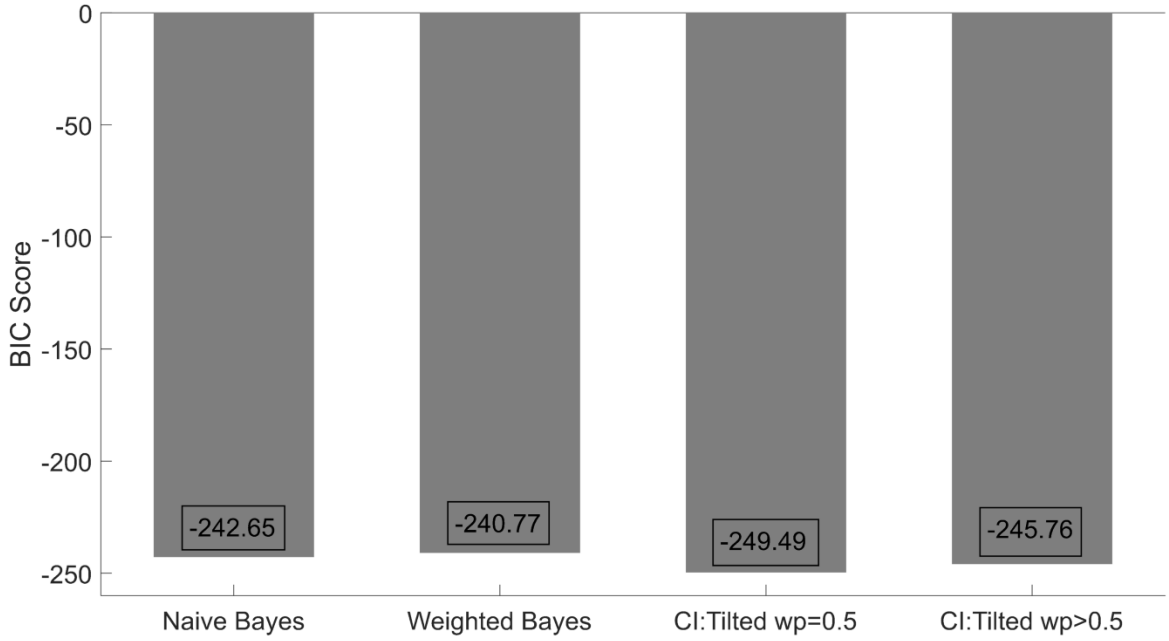


***Figure S1: Comparison of models assuming different effects of priors in the case of the tilted cube****. In the* ***main text****, we made the assumption that* $w_P = 0.5$ *when the cube is tilted, since in that case the prior becomes uninformative and thus irrelevant. An alternative would be that the prior weight remains larger than 0.5 (as in the case of the normal cube), and that only the log-prior ratio becomes equal to zero. The 2 alternatives cannot be differentiated in the cases of the NB and WB models, but they make different predictions in the case of the CI model, in which a reverberating likelihood term appears inside the prior term. This reverberating term disappears completely if we assume* $w_P = 0.5$*, whereas it remains if we assume* $L_P = 0$*. Formal comparison of the two models using the BIC score revealed that the former (*$w_P = 0.5$*) outperformed the latter (*$w_P > 0.5$*) (BIC scores: -249.49 vs. -245.76, respectively).*

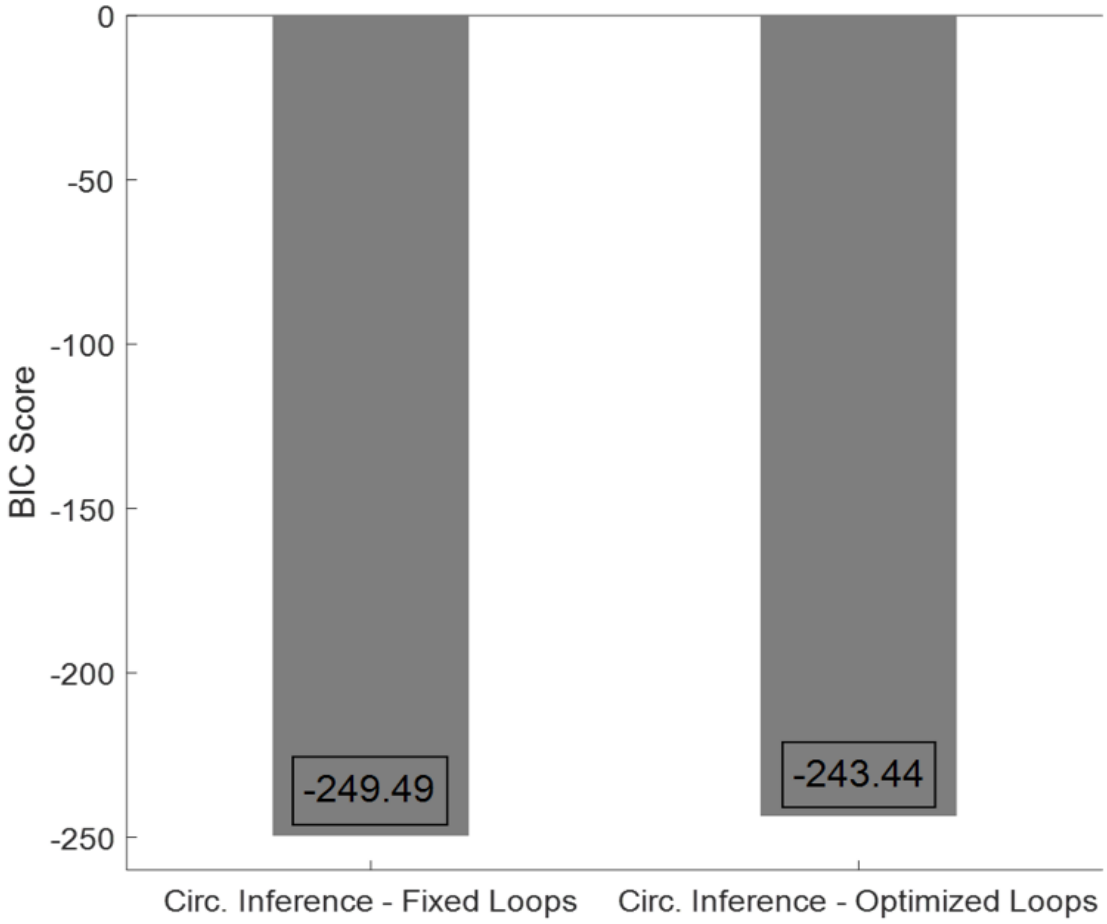


***Figure S2: Comparison of circular inference models with fixed and optimized loops****. In a previous study, Jardri and colleagues considered a CI model in which the strength of climbing loops (reverberation of sensory evidence, $a_S$) and the strength of descending loops (reverberation of priors, $a_P$) were considered free parameters and were optimize [26]. The predictions of the model were summarized by the following equation:* $L_C = F(L_S + F(a_c L_S, w_S) + F(a_d L_P, w_P), w_S) + F(L_P + F(a_c L_S, w_S) + F(a_d L_P, w_P), w_P)$*. In the current study, we fixed the values of these 2 extra parameters to 1, obtaining equation (1) (**Main Text**). The two models make the same qualitative predictions, as they both contain reverberating terms that render likelihood and prior inseparable. We quantitatively compared the two models using their BIC scores. We found that the simplified model (fixed loop strength) performed better than the full model (optimized loop strength) (BIC scores: -249.49 for the former vs. -243.44 for the latter).*

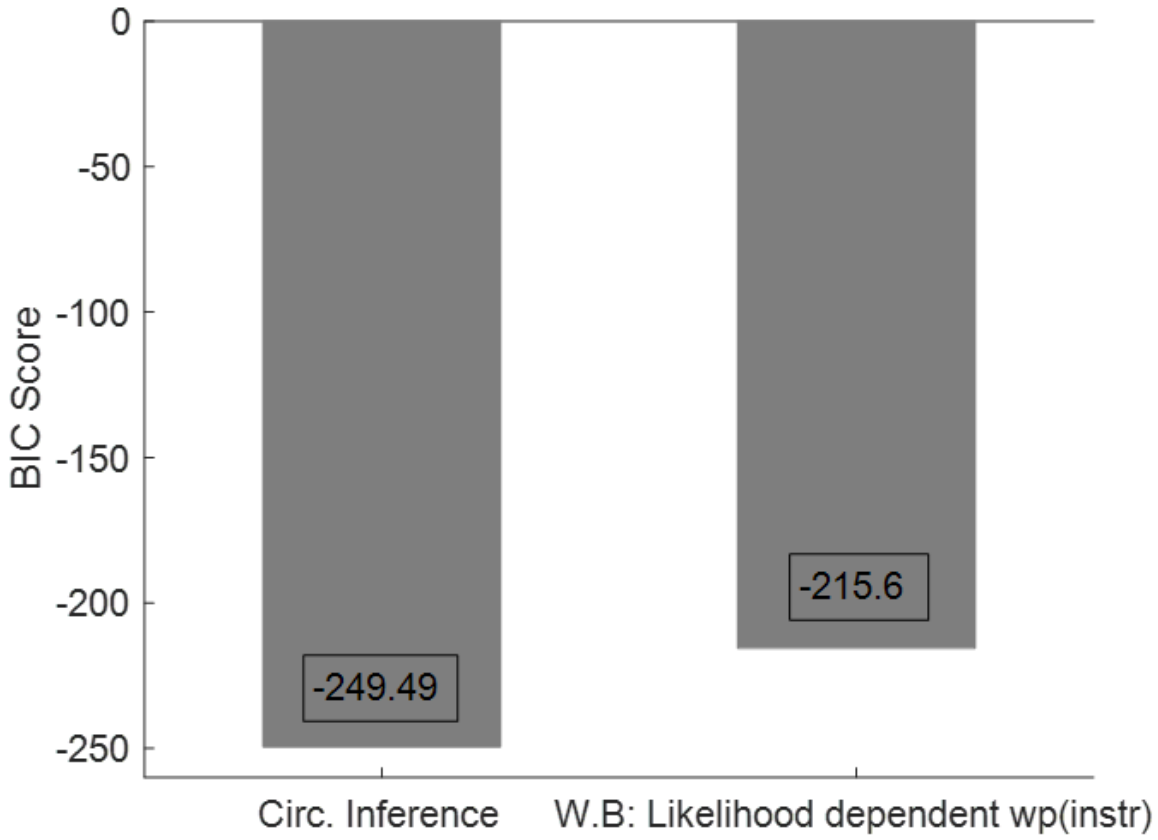


***Figure S3: Comparison of the CI model with a WB model in which the instructions are likelihood-dependent.*** *A likelihood-dependent effect of the instructions may constitute an alternative explanation for the interaction observed between the effect of the visual cues and that of the instructions/priors. In our framework, such an interpretation can be implemented as follows:* $L_{RP} = F(L_S, w_S) + F(L_{impl}, w_{P,impl}) + F(L_{expl}, w_{P,expl}(L_S))$*, where the weight attributed to the instruction is likelihood-dependent. Despite its plausibility, such an implementation drastically increases the complexity of the model, since it comprises 9 free parameters (instead of* $w_P$ *, we now have* $[w_{P,impl}, w_{P,expl}(L_{S,amb}), w_{P,expl}(L_{S,Str}), w_{P,expl}(L_{S,weak})]$*). A formal comparison of the CI model with this version of the WB model reveals a clear superiority of the former (BIC scores: -249.49 vs -215.6, respectively).*

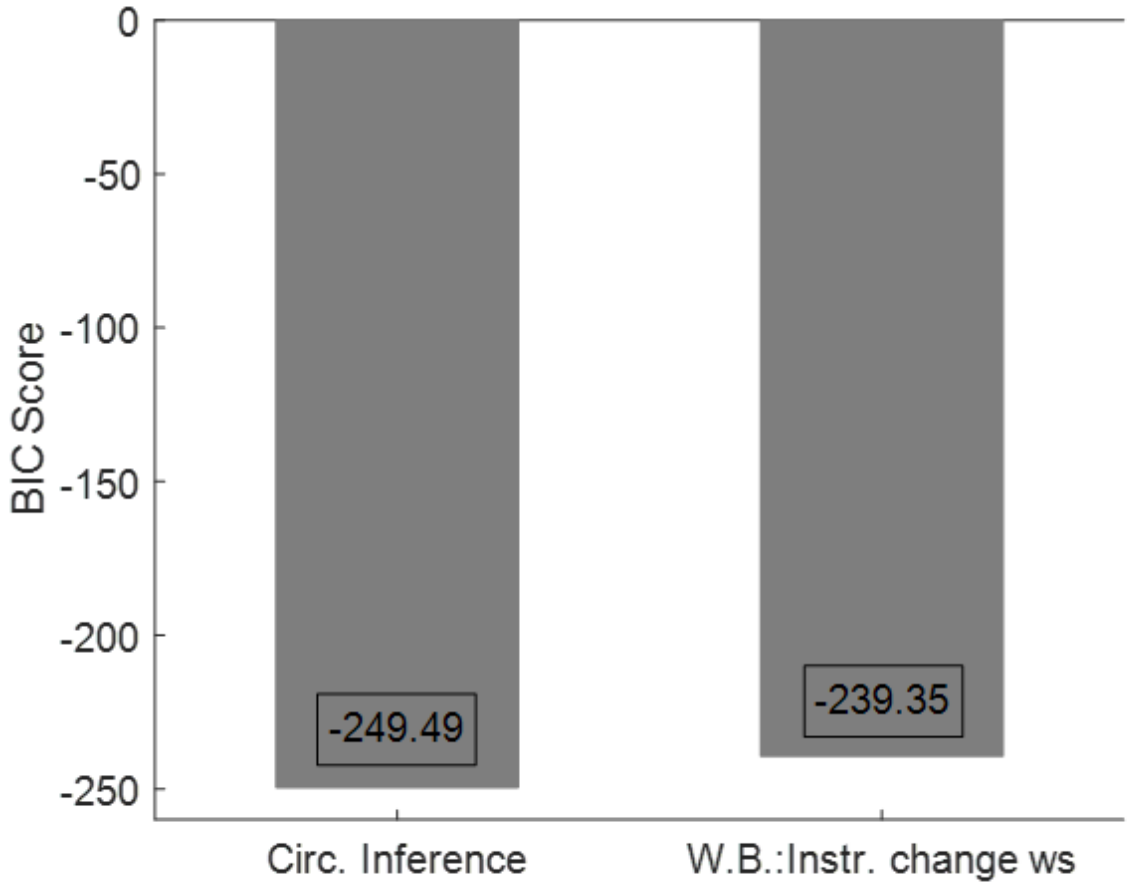


***Figure S4: Comparison of the CI model with a WB model in which the instructions directly affect the reliability of the visual cue.*** *The observed interaction between visual cues and priors could be explained by assuming that the instructions do not act as a prior but instead change the reliability of the sensory evidence. In our framework, such an interpretation could be implemented as follows:* $L_{RP} = F(L_S, w_S(Instr)) + F(L_{impl}, w_P)$*, where the sensory weight depends on the instructions. This model comprises 7 free parameters, and a formal comparison with the CI model reveals that circularity (plotted on the left) offers the best interpretation of the data (BIC scores: -249.49 vs -239.35, respectively).*

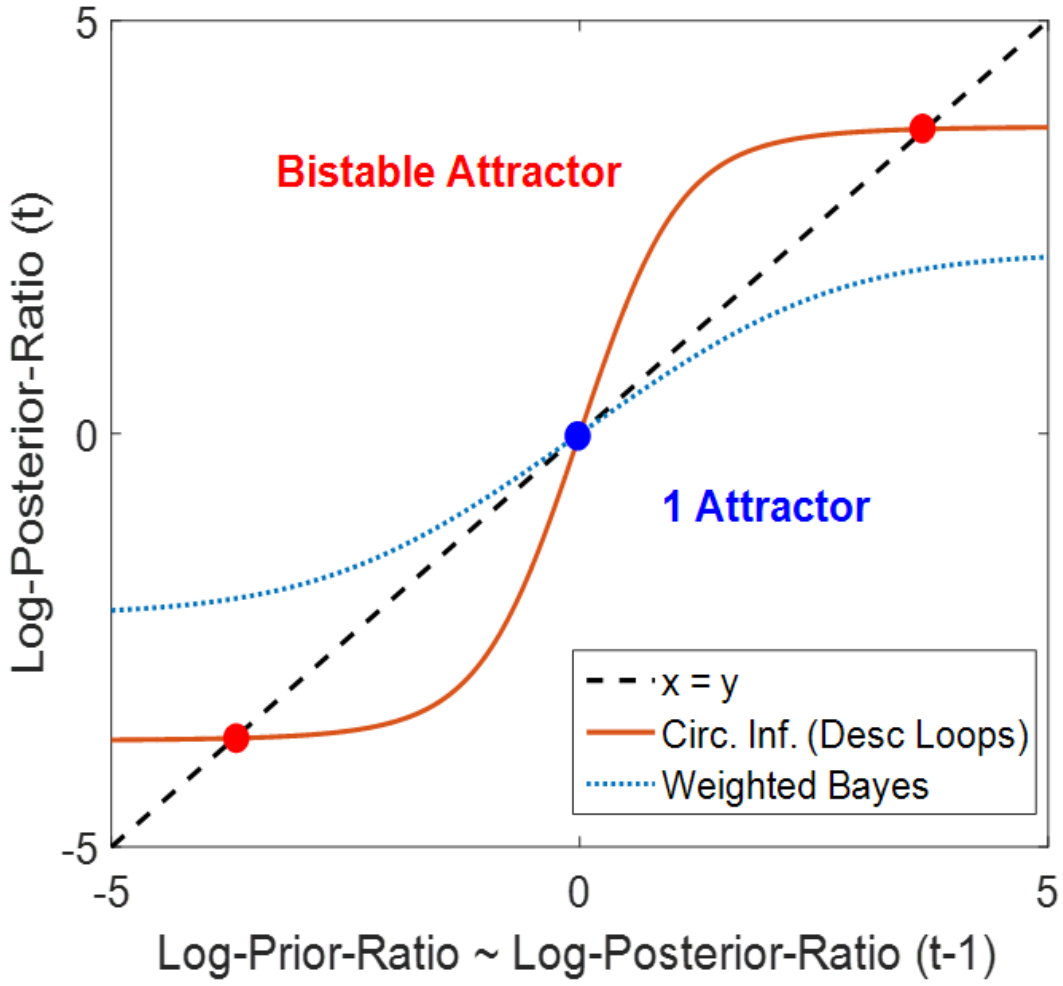


***Figure S5: Descending loops can generate a bistable attractor.*** *The present results suggest that an aberrant correlation between sensory evidence and priors might be at play in bistable perception, shaping which interpretation we see and when. In the same context, it is important to highlight that a Circular Inference model, but not a purely Bayesian model (e.g., the WB model), seems compatible with the phenomenology of bistable perception. When taking into account the dynamics, descending loops (i.e., amplifying accumulated data) introduce a positive feedback to the system, which generates a bistable attractor (2 stable states consisting of strong beliefs, one for each interpretation; red solid line). In contrast, a Bayesian model corresponds to a leaky integrator, in which the belief is similar to chance (blue dotted line) [27] and appears unable to generate bistability.*

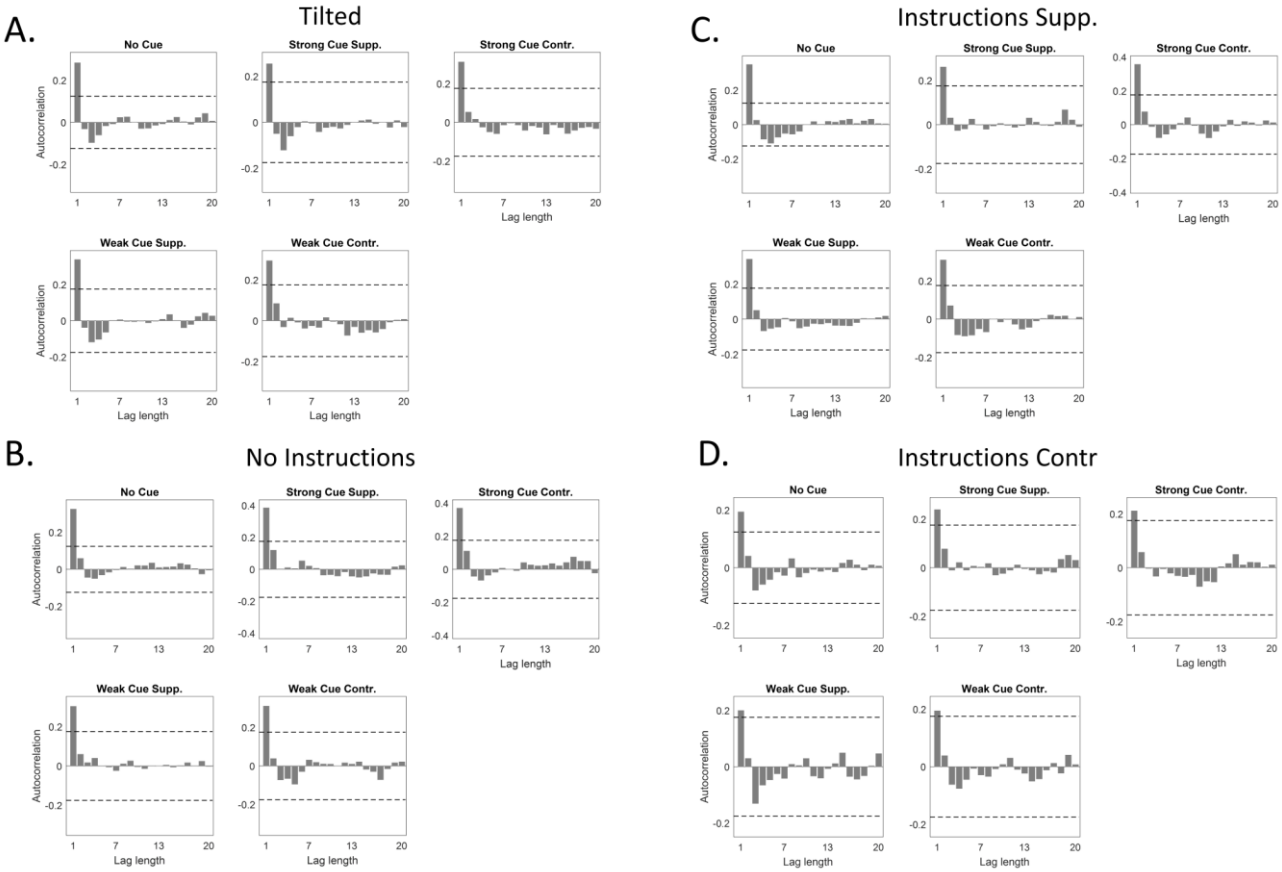


***Figure S6: Average autocorrelation functions for the different cue conditions (different subplots) and the different prior conditions*** *[**(a)**: tilted group (N=12); **(b)**: normal cube, no instructions (N=15); **(c)**: normal cube, supporting instructions (N=14); **(d)**: normal cube, contradictory instructions (N=14)]. Dashed lines correspond to 95% confidence intervals for a white noise process. Interestingly, with the exception of small lags (lag = 1), no other point in any of the autocorrelation functions exceeds the dashed lines, meaning that the effects of history (at least for the average data) are restricted to small time differences and can be neglected. Consequently, none of our models took into account temporal statistics (see* ***Discussion*** *for further consideration of this issue).*

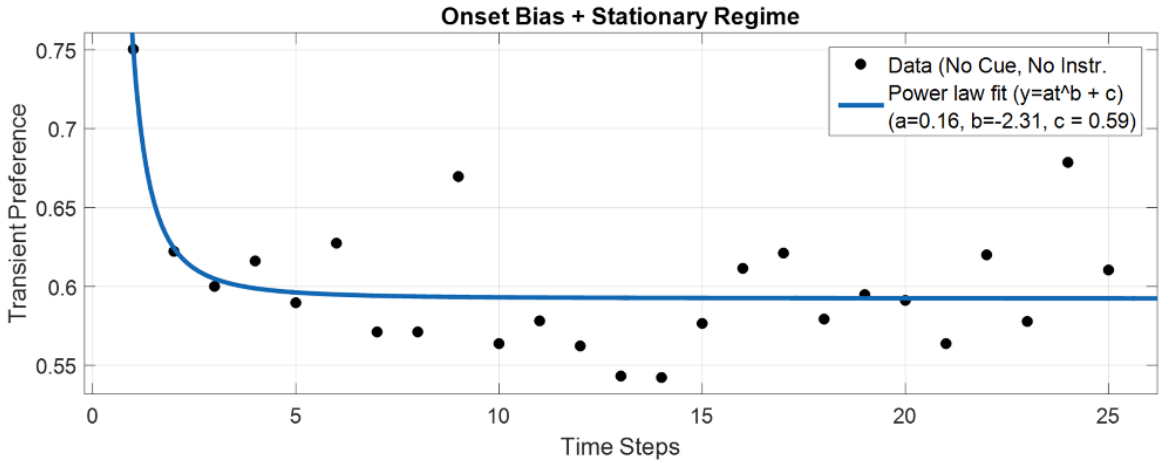


***Figure S7: Transient preference (TP) as a function of time (black dots) and fitted power law model (blue curve)****. Transient preference can be defined as the RP for different time steps. As expected, TP attains a high value (TP(1) = 0.75) at the beginning of a run (t = 1), indicating the presence of an onset bias that rapidly decreases until reaching a stationary regime (TP(st. reg.) ~ 0.6) [29]. Interestingly, TP in the stationary regime is above chance, indicating the presence of a persistent bias even after the onset bias fades out (the implicit bias that we describe in the* ***Main Text*** *is a combination of those two biases). This figure is reassuring, since absence of this pattern would indicate either a response bias or a very long inter-stimulus interval. This figure corresponds to the normal cube/no instructions/no cue condition (N=15), but a similar pattern was obtained for all other conditions as well (not presented).*

## References


1. Von Helmholtz, H. Concerning the perceptions in general. *Treatise on physiological optics* III, (1866).

2. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304 (2004).

3. Mamassian, P. & Landy, M. S. Observer biases in the 3D interpretation of line drawings. *Vision research* 38, 2817–2832 (1998).

4. Zhang, X., Xu, Q., Jiang, Y. & Wang, Y. The interaction of perceptual biases in bistable perception. *Sci Rep* 7, 42018 (2017).

5. Necker, L. A. LXI. Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *Philosophical Magazine Series 3* 1, 329–337 (1832).

6. Wheatstone, C. Contributions to the physiology of vision.–Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical transactions of the Royal Society of London* 371–394 (1838).

7. Blake, R. & Logothetis, N. K. Visual competition. *Nat. Rev. Neurosci.* 3, 13–21 (2002).

8. Dobbins, A. C. & Grossmann, J. K. Asymmetries in perception of 3D orientation. *PLoS ONE* 5, e9553 (2010).

9. Pouget, A., Dayan, P. & Zemel, R. S. Inference and computation with population codes. *Annu. Rev. Neurosci.* 26, 381–410 (2003).

10. Lochmann, T. & Deneve, S. Neural processing as causal inference. *Current Opinion in Neurobiology* 21, 774–781 (2011).

11. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* 37, 205–220 (2014).

12. Ma, W. J. Organizing probabilistic models of perception. *Trends in Cognitive Sciences* 16, 511–518 (2012).

13. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nature Neuroscience* 5, 598–604 (2002).

14. Körding, K. P. *et al.* Causal Inference in Multisensory Perception. *PLOS ONE* 2, e943 (2007).

15. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433 (2002).

16. Shen, S. & Ma, W. J. A detailed comparison of optimality and simplicity in perceptual

decision making. *Psychological Review* 123, 452–480 (2016).

17. Hudson, T. E., Maloney, L. T. & Landy, M. S. Movement Planning With Probabilistic Target Information. *Journal of Neurophysiology* 98, 3034–3046 (2007).

18. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* 74, 30–39 (2012).

19. Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the origins of suboptimality in human probabilistic inference. *PLoS Comput. Biol.* 10, e1003661 (2014).

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

21. Heeger, D. J. Theory of cortical function. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1773–1782 (2017).

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

23. Bishop, C. *Pattern Recognition and Machine Learning*. (Springer, 2006).

24. Okun, M. & Lampl, I. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat. Neurosci.* 11, 535–537 (2008).

25. Xue, M., Atallah, B. V. & Scanziani, M. Equalizing excitation-inhibition ratios across visual cortical neurons. *Nature* 511, 596–600 (2014).

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

27. Leptourgos, P., Denève, S. & Jardri, R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Curr. Opin. Neurobiol.* 46, 154–161 (2017).

28. Mamassian, P. & Goutcher, R. Temporal dynamics in bistable perception. *Journal of Vision* 5, 7–7 (2005).

29. Levelt, W. J. M. The Alternation Process in Binocular Rivalry. *British Journal of Psychology* 57, 225–238 (1966).

30. Klink, P. C., van Ee, R. & van Wezel, R. J. A. General validity of Levelt's propositions reveals common computational mechanisms for visual rivalry. *PLoS ONE* 3, e3473 (2008).

31. Nawrot, M. & Blake, R. Neural integration of information specifying structure from stereopsis and motion. *Science* 244, 716–718 (1989).

32. Pearson, J. & Clifford, C. W. G. Suppressed patterns alter vision during binocular rivalry. *Curr. Biol.* 15, 2142–2148 (2005).

33. Kanai, R., Moradi, F., Shimojo, S. & Verstraten, F. A. Perceptual alternation induced by

visual transients. *Perception* 34, 803–822 (2005).

34. Pearson, J. & Brascamp, J. Sensory memory for ambiguous vision. *Trends in Cognitive Sciences* 12, 334–341 (2008).

35. Long, G. M. & Toppino, T. C. Enduring Interest in Perceptual Ambiguity: Alternating Views of Reversible Figures. *Psychological Bulletin* 130, 748–768 (2004).

36. Orbach, J., Ehrlich, D. & Heath, H. A. REVERSIBILITY OF THE NECKER CUBE: I. AN EXAMINATION OF THE CONCEPT OF" SATIATION OF ORIENTATION". *Perceptual and motor skills* 17, 439–458 (1963).

37. Leopold, D. A., Wilke, M., Maier, A. & Logothetis, N. K. Stable perception of visually ambiguous patterns. *Nature Neuroscience* 5, 605–609 (2002).

38. Lynn, R. Reversible perspective as a function of stimulus-intensity. *Am J Psychol* 74, 131–133 (1961).

39. Babich, S. & Standing, L. Satiation effects with reversible figures. *Percept Mot Skills* 52, 203–210 (1981).

40. Van Ee, R., Van Dam, L. C. J. & Brouwer, G. J. Voluntary control and the dynamics of perceptual bi-stability. *Vision research* 45, 41–55 (2005).

41. Toppino, T. C. Reversible-figure perception: mechanisms of intentional control. *Percept Psychophys* 65, 1285–1295 (2003).

42. Denison, R. N., Piazza, E. A. & Silver, M. A. Predictive Context Influences Perceptual Selection during Binocular Rivalry. *Front Hum Neurosci* 5, (2011).

43. Chong, S. C. & Blake, R. Exogenous attention and endogenous attention influence initial dominance in binocular rivalry. *Vision Res.* 46, 1794–1803 (2006).

44. Dieter, K. C. & Tadin, D. Understanding attentional modulation of binocular rivalry: a framework based on biased competition. *Front Hum Neurosci* 5, 155 (2011).

45. Stonkute, S., Braun, J. & Pastukhov, A. The role of attention in ambiguous reversals of structure-from-motion. *PLoS ONE* 7, e37734 (2012).

46. Haijiang, Q., Saunders, J. A., Stone, R. W. & Backus, B. T. Demonstration of cue recruitment: change in visual appearance by means of Pavlovian conditioning. *Proc. Natl. Acad. Sci. U.S.A.* 103, 483–488 (2006).

47. Pearson, J., Clifford, C. W. G. & Tong, F. The functional impact of mental imagery on conscious perception. *Curr. Biol.* 18, 982–986 (2008).

48. Rock, I., Hall, S. & Davis, J. Why do ambiguous figures reverse? *Acta Psychol (Amst)* 87, 33–59 (1994).

49. Baker, D. H. & Graf, E. W. Natural images dominate in binocular rivalry. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5436–5441 (2009).

50. Zhou, G., Zhang, L., Liu, J., Yang, J. & Qu, Z. Specificity of face processing without awareness. *Consciousness and Cognition* 19, 408–412 (2010).

51. Kornmeier, J., Hein, C. M. & Bach, M. Multistable perception: When bottom-up and top-down coincide. *Brain and Cognition* 69, 138–147 (2009).

52. Intaitė, M., Noreika, V., Šoliūnas, A. & Falter, C. M. Interaction of bottom-up and top-down processes in the perception of ambiguous figures. *Vision Research* 89, 24–31 (2013).

53. Díaz-Santos, M. *et al.* Effect of visual cues on the resolution of perceptual ambiguity in Parkinson's disease and normal aging. *J Int Neuropsychol Soc* 21, 146–155 (2015).

54. Moreno-Bote, R., Knill, D. C. & Pouget, A. Bayesian sampling in visual perception. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12491–12496 (2011).

55. Schmack, K. *et al.* Delusions and the Role of Beliefs in Perceptual Inference. *Journal of Neuroscience* 33, 13701–13712 (2013).

56. Sundareswara, R. & Schrater, P. R. Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision* 8, (2008).

57. Gershman, S. J., Vul, E. & Tenenbaum, J. B. Multistability and perceptual inference. *Neural computation* 24, 1–24 (2012).

58. Brascamp, J. W., van Ee, R., Pestman, W. R. & van den Berg, A. V. Distributions of alternation rates in various forms of bistable perception. *Journal of Vision* 5, 1–1 (2005).

59. Pastukhov, A. & Braun, J. Cumulative history quantifies the role of neural adaptation in multistable perception. *J Vis* 11, (2011).

60. Li, H.-H., Rankin, J., Rinzel, J., Carrasco, M. & Heeger, D. J. Attention model of binocular rivalry. *Proc. Natl. Acad. Sci. U.S.A.* 114, E6192–E6201 (2017).

61. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879 (2006).

# Chapter 3

## A functional theory of bistable perception based on dynamical circular inference

In preparation for publication as:

Leptourgos P., Bouttier V., Jardri R., Deneve S. (In prep.). A functional theory of bistable perception based on dynamical circular inference

## Abstract

When facing ambiguous images, the brain cannot commit to a single percept and instead switches between mutually exclusive interpretations every few seconds, a phenomenon known as bistable perception. State-of-the-art mechanistic explanations suggest that bistability is the result of the combined action of three different processes, namely competition via lateral inhibition, slow adaptation and noise. Nevertheless, those models are largely based on ad-hoc assumptions, poorly adequate when it comes to functional questions. Here, we present a novel top-down approach to bistable perception, based on circular inference, a sub-optimal form of hierarchical probabilistic inference in which information is reverberated and amplified. We argue that descending loops, a type of circularity in which top-down predictions corrupt bottom-up sensory inputs, are essential for understanding both the existence and the phenomenology of bistability. More specifically, we show that descending loops result in the replacement of what is normally a temporal integration of unreliable sensory evidence by a bistable attractor switching between two highly trusted interpretations. Additionally, we demonstrate that the circular inference model is compatible with various qualitative aspects of bistability, including Levelt's laws and the stabilizing effects of intermittent presentation of the stimulus. Finally, we use this model to make new testable predictions, notably about the behaviour of individuals exhibiting aberrant percepts or beliefs such as in schizophrenia. Importantly, this is the first study to provide theoretical evidence that circularity in hierarchical neural networks, a mechanism that has been linked to the generation of psychotic symptoms, could also underlie cognitive functions in non-clinical populations, a statement with important consequences for the conceptualization of psychosis and psychopathology in general.

## Introduction

All perceptual systems have one fundamental goal: to interpret the surrounding environment based on noisy sensory evidence. In most cases, this task is performed very accurately and the correct interpretation is found. Sometimes, perceptual systems either fail to detect any meaningful interpretation (e.g. when sensory evidence is too degraded) or converge to the wrong one (e.g. visual illusions; [1,2]). Finally, a third possibility occurs (mainly in lab conditions, [3]) when ambiguity is high: the system detects more than one plausible interpretations but instead of committing to one of them, it switches every few seconds, a phenomenon known as *bistable perception* [4]. Despite ongoing scientific efforts, there has been no unanimous agreement either on the causes of bistability or its functional role.

The dominant mechanistic view of bistable perception suggests that it may result from the competition between different neuronal populations, each of them encoding a different interpretation of the sensory signal [5]. The two populations suppress each other via lateral inhibition while some form of slow negative feedback (e.g., spike frequency adaptation or synaptic depression) acts on the dominant population, weakening the interpretation that is currently perceived [6–10]. Additionally, injected noise renders switches irregular and in some models it can even be the driving force of the oscillatory behaviour [11–14]. Although these models have proven quite successful in describing different experimental observations (and linking them to the underlying neural mechanisms), they do not address functional considerations about bistable perception.

To overcome this issue, other groups suggested functional models of bistability, largely based on the idea that the brain is an inference machine and perception is equivalent to a probabilistic process ([15], e.g., Predictive Coding: [16,17] or Neural sampling: [18–20]). Compared to the mechanistic models, functional models make precise assumptions and predictions about the role and function of bistable perception (and perception in a more general sense) but they also suffer from their own shortcomings. More precisely, their inherent abstractness makes it hard to fit functional models to data but also to find links with an underlying neural implementation. Besides, one crucial question remains unanswered from a normative perspective, more particularly why would a system perceive something instead of nothing, if the sensory evidence that drives inference is completely unreliable (as is the case of completely ambiguous stimuli with flat priors).

In this paper, we tackle the problem of bistable perception by putting forward a model that gathers the advantages of both mechanistic and functional models while minimizing their

disadvantages. Based on previous experimental findings (**Chapter 2**), we suggest that bistability could be a perceptual manifestation of *circular inference*, a form of belief propagation in which priors and likelihoods are reverberated in the cortical hierarchy and consequently are overcounted and corrupted by each other (an idea similar to loopy belief propagation [21]; for a detailed description of the *circular inference* framework, see also [22,23]). In particular, we postulate that the phenomenology of bistable perception could be explained by the presence of "descending loops", a form of *circular inference* where top-down expectations (through feedback connections) corrupt the sensory representations such that the system "sees what it expects" [24]. Of note, previous work from our group linked *circular inference* with pathological brain function, as in the case of schizophrenia [22], but also with normal brain functioning [25].

In the following sections we derive the dynamics of inference in the presence of ambiguous sensory stimuli and descending loops. The result of circular inference is to replace what is normally a temporal integration of unreliable sensory evidence into a bistable attractor switching between two highly trusted interpretations. We demonstrate that this model can reproduce well known qualitative aspects of bistability, including the four Levelt's laws and the effects of intermittent presentation, while it also makes new testable predictions (e.g. about the behaviour of schizophrenia patients). Since circularity arises from an imbalance between neural excitation and inhibition in recurrent brain circuits [24], our approach bridges normative interpretations of bistable perception and possible underlying neural mechanisms.

## Methods

In this section, we introduce a simplified *circular inference* model of bistable perception and we highlight its underlying functional assumptions. For reasons of clarity, we will refer to the example of the Necker cube, an ambiguous 2D figure which is equally compatible with 2 different 3D cubes and generates bistability: a cube that is "*seen from above*" (later called the SFA Interpretation) and a cube that is "*seen from below*" (later called the SFB Interpretation) (**Figure 1a**). Note that the model can however be generalized to any other stimuli inducing perceptual rivalry.

### Generative model

Our model postulates that bistable perception is triggered by the same mechanisms and computations underlying normal perception. There is accumulating evidence that the brain

uses its cortical hierarchy to represent the causal structure of the world and then inverts this forward model in order to predict the most probable interpretation of the noisy sensory information, in other words perception can be viewed as an instance of hierarchical Bayesian inference [26,27] (**Figure 1a**). A particularly striking example of this inferential process is 3D vision (e.g., perception of the Necker cube). The brain has no direct access to the 3D structure of the perceived object. On the contrary it receives low-level 2D sensory information from the retina. In such context, the task of the perceptual system is to extract valuable depth cues from this sensory information and combine them with high level prior knowledge, to make ''educated guesses'' about the 3D object. Evidence suggests that this is a gradual process [28], with different brain regions representing features of different complexity: the lower levels of the visual cortex represent the basic features of the stimulus such as contours and orientations while higher levels are responsible for more abstract information such as the 3D organization of the stimulus [29,30].

In the case of the Necker cube, a veridical percept would correspond to a 2D object, representing a set of crossing lines. The presence of illusory depth cues forces the brain to consider the presence of depth. Nonetheless, since the cues are ambiguous and contradictory, the 2D projection of the hypothetical 3D stimulus could correspond to different 3D objects, namely the SFA and SFB interpretation mentioned before[1]. Consequently, in this work we assume that all visual information is processed by the visual system with respect to those 2 possible interpretations: sensory inputs are taken as evidence for the SFA or SFB interpretation, while all higher-level variables are considered binary, with values 1 and 0 corresponding to SFA and SFB respectively. Furthermore, the 2 interpretations are considered as mutually exclusive, an assumption compatible with the epistemological truth that 2 different 3D objects cannot occupy the same space [16].

We can represent the functional hierarchy where inference is implemented as a simple graphical model; a chain with 2 latent variables and an observation (**Figure 1a**). Roughly speaking, $X_{2D}$ could correspond to the 2D image, with orientation and depth of surfaces being explicitly represented ($2\frac{1}{2}$D sketch in Marr's terminology) while $X_{3D}$ represents the 3D interpretation. Sensory information ($S$) on the other hand can be regarded as the basic features of the image (or the primal sketch), where we receive a noisy measurement. The ultimate goal

[1] It's interesting to highlight that in a more general sense, the Necker cube is compatible with an infinity of 3D objects, among which the brain represents only the 2 symmetrical cubes. This reduction of possible causes could be the result of hyperpriors used by the brain and will not be considered by the current model.

of the perceptual system is to infer the 3D interpretation ($X_{3D}$) using the noisy measurement and any available prior knowledge (for more information about the generative model, see **Supplementary Material**).

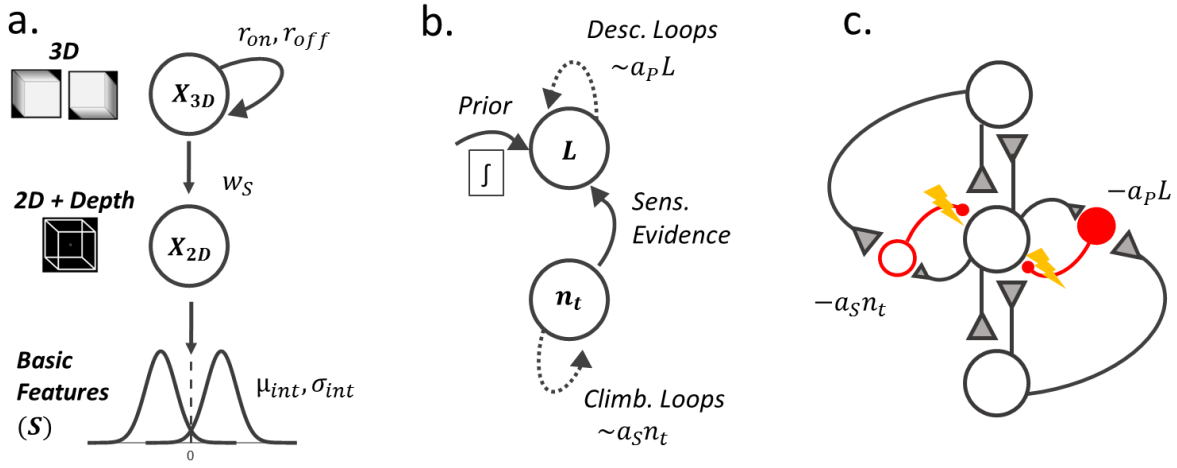


***Figure 1: Graphical representation of the model and neural implementation.*** *(a.): The internal model is a simple pairwise graph with binary variables (except for the sensory evidence which is normally distributed). The brain interprets the depth cues (basic features) as indication of real depth. Consequently, it first reconstructs the 2D figure and from that, it predicts the 3D object. Note that in reality (real forward model) there is one single 2D stimulus containing contradictory depth cues.* ***(b.):*** *From the Bayesian model (a.) we derived an attractor model that does inference in the presence of loops. The model accumulates noisy evidence, while descending loops add a positive feedback and climbing loops increase the sensory gain.* ***(c.):*** *A neural implementation of the circular inference model. Reciprocal excitatory connections generate loops which are balanced by strong inhibitory connections. Dysregulation of this balance (too much excitation or not enough inhibition) results in amplification of information and sub-optimal inference.*

As usual, we assume that sensory information is corrupted by Gaussian noise, with mean $\mu_{noise}$ ($\mu_{noise} = 0$ if the cube is completely unbiased and $\mu_{noise} \neq 0$ if there is a visual cue, e.g. contrast) and variance $\sigma^2_{noise}$ (**Figure 2**; black and grey distributions, corresponding to $P(S|X_{real})$). Crucially, the sensory evidence is ambiguous; the stimulus contains contradictory depth cues, although there is no real depth (2D structure). This implies that if the brain was using the correct model to do inference, then it would perceive the Necker cube as it is, a 2D object with illusory contradicting depth cues.

This is clearly not the case. The brain interprets the depth cues as meaningful (i.e., "depth cues signify the presence of depth"), and generates 3D explanations (SFA or SFB) of the sensorium which are driven by noisy inputs [11,13,31]. This also presupposes the representation

of a likelihood function containing two Gaussian distributions (internal mode), as illustrated in **Figure 2** (red and blue dotted distributions; $P(S|X_{2D} = 1) \neq P(S|X_{2D} = 0)$). Those distributions are assumed symmetrical, with mean $\pm\mu_{int}$ (independent of the type of the presented cube) and variance $\sigma_{int}^2$.

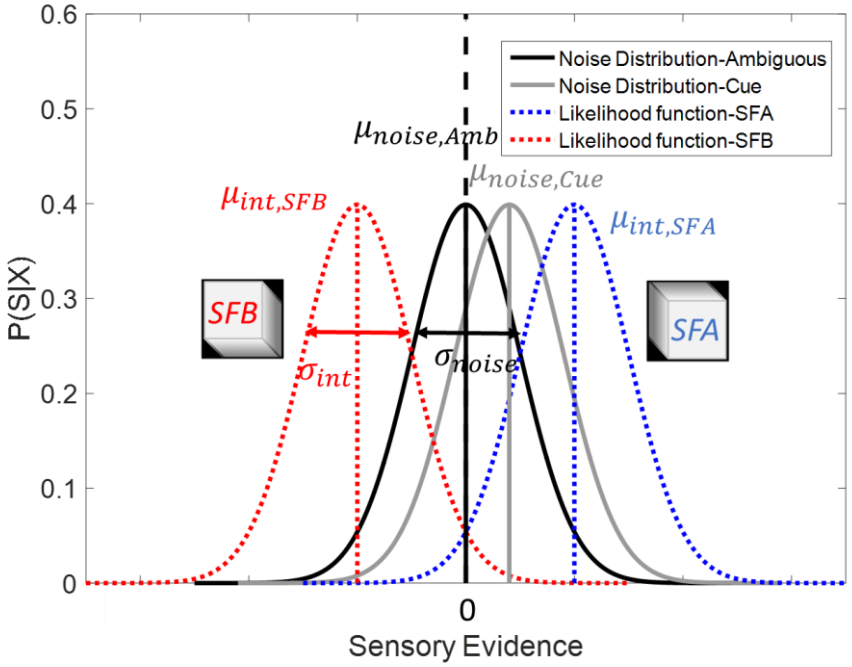


***Figure 2: Sensory Evidence.*** *In reality, there is a 2D stimulus that generates noisy sensory evidence (measurements). In the totally ambiguous case (cube with no extra cues), the noise distribution is a Gaussian with mean 0 (black). Visual cues add a bias, which is equivalent to a shift of the Gaussian (grey). Conversely, the brain interprets the depth cues as meaningful, constructing separate representations for the two cubes (SFA, SFB; two objects cannot occupy the same space). That corresponds to a likelihood function with two non-overlapping Gaussian distributions in the internal model (dotted red and blue distributions). Consequently, there is a discrepancy between the real model and the internal model.*

## Markovian statistics

Importantly, to make optimal decisions, the brain should not only count on current evidence, but also consider past perceptual decisions, in other words it should also take into account temporal regularities. Particularly in bistable perception, the dominant percept persists for a few seconds before switching back to the opposite interpretation, a behaviour attributed to correlated perceptual decisions [32].

This can be captured by *Markovian* statistics over variable $X$ (from now on used interchangeably with $X_{3D}$), resulting in a Hidden Markov Model (HMM; **Figure 1a**). Note that Markovian statistics constitute the simplest form of temporal dynamics and suggest that the current percept only depends on its very recent history. Formally, this can be expressed by the following equation:

$$P(X_t, X_{t+dt}) = P(X_{t+dt}|X_t) \times P(X_t) \neq P(X_{t+dt}) \times P(X_t) \quad (1)$$

where $X_t$ is the 3D interpretation at time t, and dt is some arbitrary time-step. The *Markovian* temporal statistics can be quantified by the 2 transition rates $r_{on}dt$ and $r_{off}dt$ given by the following formulas:

$$r_{on}dt = P(X_{t+dt} = 1|X_t = 0) \quad (2)$$

$$r_{off}dt = P(X_{t+dt} = 0|X_t = 1) \quad (3)$$

Intuitively, the HMM is similar to a leaky integrator, where new information is accumulated over time while old information is discarded with a rate that depends on the transition rates (see also **Continuous Model**). Namely, the posterior probability at time t becomes the prior probability for time t+dt (**Figure 1b**).

Note that the two transition rates are not necessarily equal. Assuming $r_{on} \neq r_{off}$ introduces a bias in the system, which we could interpret as an implicit preference for one of the two interpretations, unrelated to the presence of cues in the sensory evidence. This is very useful in the case of the Necker Cube, where people usually prefer the SFA interpretation, according to a general prior to view things from above [33].

## Circular inference

Once built, the generative model needs to be inverted, to give a posterior probability. An intuitive and biologically plausible way to invert such a model is through the use of message-passing algorithms [21]. One of these algorithms is belief propagation (BP), in which nodes exchange information reciprocally in the form of local messages and calculate beliefs (posteriors) simply by multiplying all the incoming information. Sensory evidence, in the form of a likelihood function, is propagated upwards, in a bottom-up fashion, while priors are sent in the opposite direction, in a top-down fashion. Because of the hierarchical structure of the

model, information in one node constitutes the sensory evidence of the node above and the prior information of the node below, leading to a recursive process.

It's interesting to highlight that the reciprocal connectivity of the nodes in the algorithm closely resembles the recurrent structure (recurrent excitation) of the brain [34]. An important consequence of such reciprocity is that it creates loops in which information can be counted multiple times, resulting in overestimations in the posterior probabilities [22]. To avoid that, one must carefully remove all the redundant information that has already been taken into account (for more information and a mathematical description of Belief Propagation, see **Supplementary Material**). Crucially, the algorithm performs exact inference, provided that the generative model is an acyclic graph (as is the case here; [21]) and subtraction of the redundant information is done properly.

It has been hypothesized that inhibition might be tracking and removing all redundant information and that a tight balance between neural excitation and inhibition is fundamental for the brain to perform exact inference (**Figure 1c**). The slightest imbalance in favour of excitation could cause inefficient removal of redundant information, leading to a form of approximate inference known as *circular inference* [22]. Motivated by the understanding of psychotic symptoms in schizophrenia, like hallucinations and aberrant beliefs, this framework introduced a new parameter to the belief propagation algorithm, accounting for redundant exchange of information between the different nodes (see **Continuous model**). This means that a single piece of information (e.g. sensory evidence or prior) is integrated several times, due to "loops of information": descending loops on one hand in which sensory evidence is corrupted by priors, generating a system that "*sees what it expects*"; and climbing loops on the other hand in which the priors are corrupted by sensory evidence and the system "*expects what it sees*" [23]. This amplification of information leads to over-accumulation of evidence up to certainty (as for instance in the jumping-to-conclusions phenomenon observed in delusional individuals) [22]. Interestingly, recent evidence suggests that this altered form of belief propagation can be found in healthy populations as well, explaining their behaviour in probabilistic reasoning tasks [25] but also, most importantly, the integration of priors and sensory evidence in bistable perception tasks (**Chapter 2**).

Based on these findings, we assume that inference in our hierarchical model is also corrupted by loops (**Figure 1b**). As it will be explained later, only a system with loops (primarily descending loops) can generate a behaviour compatible with the phenomenology of bistability.

## Decision criterion

Finally, no attempt to model perceptual decision making is complete without defining the decision criterion. In agreement with most studies, we consider "maximum a posteriori" (MAP), which is viewed as the optimal strategy in perceptual problems, since it maximizes accuracy. In our case, that means that a switch occurs when the logit of $X$ ($L$) crosses 0. In addition to that, we also considered a more conservative criterion (see **Supplementary Material)**, according to which the threshold depends on the current percept and switches only occur when there is substantial evidence in favour of the opposite interpretation. Compared to MAP, such a decision criterion has the additional advantage of making the perceptual system more robust to noise, allowing for the generation of gamma distributions of phase durations (see also the section on **Distribution of phase durations** and **Supplementary Material**).

## The continuous model

In the previous section we introduced a functional model able to use belief propagation to do probabilistic inference in a HMM, while information is amplified due to *circular inferences*. In a previous study, Jardri et al suggested that circular inference can be well approximated by a single discrete equation, which ignores the hierarchy and presents a belief as a sum of 2 terms: a likelihood term and a prior term, both corrupted by loops [25]. Based on this, we suggest that our *dynamical circular inference* model can also be approximated by a similar discrete equation (**Figure 1b;** see **Supplementary Material** for a detailed description of the discrete model).

Considering infinitesimally small time-steps ($dt \to 0$), the discrete model turns into the following stochastic equation, which describes how the belief about the 3D interpretation of the Necker cube changes over time (see derivation in **Supplementary Material**):

$$\frac{dL}{dt} = 2w_S a_P L + (r_{on}e^{-L} - r_{off}e^{L}) + (r_{on} - r_{off}) + w_{int}(2w_S - 1)(1 + 2w_S a_S)n_t = f(L) \qquad (4)$$

where $L = log(\frac{P(X=1|S_{0\to t})}{P(X=0|S_{0\to t})})$ is the belief and $n_t$ is the noisy sensory evidence, modelled as a Gaussian process (with or without drift, depending on the $\mu_{noise}$):

$$n_t \sim N(\mu_{noise}dt, \sigma^2_{noise}dt) \qquad (5)$$

$r_{on}$ and $r_{off}$ represent the 2 transition rates, while $a_S$ and $a_P$ quantify the strength of the loops (climbing and descending respectively) and represent the amount of amplification of the feedforward/sensory and feedback/temporal information [25].

Finally, we can define the gain of the sensory evidence as follows:

$$v = w_{int}(2w_S - 1)(1 + 2w_S a_S) \quad (6)$$

where $w_{int} = \frac{2\mu_{int}}{\sigma_{int}^2}$ is the reliability of the sensory evidence, according to the internal model, and $w_S$ corresponds to an additional feedforward weight. Interestingly, $w_S$ also appears in the first term, where it constrains the descending loops, an indication of the interaction between sensory evidence and priors in *circular inference*. **Table 1** contains a summary of all the parameters of the model.

| Variable | Description | Link to other variables |
|---|---|---|
| $\mu_{noise}$ | Drift of sensory evidence | - |
| $\sigma_{noise}$ | Standard deviation of sensory evidence | - |
| $\mu_{int}$ | Mean of likelihood function | - |
| $\sigma_{int}$ | Standard deviation of likelihood function | - |
| $w_S$ | Feed-forward weight | - |
| $a_P$ | Descending loops | - |
| $a_S$ | Climbing Loops | - |
| $r_{on}$ | Transition rate (0→1) | - |
| $r_{off}$ | Transition rate (1→0) | - |
| $w_{int}$ | Reliability of sensory evidence | $w_{int} = \frac{2\mu_{int}}{\sigma_{Int}^2}$ |
| $v$ | Sensory gain | $v = w_{int}(2w_S - 1)(1 + 2w_S a_S)$ |
| $b$ | Bias | $b = r_{on} - r_{off}$ |

***Table 1: The parameters of the model***

It's important to mention that the derivation of (4) requires $r_{on}$, $r_{off}$ and $a_P$ but not $a_S$ to be proportional to $dt$. This means that amplification of priors is considered as a gradual process, which builds up with time, while amplification of sensory evidence is considered

instantaneous. Note however that assuming $a_S$ proportional to $dt$ would only result in $(1 + 2w_S a_S) \approx 1$, meaning that the effect of the climbing loops would be negligible.

Taking a closer look at (4), we see that it consists of 4 terms, from which only the first two depend on $L$. The first term implements the effect of the descending loops and has the tendency to stabilise perception and push $L$ towards extreme values (positive or negative; as we will describe later, descending loops turn the system into a bistable attractor and the extreme values correspond to the new attractor states). This stabilising effect is balanced by the second term, which implements a leak towards the prior of the system (described below) and is quantified by the two transition rates. The third term adds a bias to the system (a non-zero prior), which is independent of the sensory evidence (an example of that could be the implicit preference of people to see things from above, creating a preference for the SFA interpretation of the Necker Cube; [33]).

Finally, the noise term is a Gaussian process (with or without drift) multiplied by a weight that depends on the climbing loops and the internal model. The noise pushes the belief $L$ away from its stable states, forcing it to explore the energy landscape imposed by the three first terms. We highlight that in case of $\mu_{noise} \neq 0$ (e.g., after adding visual cues), a second bias, equal to the drift $\mu_{noise}$, is added to the first [11,35,36]. Although they seem similar, the 2 biases have very different interpretations and functions. Contrary to the first bias, which depends on the difference of the transition rates and illustrates some prior knowledge about the statistics of the environment, the second term depends on the sensory evidence. As a result, it disappears when there is no stimulation, as in the case of intermittent presentation (see section about **Intermittent presentation**).

To summarise, eq. (4) corresponds to an attractor model, where switches are driven by noise [11]. Importantly, it's entirely based on functional assumptions and all the parameters have a probabilistic interpretation. We note that the model does not contain adaptation (or any other form of slow negative feedback), as such a mechanism is not necessarily implied by the underlying function (see also **Supplementary Material**). Crucially, the presented model is a functional model with a straightforward mechanistic interpretation, enjoying advantages from both classes of models.

## Simulations

For all the simulations later presented, we used the Euler – Maruyama algorithm. The time step was fixed at $dt = 0.01s$. Both standard deviation of the noise (real model) and of the likelihood function (internal model) were taken equal to 1. The mean of the likelihood function was also fixed at $\pm 1$. $\mu_{noise} = 0$ for the completely ambiguous case and $\mu_{noise} \neq 0$ when sensory evidence was biased, ranging between -1 and 1. Climbing loops were always considered as absent ($a_S = 0$) and descending loops were fixed at $a_P = 1$, except if mentioned otherwise. $w_S$ was between [0.7, 1] and transition probabilities between [0.25, 1]. The difference between the 2 transition probabilities was taken between [0, 0.25] (0 corresponds to the unbiased case). The initial belief in all simulations was $L_0 = 0$.

## Results

In the previous section we presented an attractor model that describes the evolution of a system's belief in a bistable context. The model was derived from first (normative) principles and illustrates a system that uses belief-propagation (more particularly a belief-propagation proxy [25]) to perform probabilistic inference in an internal, hierarchical model with Markovian statistics. Crucially, inference in the model is corrupted by the presence of loops, resulting in overcounting of information.

In this section, we explore the properties of the model in terms of dynamics but also in terms of behavioural predictions. As a first step, we highlight the importance of the descending loops in the generation of bistable perception, from a phenomenological and from a mechanistic point of view. Subsequently, we illustrate how the model can reproduce some of the most seminal features of bistable perception, like Levelt's laws but also some counterintuitive findings, including stabilization of perception after a brief disappearance of the stimulus (for shorter intervals, bistable perception is known to get destabilized, a behaviour that is also predicted by our model, when considering state dependent transition rates; see **Supplementary Material**). Finally, we present further consequences of the model, notably detailed predictions about the performance of schizophrenia patients exposed to bistable stimuli. **Table 2** summarises the characteristics / predictions of the model (with and without loops as well as for different decision criteria).

| | Functional Interpr. | Selection Problem | Alternation Problem | Confidence Problem | Bistable Attractor | Robustness | Levelt's Laws | Gamma Distribution | Intermittent-Stabilisation | Intermittent-Destabilisation |
|---|---|---|---|---|---|---|---|---|---|---|
| HMM, $a_P = 0, MAP$ | + | + | + | - | - | - | + | - | - | - |
| HMM, $a_P > 0, MAP$ | + | + | + | + | + | - | + | - | + | - |
| HMM, $a_P > 0, \varepsilon > 0$ | + | + | + | + | + | + | + | + | + | - |
| HMM, $a_P > 0, \varepsilon > 0,$ r(t) | + | + | + | + | + | + | + | + | + | + |

***Table 2: Model Predictions for the different versions of the model.*** *Robustness refers to how robust against noise are perceptual decisions.*

## Preamble: the phenomenological features of bistable perception

From a phenomenological point of view, bistability is a unique experience. Prolonged viewing of an ambiguous stimulus generates unstable percepts that switch between two configurations, despite the stimulus being exactly the same. Perception in that case seems to be dissociated from stimulation. Additionally, despite the ambiguous and unreliable nature of the stimulus, each interpretation usually persists for many seconds before switching back, and most of the time, it is perceived with high levels of confidence. This fascinating phenomenon leads to several unresolved questions. For instance, why would a perceptual system choose to favour one interpretation instead of the other or, if we go a step further, why choose one interpretation instead of perceiving both at the same time? In addition to that, why would that system change its mind? Finally, how does this system generate strong beliefs in the absence of strong evidence? Following Hohwy and colleagues, we call the first question the "*selection problem*" and the second question the "*alternation problem*" [16]. Moreover, we call the third question the "*confidence problem*".

In agreement with previous studies, we argue that "selection" is a simple consequence of the brain's function: to make "perceptual" predictions under uncertainty [15–17]. According to this hypothesis, the brain chooses the cause that best explains its sensory evidence. If at a given moment the sensory evidence, combined with any available prior knowledge (e.g., SFA preference) points more strongly to a certain interpretation, this interpretation will be picked and perceived by the brain.

Furthermore, apart from the sensory evidence and the priors, additional information (e.g., epistemological truths) might have to be considered when solving ill-posed problems, such as the problem of 3D perception. For example, common sense dictates that two different objects cannot occupy the same part of the visual space [16]. We postulate that such a hyperprior

renders the 2 interpretations mutually exclusive, and consequently impossible to be perceived at the same time.

We further argue that the "*alternation problem*" is a consequence of another property of the brain: evidence accumulation. In the model, this is captured by the Markovian statistics and it is regulated by two transition rates. More particularly, the new noisy evidence is integrated into the past accumulated evidence and pushes the belief away from its stable state. Every time the belief crosses the threshold, a switch occurs.

Although the "*selection*" and the "*alternation*" problems can be solved rather easily by referring to general functions of the brain, the "*confidence*" problem seems more difficult to deal with. Despite being unstable, percepts usually persist many seconds before switching, while they also enjoy high levels of confidence. On the contrary, a system doing exact inference in the presence of some environmental volatility (i.e., a system without loops and with non-zero transition rates) would switch very often, since the belief would have the tendency to hover around the prior (**Figure 3a,b (red)**). This problem could be solved by considering a perfect integrator ($r_{on} = r_{off} = 0$). Nevertheless, such a system would perform suboptimally in an unstable environment. In addition to that, even a perfect integrator is not able to generate strong beliefs in the absence of strong and reliable data, as for example when facing bistable stimuli. Crucially, both persistence and high confidence would be expected from a system that over-counts its accumulated evidence (prior knowledge), i.e. a system with descending loops. As shown in **Figure 3a (blue),** even weak descending loops increase the beliefs to extreme values, generating a system that is very confident about what it perceives, even if information is too noisy and ambiguous. At the same time, descending loops increase the persistence of the percepts, by amplifying the stabilizing effect of past information. In the next section, we will demonstrate that this dual effect of the descending loops is due to their ability to change the dynamics of the system by transforming it into a bistable attractor.

## Dynamical Systems Analysis

Let us now explore the dynamics of the model, by considering particular cases for the parameters of the system. First of all, let's examine the case in which there are no descending loops. Then, $a_P = 0$ and the first term disappears completely. The resulting system is equivalent to a HMM with transition rates $r_{on}$ and $r_{off}$ [37]. An example of such a system is presented in

**Figure 3a (red)**, while dynamics is illustrated in **Figure 3c,d.** Importantly, a HMM has only one stable fixed point (the prior) ($f(L) = 0, f'(L) < 0$) that depends on the 2 rates:

$$L_{St,a=0} = \log\left(\frac{r_{on}}{r_{off}}\right) \tag{7}$$

$L_{St,a=0}$ is 0 if the 2 rates are equal (e.g., structure from motion) and non-zero if they are not equal (e.g., Necker Cube). Consequently, the belief hovers around the prior without getting large positive or negative values (**Figure 3a (red)**), especially if sensory evidence is rather unreliable ($w_S$~0.5 or $\mu_{int}$ ~0, as in bistable perception) and the world relatively unstable ($r_{on}, r_{off} > 0$). Such a system would live in constant uncertainty, it would be overly affected by noise and would face serious difficulties in making perceptual decisions and acting upon them.

Descending loops push the belief away from the prior and towards more extreme values (**Figure 3a, (blue)**). The dynamics of a system with positive descending loops ($a_P > 0$) is shown in **Figure 3c,d**. If there is no leak ($r_{on} = r_{off} = 0$), the system is inherently unstable and the slightest noise induces complete certainty ($L = \pm\infty$, see **Figure 3c (orange)**). For non-zero transition rates, the effect of the loops is constrained. In that case, for certain values of $a_P, r_{on}$ and $r_{off}$, they give rise to a bistable attractor (**Figure 3c,d**).

In the unbiased case ($r_{on} = r_{off} = r$; e.g. structure from motion or 45°-tilted Necker cube), $L = 0$ is always a fixed point (**Figure 3c**). When the descending loops are weak compared to the leak, it is the only fixed point of the system and it is stable. That is true up to the value:

$$a_P^{Pf} = \frac{r}{w_S} \tag{8}$$

At this value, the system undergoes a Pitchfork bifurcation (**Figure 4a,b**): The existing fixed point becomes unstable and 2 additional attractors are generated, given by the 2 symmetrical, non-zero solutions of the equation $f(L_{st}^{Bist}) = 0$. The stronger the descending loops (or the weaker the leak), the further apart the 2 symmetrical attractors are (**Figure 4a,b**). Furthermore, (8) means that it is harder to get a bistable attractor (more loops are needed) in cases of higher volatility (larger $r$) or more unreliable sensory evidence ($w_S$ closer to 0.5).

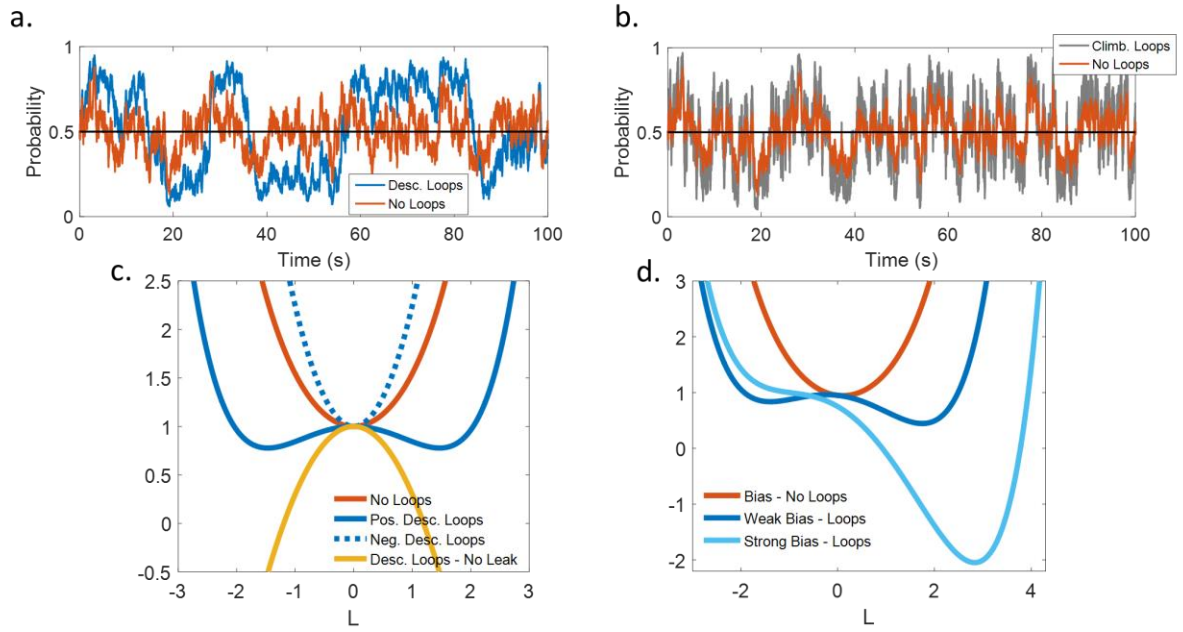


***Figure 3: Dynamics. (a.,b.):*** *Without loops, the system is equivalent to a leaky integrator (red). The accumulation of noisy evidence results in a belief that hovers around the prior, indicating a system that is constantly uncertain about the environment and switches very often. The addition of descending loops (positive feedback) pushes the beliefs towards more extreme values and increases persistence, in agreement with the phenomenology of bistable perception* ***(a.****; blue). On the other hand, climbing loops increase the gain of the noise, resulting in more extreme beliefs combined with more switches (****b.****; grey).* ***(c., d.):*** *From a mechanistic point of view, (positive) descending loops generate a bistable attractor, whose stable fixed points correspond to (strong beliefs about) the two interpretations (blue). On the contrary, a balanced system (no loops) has only one attractor, the prior (red). For the sake of completeness, we also present the effect of negative descending loops which act as a secondary leak (dashed blue) while (positive) descending loops in the absence of leak (zero transition rates) result in an inherently unstable system. The energy landscape can be symmetrical* ***(c.)*** *or assymetrical* ***(d.)*** *depending on whether the two rates are equal. Note that a strong bias forces the system to get stuck to one interpretation (light blue).*

Adding a bias to the system ($r_{on} \neq r_{off}$; e.g., SFA bias in Necker cube) creates an asymmetry in the energy landscape and shifts the fixed point away from 0 (**Figure 3d**). A Saddle Node (SN) bifurcation occurs when one of the 2 local extrema of $f(L)$ touches the x axis (**Figure 4c,d**; for a mathematical description of the SN bifurcation, see **Supplementary Material**). Qualitatively speaking, a SN bifurcation retains the stability of the existing fixed point (contrary to the symmetrical case, here the position of the fixed point is a function of $r_{on}, r_{off}$ and $a_P$) and on top of that it generates an additional pair of stable and unstable fixed points (**Figure 4c**). A bistable attractor can exist only in a narrow range of biases (difference between $r_{on}$ and $r_{off}$), more particularly in the range constrained by the 2 SN bifurcation points (one for $r_{on} > r_{off}$ and one for $r_{on} < r_{off}$; **Figure 4d**). Given that in reality a combination of weak sensory evidence

and weak priors is not very frequent, it is not strange that bistability is rather uncommon in everyday life.

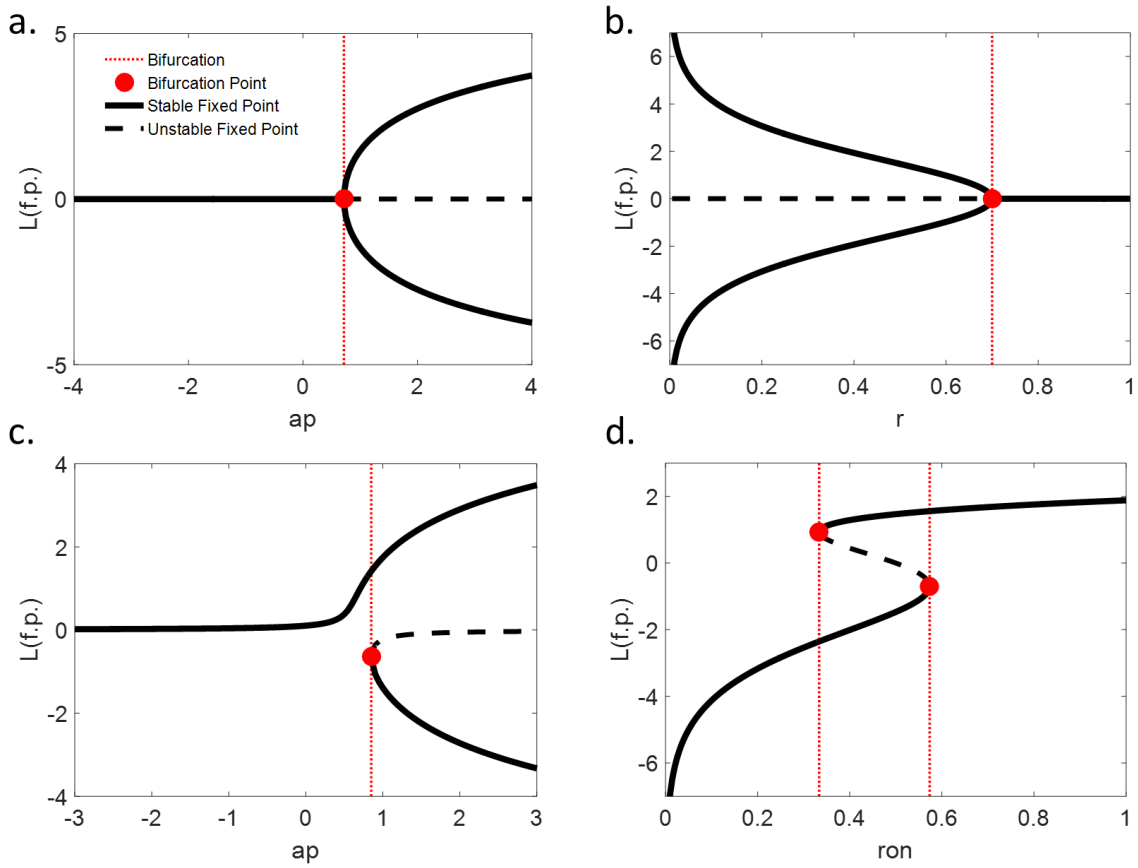


***Figure 4: Bifurcation Diagrams.*** *Descending loops change the dynamical behaviour of the system (**(a.,b.):** Pitchfork bifurcation for symmetrical systems; **(c.,d.):** Saddle-Node bifurcation for asymmetrical systems). When transition rates become too small, the fixed points go to infinity and the system becomes unstable **(b.)**. Additionally, bistability can exist in a narrow range around symmetry **(d.)**.*

Along with the effect of the positive descending loops, it's worth considering what would happen if we added negative loops to the system ($a_P < 0$). While positive descending loops mean that redundant top down messages are propagated in the hierarchy, corrupting new sensory evidence, negative loops mean exactly the opposite: the system corrects too much for the redundant messages, to a point where it starts removing useful information as well (throwing away part of the prior messages). Despite the loss of information, the effect of the negative loops on dynamics is less overwhelming than the effect of the positive loops, resembling a secondary leak mechanism (they push $L$ towards 0, independently of the prior, **Figure 3c (dotted)**).

Until now, our analysis focused mainly on the effects of the descending loops. Indeed, their role in bistable perception, according to the model, appears crucial, since they are the cause of the bistable attractor (it's important to repeat that the model makes no assumptions about the dynamics; the presented dynamics is a necessity, imposed by the functional mechanisms). On the other hand, climbing loops play a less important role. According to (6), climbing loops increase the gain of the sensory evidence (noise) (**Figure 3b**) and consequently they act by destabilising perception and reducing the effect of the bias on predominance (for more details, see section on **How schizophrenia patients perceive bistable stimuli**). As a result, it is not possible to fully separate the effect of climbing loops from the effect of the reliability of sensory evidence (the other term in (6)).

To conclude, descending loops could constitute a crucial part of the machinery of a system exhibiting bistable perception. When they are strong enough to overcome the effect of the leak, they generate a bistable attractor, implementing a memory-like mechanism that pushes the belief towards more extreme values, based on the previous observations. This helps the system make decisions and act upon them in the absence of fully convincing evidence. In the next sections, we explore the predictions of the model regarding well known psychophysical features of bistable perception.

## Levelt's Laws

An important qualitative feature of bistable perception is Levelt's laws. These laws constitute a set of 4 psychophysical propositions relating the strength of the bistable stimulus to the phenomenology of binocular rivalry [35], and more generally of bistable perception [36]. Despite some recent modifications in their formulation (to account for new experimental data [38,39]), Levelt's laws remain fundamental for our understanding of the machinery of bistability and an important crash-test for any potential model. We will present one by one the four revised propositions and will critically discuss them through the prism of the *dynamical circular inference* model.

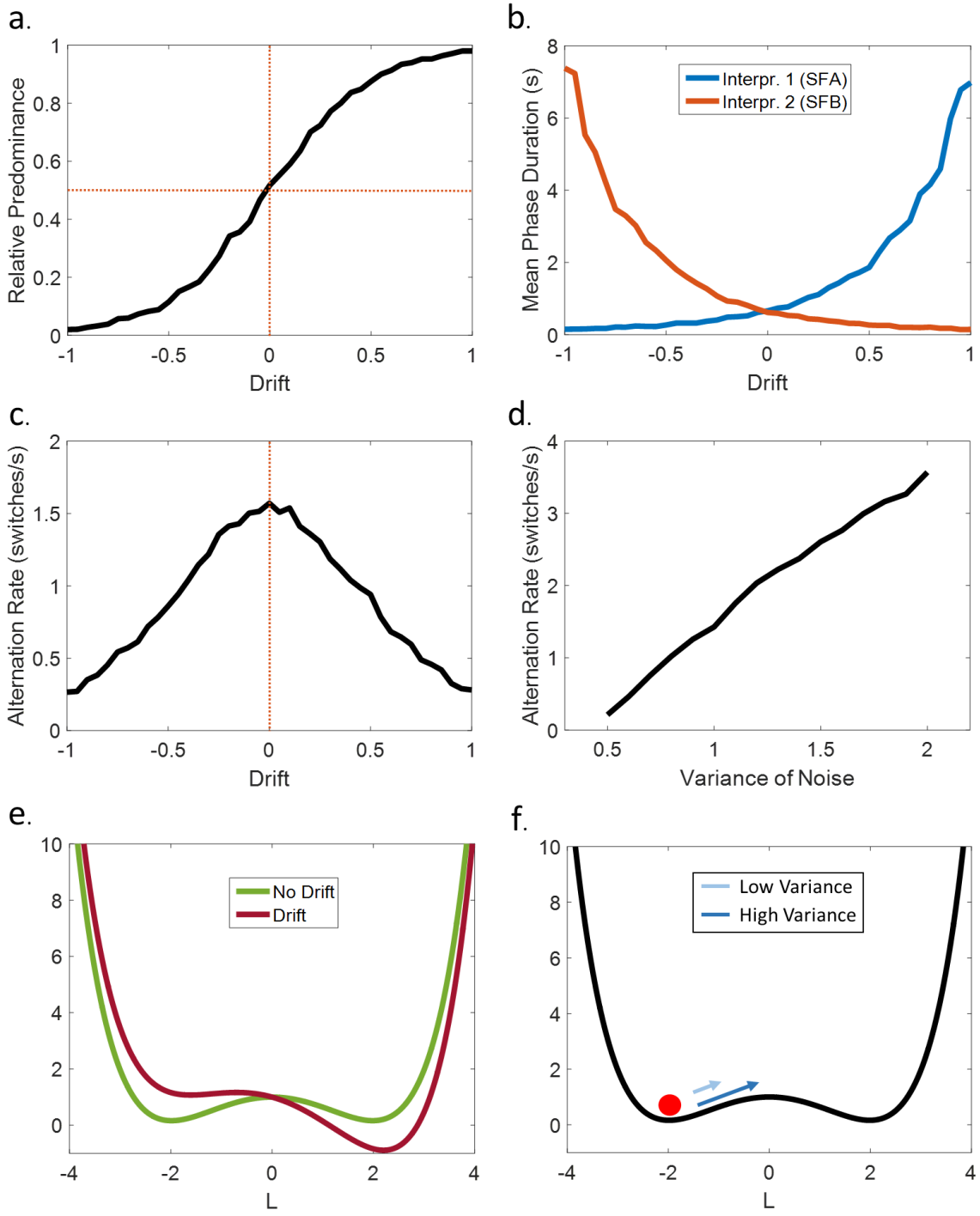


***Figure 5: Levelt's Laws.*** *The circular inference model qualitatively reproduces the 4 Levelt's propositions.* ***(a.):*** *1st proposition - Increasing the stimulus strength of one perceptual interpretation increases the predominance of this perceptual interpretation.* ***(b.):*** *2nd proposition - Manipulating the stimulus strength of one perceptual interpretation of a bistable stimulus doesn't influence equally the average dominance duration of both interpretations, but mainly affects the persistence of the stronger interpretation.* ***(c.):*** *3rd proposition - Increasing the difference in the stimulus strength between the 2 perceptual interpretations should result in a decrease in the perceptual alternation rate.* ***(d.):*** *4th proposition - When we increase the strength of both interpretations, the number of switches increases.* ***(e.,f.):*** *Interpretation of the 4 laws according to the present framework. Adding a visual cue (drift) affects the stability of both attractors. The effects are not equal due to the non-linearities* ***(e.;*** *Propositions 1-3).* *In addition to that, increasing the*

*variance of the noise results in stronger destabilization from the stable states, hence more switches (**f.**; Proposition 4).*

### *1st Levelt's law*

The first proposition links the stimulus strength with the predominance of each interpretation. It postulates that increasing the stimulus strength of one perceptual interpretation increases the predominance of this perceptual interpretation [36]. For example adding a cue to the Necker cube helps the relevant interpretation gain more perceptual dominance compared to its rival. Although in modern terminology proposition 1 sounds more like a tautology, it is still useful in order to detect stimulus features (or parameters of the model) that affect the strength of an interpretation [39]. Within our model, we can parameterize the strength of the sensory evidence by adjusting the drift $\mu_{noise}dt$ of the Gaussian process, which biases the sampling of evidence (**Figure 2**). As expected, the more positive the drift the closer the relative predominance goes to 1 (the opposite for negative drift) (**Figure 5a**), in agreement with the first proposition.

### *2nd Levelt's law*

The second proposition is less intuitive than the first and posits that manipulating the stimulus strength of one perceptual interpretation of a bistable stimulus doesn't influence equally the average dominance duration of both interpretations, but mainly affects the persistence of the stronger interpretation [36,40]. For example, increasing the strength of a visual cue in the Necker Cube example will mainly affect the mean dominance duration of the corresponding interpretation. The *dynamical circular inference* model is fully compatible with Levelt's second law, as presented in **Figure 5b**: making the drift more positive (bias for SFA) will predominantly affect the mean phase duration of the SFA interpretation (the opposite happens for a negative drift and the SFB interpretation). Indeed, the drift acts as an additional bias term in (4), that deepens the well of the strong interpretation, while at the same time it makes the other well shallower (**Figure 5e**). This dual effect of the drift (not obvious in other models in which different variables represent the different interpretations, see also [11]), along with the model's inherent non-linearity can explain Levelt's second law [40].

*3rd Levelt's law*

Levelt's third proposition is closely related to the second proposition [39] and suggests that increasing the difference in the stimulus strength between the 2 perceptual interpretations should result in a decrease in the perceptual alternation rate [36]. In the Necker Cube example, this proposition implies that adding a visual cue results in fewer switches. Importantly, the *dynamical circular inference* model behaves exactly as the third proposition dictates. As shown in **Figure 5c**, alternation rate gets its maximum value for drift = 0 (completely ambiguous stimulus) and decreases symmetrically as the drift becomes more positive or negative, a direct consequence of the third law [40].

*4th Levelt's law*

Finally, the fourth proposition goes one step further and discusses what happens to the alternation rate if we equally increase the strength of both interpretations. In this case, the number of switches increases, resulting in a higher alternation rate. Contrary to the 3 first propositions, the fourth one illustrates the effect of a simultaneous manipulation of both interpretations (global stimulus strength). In the model, this global manipulation can be captured by a change in the variance of the noise distribution $\sigma_{noise}$. A higher variance corresponds to increased sensory gain which results in more exploration of the energy landscape due to the noise (**Figure 5f**). Consequently, as illustrated in **Figure 5d**, increasing $\sigma_{noise}$ results in more switches, in agreement with Levelt's fourth law.

## Intermittent presentation

When an ambiguous stimulus is presented continuously, switches between competing interpretations occur randomly every few seconds, with consecutive phase durations being largely independent [41]. Based on this observation, many researchers came to the conclusion that bistable perception is principally a memoryless process ([42], see also [43,44]). Nevertheless, this conclusion contravenes with another observation: the fact that people tend to perceive the same interpretation repeatedly, when ambiguous stimuli are presented intermittently, for a wide range of OFF-durations (intervals during which stimulus is absent) [45,46]. This second observation forced researchers to consider the presence of some perceptual memory [47], which manifests itself when the stimulus disappears from the screen. A variety of

mechanisms implementing this memory have been proposed, including low-level mechanisms such as adaptation (combined with sub-threshold effects; [9]), or high level memory mechanisms located outside the extrastriate cortex [46,48,49]. The *dynamical circular inference* model offers a different explanation for this stabilization effect, based on the descending loops.

In agreement with previously published experimental observations, our model predicts no significant correlation in the duration of successive phases [41,42], as we would expect from a model that doesn't contain adaptation (or adaptation-like) mechanisms [44]. On the other hand, the model should be able to predict a stabilization effect, when the stimulus disappears for brief durations. In order to quantify stabilization, many studies referred to the alternation rate, which is the number of switches in a time interval [45,46,50]. However, this measure is not ideal as it can be affected by various confounding factors including different presentation durations and switches occurring during ON-duration (interval during which stimulus is present). Moreover, the alternation rate considers both interpretations together and obscures any possible asymmetries. Instead, we used the survival probability (SP) of each interpretation, which is the probability that the dominant percept at the end of an ON-duration would be dominant again when the stimulus reappears after the OFF-duration.

**Figure 6a** illustrates our interpretation of the phenomenon (5 ON-OFF cycles, $a_P > 0$). In order to understand the behaviour of the model during intermittent presentation, we need to look at the phase portrait of the dynamical system (trajectories in absence of stimulation). Without descending loops ($a_P = 0$), the belief returns to its prior value ($log(\frac{r_{on}}{r_{off}})$), due to the leak (**Figure 6b,f).** The longer the OFF-duration, the more the belief approaches the prior. The resulting effect on SP is presented in **Figure 6c,g** (solid lines). For the unbiased system, the model predicts that both SP will decrease towards 0.5 (chance), with a time constant that depends on the transition rates.

On the contrary, the SP in a biased system would reach symmetrical points above and below chance, with the values depending on the strength of the bias. Additionally, SP for the continuous case (stimulation is not interrupted; in that case we measure the survival probability in constant intervals) are also presented in **Figure 6c,g** (dashed lines). Importantly, we observe that even without descending loops, we can still get a relative stabilization in intermittent presentation, compared to continuous presentation (dashed lines are below solid lines), due to the accumulation of noise in the continuous case. On the other hand, this version of the model does not predict an increasing SP with the OFF-duration, while at the same time it imposes symmetrical convergence points for large OFF-durations.

The descending loops ($a_P > 0$) change the behaviour of the system. The phase portrait of this system is presented in **Figure 6d,h**. Instead of one single point where all the trajectories meet, now we observe 2 clearly distinct basins of attraction, symmetrical for an unbiased system and asymmetrical for a biased system. Importantly and contrary to what we previously observed, the SP do not converge to symmetrical values. Instead, in an unbiased system the two probabilities reach the same value, which is not chance-level (they can increase or decrease with the OFF-duration, depending on the parameters) whereas in a biased system they converge to non-symmetrical values, which depend on the interaction of descending loops and transition rates (**Figure 6e,i**). Additionally, we observe again a relative stabilization compared to the continuous presentation (dashed lines).

An important comment needs to be made. The current version of the model does not predict a destabilization occuring for small OFF-durations, usually for values below 500ms. Other models have attributed this observation to a combined effect of adaptation and sub-/near-threshold signals [9]. Another possibility in the present context is that, destabilization could be obtained by considering history-dependent transition rates (see **Supplementary Material** for more details). More particularly, when the system perceives interpretation 1 (e.g., SFA), the probability of switching from 1 to 0 (i.e, $r_{off}$) increases exponentially towards a high value, while the probability of switching from 0 to 1 (i.e., $r_{on}$) decreases exponentially towards a baseline value. The opposite happens when interpretation 0 (e.g., SFB) is the dominant interpretation. When none of the 2 interpretations is dominant (e.g. during OFF-durations), both rates go back to baseline. This adaptation-like mechanism destabilizes the dominant percept and boosts exploration of the whole energy landscape.

From a functional point of view, such a mechanism forces the system to explore non-optimal perceptual choices, like a non-optimal decision criterion (e.g Softmax). Alternatively, it could simply implement a hyperprior that things do not remain the same for a long time. Independently of the interpretation, changing rates result in time-dependent attractor states, which under certain circumstances can cause a destabilization for short OFF-durations (see **Supplementary Material**).

To summarise, *dynamical circular inference* predicts stabilization of bistable perception for longer OFF-periods. In addition to that, it makes specific predictions about the persistence of each interpretation separately, which could help to experimentally distinguish a system with loops from a system without loops.

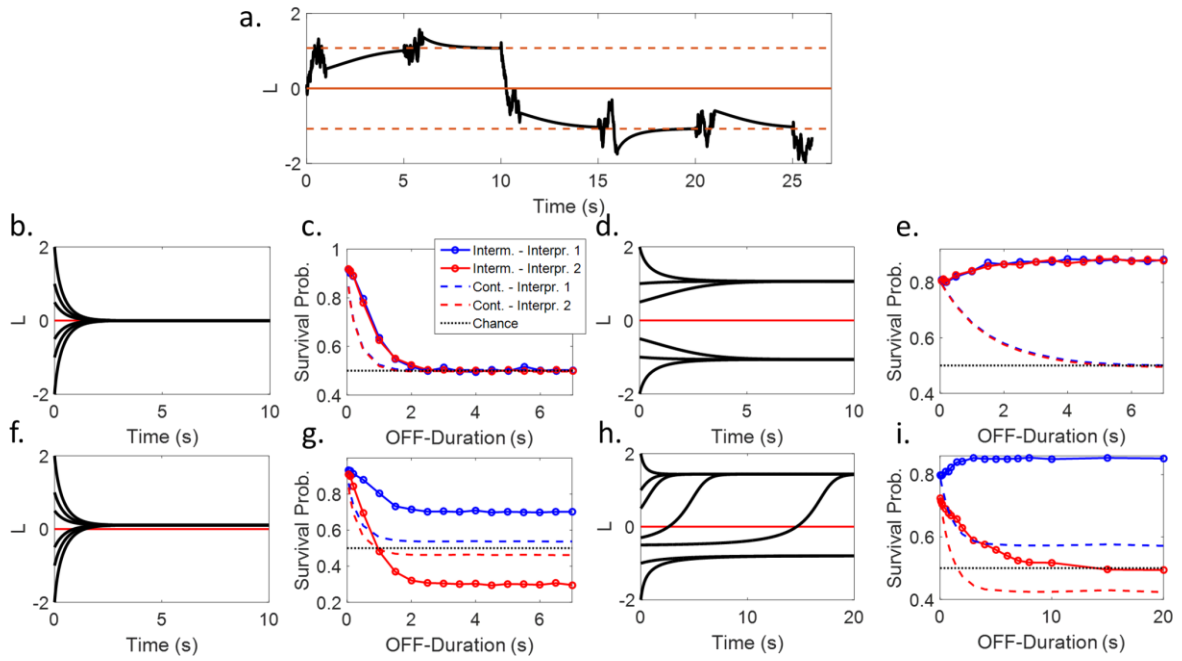


***Figure 6: Continuous vs Intermittent presentation.*** ***(a.):*** *An interpretation of the phenomenon, based on the circular inference framework. When the stimulus disappears, the belief converges to an attractor. The behaviour of the system depends on the number and the value of the fixed points.* ***(b.,c.,f.,g.): No loops*** *- If there are no (descending) loops, when the stimulus disappears the beliefs converge to the prior* ***((b.): No implicit preference; (f.): Implicit preference)****. Consequently, for longer OFF-Durations, the 2 survival probabilities (blue and red solid lines) either converge to 0.5* ***((c.): No implicit preference)*** *or to symmetrical values* ***((g.): Implicit preference).*** *In both cases, the stimulus is not stabilised for longer intervals. Interestingly, it is more stable compared to continuous presentation (dashed lines).* ***(d.,e.,h.,i.): Descending loops*** *– Descending loops generate a bistable attractor* ***((d.): No implicit preference; (h.): Implicit preference)****. Crucially, when they are strong enough, they cause stabilisation for longer intervals* ***((e.): No implicit preference; (i.): Implicit preference)****. Furthermore, in the biased case, survival probabilities converge to assymetrical values.*

## Distribution of phase durations

Another important feature of bistable perception, shared by human and non-human observers is the distribution of dominance durations. Surprisingly, although there is huge variability in the mean phase-duration between participants (but also within participants and between conditions or stimuli), there is an impressive similarity in the shape of the distribution of phase-durations, which is well approximated by a gamma or log-normal distribution [51–53] (but see also [54]).

It is well established theoretically that models in which switches are driven solely by noise do not generate gamma distributions, instead they produce histograms that look like exponentials [11]. For this reason, many attractor models assume some amount of adaptation

acting on top of the noise [11,14]. Adaptation is not strong enough to trigger a switch; It is sufficient though to push the mode of the distribution to higher durations. The *dynamical circular inference model*, is a functional model in which switches are only triggered by noise [14]. Crucially, the model, in its simplest form, does not contain any adaptation-like mechanism, since such a mechanism is not a necessary consequence of the function. It is thus not surprising that it produces histograms approximated by exponential distributions (**Figure 7**). This happens because when a switch occurs and while the belief is still close to the threshold, the slightest noise can push the belief back to the other side, causing an instantaneous switch. Indeed, no mechanism can prevent such a rapid alternation, leading to very high frequencies for short phase-durations.

Various additional mechanisms could prevent those rapid switches, generating gamma distributions of phase-durations. Such a mechanism, based on a more conservative decision criterion, is presented in **Supplementary Material**. Whenever a switch occurs, the threshold jumps to its symmetrical value. Such hysteresis is sufficient to prevent noise from causing an instantaneous switch. In particular, a switch occurs only when there is substantial evidence against the current interpretation, which makes the system robust and results in very infrequent short phase-durations and thus, gamma-distributed histograms (**Figure S4**).

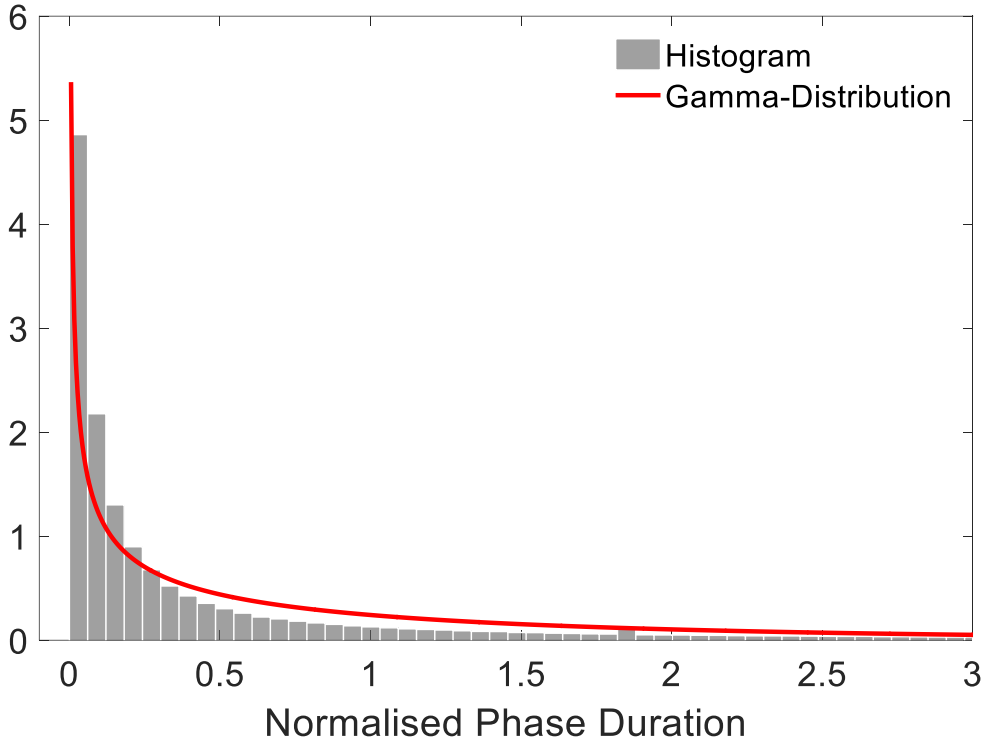


***Figure 7: Phase duration histogram.*** *In agreement with previous results, our nosie-driven (switches are triggered by noise), circular inference model predicts exponential-like phase duration histograms, instead of the experimentally-observed gamma distributions. Importantly, a more conservative decision criterion generates gamma distributions (see* ***Supplementary Material, Figure S4****).*

## How Schizophrenia patients perceive bistable stimuli?

So far, we have described a functional model of bistable perception, based on the notion of *circular inference*. Accumulating evidence supports the idea that circularity (and especially a small amount of descending loops) is a common property of the human brain, reflecting some inherent limitations of neural circuits [25] (see also **Chapter 2**). However, it has also been suggested that *circular inference* could be the cause of several cognitive and/or perceptual disorders, including schizophrenia [22,24]. In a previous study, Jardri et al found that on average, patients with schizophrenia have stronger climbing loops compared to a group of matched healthy controls [25]. Additionally, it was evidenced that "positive" (i.e., psychotic) symptoms, including hallucinations and delusions, correlate with the amount of climbing loops (i.e., sensory evidence amplification), "negative" symptoms, including lack of motivation and anhedonia, correlate with the amount of descending loops (i.e., prior amplification) and finally cognitive disorganization correlates with the total amount of loops ($a_S + a_P$). Considering these

previous findings, an interesting question is what does the current *dynamical circular inference* model predict about the behaviour of schizophrenia patients exposed to bistable stimuli?

**Figure 8c,d,g,h** illustrates the effect of climbing loops on bias (Relative Predominance; **Figure 8c,g**) and stability (Mean Phase-Duration; **Figure 8d,h**). As previously shown, climbing loops increase the gain of the noise, facilitating the jumps between the 2 attractors (**Figure 3b**). Consequently, our model predicts that patients with more severe hallucinations and delusions should be less biased in their responses (both due to inherent priors and visual cues) but also less stable (especially the interpretation that is supported by the visual cue). Especially the effect of climbing loops on Relative Predominance, although it might seem counterintuitive (over-counting of sensory evidence leads to a smaller effect of that evidence), illustrates the detrimental effect of the higher gain of noise on the accumulation of evidence.

Descending loops deepen the wells of the energy landscape and consequently, they produce the exact opposite effects. As shown in **Figure 8a,b,e,f**, they increase both the bias (**Figure 8a,e**) and the stability (**Figure 8b,f**) of schizophrenia patients with more severe negative symptoms.

Our model thus predicts different but specific patterns of behaviour when schizophrenia patients are exposed to bistable stimuli that depend on the predominance of their positive and negative symptoms.

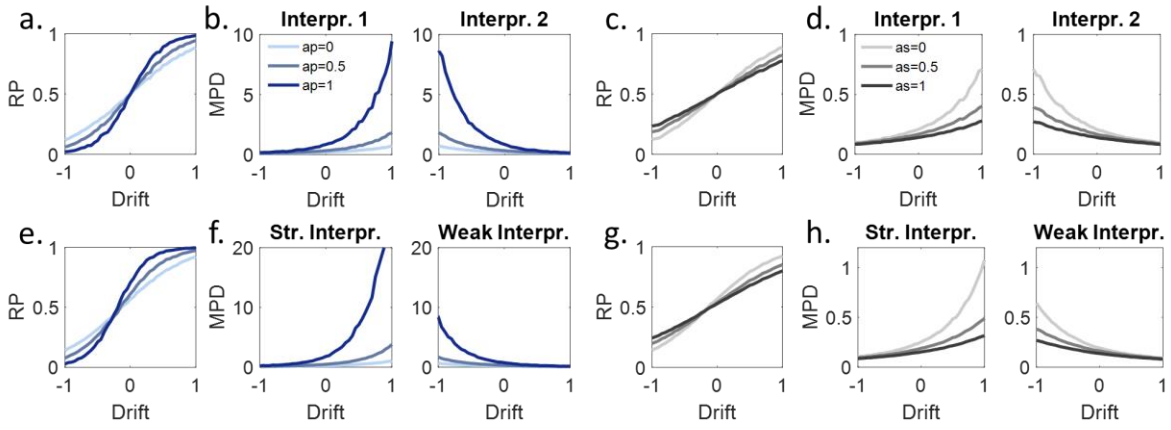


***Figure 8: Predictions for Schizophrenia Patients. (a.,b.,e.,f.): Descending loops*** *– Previous work [25] related the negative symptoms of schizophrenia with stronger descending loops. The current model predicts that patients with stronger descending loops (darker blue), or more severe negative symptoms, will exhibit a stronger bias (Relative Predominance) due to visual cues / drift* ***((a.): No implicit preference; (e.): Implicit preference)****, a stronger bias in the ambiguous condition in the assymetrical case (**e.**; RP at 0) and they will be more persistent (Mean Phase Duration)* ***((b.): No implicit preference; (f.): Implicit***

***preference)**. **(c.,d.,g.,h.): Climbing loops** - The same work [25] related the positive (psychotic) symptoms of schizophrenia with stronger climbing loops. The prediction for the patients with stronger climbing loops (more severe psychotic symptoms; darker grey) is exactly the opposite: weaker bias due to visual cues **((c.): No implicit preference; (g.): Implicit preference)** or implicit prior (**g.**; RP at 0) and less persistence **((d.): No implicit preference; (h.): Implicit preference)**.*

## Discussion

We introduced a novel functional approach to bistable perception, notably based on the idea that bistability could be the result of priors' over-counting due to *circular inferences* [22]. More specifically, we presented a dynamical model with Markovian statistics, whose dynamics implements inference in a hierarchical representation of the world [21,26]. We postulated that due to inherent limitations of neural networks implementing inference (see also below for an alternative interpretation), priors are propagated between layers without control, forming loops of information (descending loops). The direct consequence of this phenomenon is that feedback messages moving from high-level association areas to lower-level sensory areas corrupt the feedforward messages moving in the opposite direction, causing the system to «see what it expects» [24]. This idea is in agreement with previous results (**Chapter 2**), suggesting that the way people integrate priors and sensory evidence in bistable perception cannot be fully captured by the Bayes theorem and instead it could be better explained by a *circular inference* model.

From the point of view of the underlying dynamics, the descending loops have equally important consequences: Due to their inherently stabilising effect, they push the system to undergo a Saddle-Node bifurcation (or Pitchfork, in the case of an unbiased system), resulting in the creation of a bistable attractor. The emerging dynamical system can explain various intriguing features of bistable perception, primarily its mere existence. The descending loops overcount and artificially inflate the accumulated noisy information. Hence, they lead to a system with high levels of conviction that perceives clearly, persistently and in alternation the 2 potential interpretations.

Crucially, although descending loops are necessary for bistability, they are not sufficient. Even in the presence of descending loops, a system using the correct generative model would completely discard the visual information as completely unreliable, in other words it would ignore the depth cues, resulting in 2D percepts. In addition, we thus assumed that the system attributed some reliability to the sensory evidence (use the depth cues to infer depth), which corresponds to the system using the wrong internal model (**Figure 2**). One explanation is that

in everyday life, although ambiguity is a common obstacle for our perceptual systems, completely ambiguous stimuli are very rare [3,55]. In the general case, depth cues are tightly related to the presence of depth. As a result, in everyday situations, completely discarding sensory evidence would be suboptimal and equivalent to throwing away important information.

In addition to the existence of bistable perception, the *dynamical circular inference* model can also explain several qualitative features of bistability. It is compatible with Levelt's four laws, while the descending loops also generate a memory-like mechanism that allows for stabilization of the percepts when the stimulus is presented intermittently.

Beyond our model, various other implementations have been proposed to account for bistability and some of its characteristics. On one hand, people have proposed mechanistic models and have either focused on the neural mechanisms [7,8,10] or the dynamics [6,9,11]. Typically, those models assume two competing populations of neurons and a combination of mechanisms including lateral inhibition, adaptation and noise. Despite their critical importance for our quantitative understanding of the phenomenon and its neural implementation, those models remain largely descriptive, a fact with important implications for their explanatory power. First, they are usually designed on an ad-hoc basis and their mechanisms are adjusted (without formal constraints) depending on the phenomenon they are trying to explain (e.g. [56]). On top of that, mechanistic models, with few exceptions (e.g. [40]), remain agnostic regarding the functional role of bistable perception. More particularly, although they are very good in answering the «what» questions (mechanisms and implementations), they are not appropriate for the «why» questions (epistemological questions).

To answer the second type of questions, other groups have put forward functional models of bistable perception, which approach the problem in a top-down fashion [16–20,57]. Essentially, those approaches focus on the type of problems that perceptual systems usually encounter (e.g., deal with uncertainty) and adequately impose functional constraints on the structure of the models (e.g. optimality constraints in Bayesian models). Although these models are equally important and complementary to the mechanistic models described above, they also have their own disadvantages: Functional models are more abstract models whose links with the real neural systems (and with the mechanistic models) are usually vague. Therefore, they are rarely fitted to real data and their predictions remain largely qualitative. Moreover, finding closed form solutions for those models is rarely possible.

The *circular inference* model presented in this paper reconciles those 2 approaches, keeping most of their advantages and solving some of their disadvantages. Basically, it's an

attractor model based entirely on normative assumptions, whose parameters have a functional (probabilistic) interpretation. Similar to other attractor models, it assumes that switches in perceptual bistability are driven by noise, in agreement with existing evidence [12–14]. Additionally, its dynamical behaviour has important similarities with that of other attractor models [11].

Crucially, despite apparent similarities, our model has fundamental differences from other attractor models. Most importantly, its properties are not ad-hoc, on the contrary everything is imposed by the underlying function of the system: to use belief propagation to do probabilistic inference in an internal model with loops. For example, the bistable attractor is not imposed to explain certain features of bistability, instead it's a direct consequence of the descending loops. In the same vein, our model makes a clear distinction between a bias induced by sensory evidence and a bias resulting from the system's implicit preference (prior knowledge), thus enabling the generation of asymmetries in the absence of stimulation (intermittent presentation).

Additionally, our model is different from other functional models, although they share some key assumptions, in particular the fact that perception is probabilistic in nature. Many functional models of bistable perception are based on the idea that inference is approximated by a sampling process, without explicit calculation and knowledge of the exact posterior distribution [18–20]. Although such an idea appears plausible, many sampling algorithms (e.g. Markov Chain Monte Carlo) require long sampling times to generate accurate results, making those solutions unrealistic. Additionally, it remains unclear whether those models could account for less trivial experimental results, including stabilization under intermittent presentation.

In the present model, we assume that sensory evidence is corrupted by Gaussian noise with zero-mean in the case of complete ambiguity. Other studies have considered more complex noise distributions, e.g. bimodal distributions which can be described as mixtures of 2 symmetrical Gaussians [17]. We can note that despite their differences, considering alternative noise distributions has no effect on the qualitative results presented in the paper, provided that the average evidence is zero in case of perfect ambiguity and non-zero when there are additional visual cues.

It's also important to highlight that contrary to other functional models, *circular inference* can be fitted to real data. On one hand, we were able to formalize the belief propagation algorithm with loops using a single stochastic differential equation, which was then

used to derive analytical and semi-analytical results. On the other hand, without important consequences, one can significantly reduce the number of free parameters to the following set: [sensory gain $v$, functional descending loops ($a'_P = 2w_S a_P$) and bias ($b = r_{on} - r_{off}$)], with the possible addition of the drift $\mu_{noise}$. Those parameters can then be fitted to data, even with a brute-force approach.

It has been argued that *circular inferences* are linked, at the neurophysiological level, to an imbalance between neural excitation and inhibition in favour of excitation [24,58]. This imbalance might concern only local microcircuits, encompassing pyramidal cells and local interneurons (**Figure 1c**), or more global networks, potentially involving thalamocortical or cortico-striatal long range connections [24]. Although both are plausible neural implementations of loops, local interneurons make a better candidate in the particular case of bistable perception. Indeed, it has been argued that bistability is a rather low level process mainly occurring within the visual cortex ([4,59,60]; but see [61,62], arguing for the involvement of high-level areas) while the involvement of local inhibition is also supported by pharmacological evidence [63].

Apart from normal brain functioning, *circular inference* has been used to account for clinical dimensions in schizophrenia [22,25]. Combined with those studies, the results presented here could have further consequences regarding our understanding and definition of psychopathology. More particularly, our model of bistable perception implies that the same generic mechanisms could be involved in severe symptoms, such as hallucinations and delusions in schizophrenia, but at a lesser degree could also explain common perceptual phenomena, such as bistable perception. Such a blurring of the boundaries between normal and pathological brain functioning appears in agreement with the idea that psychosis may exist along a continuum with normal experience [64–68]. Still, when and how exactly those mechanisms go awry and generate pathological symptoms remains an open question. In addition to that, the present model constitutes an extension of previous *circular inference* models, reinterpreting the framework in dynamical systems' terms and relating it to other influential frameworks [69,70].

Could circularity offer a relative advantage to perceptual systems using them or is it simply a manifestation of the inherent limitations of neural systems? According to the present results, a system performing exact inference in the face of very unreliable evidence would be too vulnerable against noise and might have difficulties in making decisions and acting. On the contrary, moderate descending loops could give a boost to the system by slightly amplifying its

priors and help it make robust decisions even when evidence is absent (after all, both "fighting" and "flying" are better than standing still; a similar explanation was suggested by Moreno-Bote and colleagues, who interpreted bistability as exploratory behaviour under uncertainty [40]). Moving a step further, a system with flexible descending loops (e.g. a system that can regulate its E/I balance through neuromodulators, such as dopamine, serotonin or acetylcholine [71,72]) could still use exact inference under most circumstances and only recruit the loops when there is a need (e.g. when the stimulus is completely unreliable but there is a need to decide, as in the case of bistable perception). This suggestion, although speculative, could reconcile the present results with evidence showing perfect balance between excitation and inhibition at different scales [73–75] and is furthermore easily testable (e.g., by measuring E/I balance during bistability and during stimulation with unambiguous stimuli).

Some limitations of the model presented here must be acknowledged. First of all, while the model tackles the problem of inference under ambiguity, it contains no learning. In fact, we assume that the internal model has already been learnt in advance and doesn't change during the experiment, even after the addition of visual cues (**Figure 2**; only the drift changes, but not the internal model). Although this is a relatively strong assumption, it can be justified if we assume that enough training has taken place before the hypothetical experiment.

Finally, some small improvements to the model could be imagined in the near future. For instance, the current version does not explain why in certain cases, although there is evidence in favour of one interpretation, this interpretation does not gain access to consciousness. (e.g., during OFF-periods in intermittent presentation). In other words, although our simplified model predicts well perceptual behaviour when the 2 interpretations are the only relevant choices (e.g., when there is presentation of an ambiguous stimulus), it fails to predict perceptual decisions when additional interpretations become relevant (e.g., absence of cube when there is no stimulus, during OFF-periods). Despite that, the model is still able to explain the priming (and suppressing) effect observed when the stimulus is briefly removed.

In conclusion, we described bistable perception as a probabilistic inference process, under the influence of amplified priors due to the presence of descending loops in the cortical hierarchy. The model explains why bistable perception occurs in the first place and qualitatively predicts several of its properties. Additionally, it has important implications for the neural correlates of bistability and the relation between normal brain functioning and pathology, ultimately linking computation, behaviour and neural implementation.

## Supplementary Material

### From belief propagation to the discrete model

Belief propagation (BP; or Sum-Product algorithm) is a general and efficient algorithm that performs inference in directed acyclic graphs (DAG) (**Figure S1a**; [21]). In the most general case, the DAG has first to be transformed into a factor graph (FG), which elucidates the factorization properties of the underlying joint distribution (**Figure S1b**). A FG contains 2 types of nodes: variable nodes represented by lower-case indices (e.g., $i, j$) and factor nodes represented by upper-case indices (e.g., $I, J$), connected with each other through edges. All the variable nodes directly connected to a factor node (e.g., $I$) appear in the corresponding factor ($f_I$) and are represented by $\boldsymbol{x}_{N_I}$.

BP works by propagating local messages between neighbouring nodes. There are 2 types of messages: messages going from variables to factors and messages going from factors to variables. Those messages can be computed recursively by the following equations (discrete variables):

$$\mu_{j\to I}(x_j) = \prod_{J\in N_j\setminus\{I\}} \mu_{J\to j}(x_j) \qquad (S1)$$

$$\mu_{I\to i}(x_i) = \sum_{x_{N_I}\setminus\{i\}} f_I\left(x_{N_I}\right) \prod_{j\in N_I\setminus\{i\}} \mu_{j\to I} \qquad (S2)$$

$N_j$ is the set of all factors directly connected to variable $j$. Given that the FG is derived from a DAG, the factor $f_I(x_{N_I})$ is a conditional probability and represents the strength of the connection between the variables in $\boldsymbol{x}_{N_I}$. Eq. ($S1$) simply means that the message from a variable $j$ to a factor $I$ is the product of all the messages arriving at $j$, except for the message sent from this particular factor $I$. Similarly, a message sent from a factor $I$ to a variable $i$ is the product of all the messages arriving at $I$, except for the message sent from $i$, weighed by the factor $f_I$ and marginalized over all the other variables in $x_{N_I}$ (($S2$)). Once all messages have been propagated in both directions (one pass is sufficient for convergence to the correct posterior, when there are no loops), posterior probabilities can be computed as follows:

$$b_i(x_i) = \frac{1}{Z}\prod_{I \in N_i} \mu_{I \to i}(x_i) \qquad (S3)$$

where $Z$ is the normalization constant. Consequently, posteriors can be calculated as the products of all the messages arriving at each node.

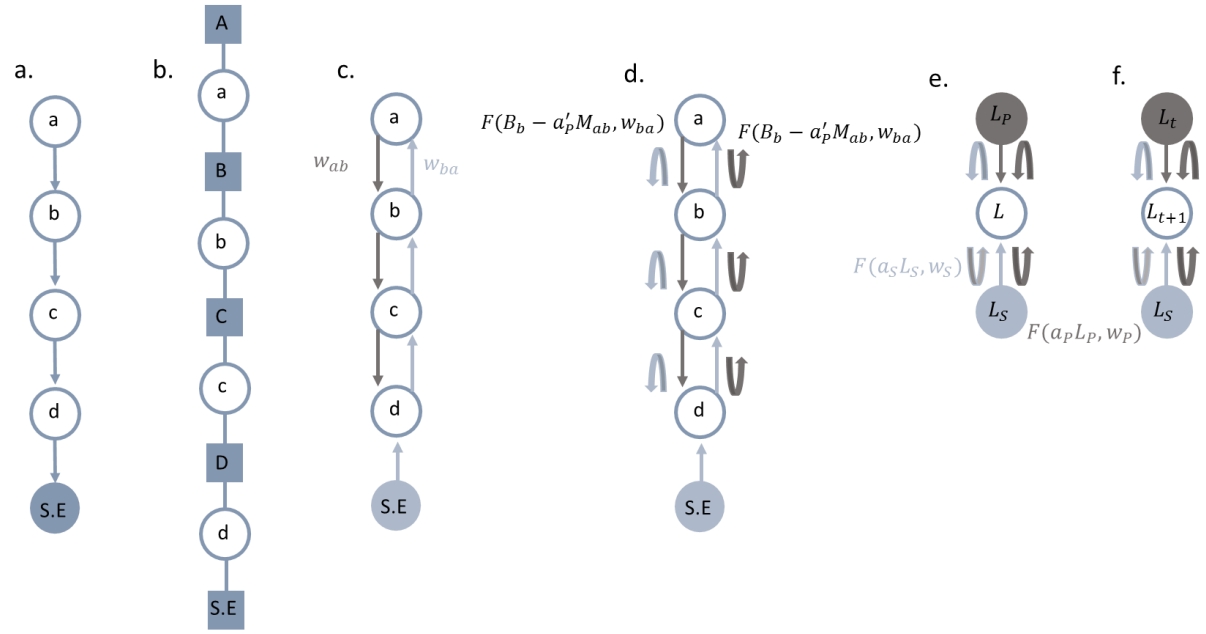


***Figure S1**: **BP and discrete model.** **(a.):** The generative model as Bayesian network. **(b.):** The factor graph. **(c.):** Belief propagation. **(d.):** Circular inference. **(e.):** A simplified model for circular inference [25] . **(f.):** The discrete model.*

In a previous paper, Jardri and Denève drew an analogy between BP and neural processing in recurrent, hierarchical networks (**Figure S1c;** [22]). They showed that, when considering pairwise graphs and binary variables and taking the log-ratios, factors can be omitted and BP recursive equations can be rewritten as follow:

$$B_i = \sum_j M_{j \to i} \qquad (S4)$$

$$M_{j \to i} = F\left(B_j - M_{i \to j}, w_{ji}^1, w_{ji}^0\right) \qquad (S5)$$

In $(S4)$ and $(S5)$, $M_{j \to i} = \log\left(\frac{\mu(x_i=1)}{\mu(x_i=0)}\right)$ and $B_i = \log\left(\frac{b_i(x_i=1)}{b_i(x_i=0)}\right)$. $F()$ on the other hand corresponds to a sigmoid function, defined in the following way:

$$F(B, w^1, w^0) = \log\left(\frac{w^1 e^B + w^0}{(1-w^1)e^B + (1-w^0)}\right) \qquad (S6)$$

$w_{ji}^1$ and $w_{ji}^0$ correspond to the strength of the $(j \to i)$ connection and are defined as the following conditional probabilities:

$$w_{ji}^1 = P\big(x_i = 1 | x_j = 1\big), \qquad w_{ji}^0 = P\big(x_i = 1 | x_j = 0\big) \qquad (S7)$$

In equation $(S5)$, the message is defined as a function of the belief of the variable that sends the message, minus the message sent in the opposite direction. This correction (similar to the correction in initial equations $(S1), (S2)$) is crucial, since it prevents the formation of circularity, i.e. reverberation of messages which are counted many times. Failure of the corrective mechanism (hypothesized to be implemented by inhibition) to subtract efficiently all the redundant information (e.g. due to E/I imbalance in favour of excitation) leads to circular inference (CI) (**Figure S1d**), in which case equation $(S5)$ is written as follows:

$$M_{j\to i} = F\big(B_j - a_S' M_{i\to j}, w_{ji}^1, w_{ji}^0\big)\,, if\ i\ is\ below\ j \qquad (S9)$$

$$M_{j\to i} = F\big(B_j - a_P' M_{i\to j}, w_{ji}^1, w_{ji}^0\big)\,, if\ i\ is\ above\ j \qquad (S10)$$

Parameters $a_S'$ and $a_P'$ quantify the strength of the loops (climbing and descending respectively) per layer, in other words what part of each message gets reverberated within a single connection and take values between 0 and 1 (1 corresponds to exact inference, i.e. E/I balance).

Although the outcome of BP (and of CI) can be calculated recursively via eq. $(S4), (S5)$ $((S4), (S9), (S10)$ for CI), the presence of the non-linear $F$ term makes it difficult to derive general and handy closed form solutions. For this reason, a simplification of the BP scheme was recently suggested, that keeps all the essential features of CI while presenting it in a more operational form (**Figure S1e;** [25]). In particular, hierarchy is reduced to a 3-node graph, comprising the prior information, the sensory information and the variable whose posterior needs to be calculated. Because the total effect of the loops depends on the number of possible reverberations (i.e. the number of layers in the hierarchy), parameters $a_S'$ and $a_P'$ where replaced by $a_S$ and $a_P$, which can take any value above 0 (with 0 corresponding to exact inference; values below 0 are also possible and signify an increased strength of inhibition, **see Main Text**) and quantify the overall amplification of information in the whole original hierarchy. This simplified CI model can be written as follows:

$$L = F(L_S + R_S + R_P, w_S, 1 - w_S) + F(L_P + R_S + R_P, w_P, 1 - w_P) \qquad (S11)$$

where $L$ is the log-odds ratio, $L_S$ is the log-likelihood ratio, $L_P$ is the log-prior ratio, $w_S$ is the feedforward weight, $w_P$ is the feedback weight and $R_S, R_P$ are the reverberated terms, computed as follow:

$$R_S = F(a_S L_S, w_S, 1 - w_S) \quad (S12)$$

$$R_P = F(a_P L_P, w_P, 1 - w_P) \quad (S13)$$

For Gaussian noise, the log-likelihood ratio can be written in the following way:

$$L_S = \frac{2\mu_{int}}{\sigma_{int}^2} S_t = w_{int} S_t \quad (S14)$$

where: $S_t \sim N(\mu_{noise}, \sigma_{noise}^2)$

and $(\mu_{int}, \sigma_{int}, \mu_{noise}, \sigma_{noise})$ are parameters of the internal / noise model (**Main Text**).

In the discrete case, that corresponds to a random walk around $\mu_{noise}$.

In eq. $(S11)$ the first term represents the total bottom up information and the second term the total top down information. That comprises the original information (likelihood and prior respectively) along with the reverberated terms, that means the likelihood corrupting the prior and vice versa. Furthermore, an additional term is considered inside each component (a reverberated likelihood for the likelihood term and a reverberated prior for the prior term) that renders the 2 streams practically indistinguishable.

Both implementations of CI make the same assumptions and generate the same qualitative predictions. More particularly, they both hypothesize that redundant information is not fully removed from the propagated messages, leading to amplification of sensory evidences and/or priors [22,24]. Additionally, bottom-up information is corrupted by top-down information and vice versa, creating aberrant correlations between sensory evidence and priors ([24,25]; Leptourgos et al, submitted).

The model presented here is similar to the model described by eq. $(S11)$ but it comprises dynamics too, resulting in the following equation (**Figure S1f**):

$$L_{t+1} = F(L_S + R_S + R_P, w_S, 1 - w_S) + F\big(L_t + R_S + R_P, 1 - r_{off}, r_{on}\big) \quad (S15)$$

In $(S15)$ we take $dt = 1$ (discrete model). $L_t$ is the log-posterior ratio of variable $X$ (3D interpretation) at time t, which becomes the prior for the next time step. $r_{on}, r_{off}$ are the transition rates (see **Main Text**), which for simplicity and without loss of generality have been taken equal to the feedback weights ($r_{on} = w_P^0, r_{off} = 1 - w_P^1$; in $(S13)$ $w_P^1 = w_P, w_P^0 = 1 - w_P$).

Note that the 2 rates are not necessarily equal to each other. In that case, $F(0{,}1 - r_{off}, r_{on}) \neq 0$. In order to avoid reverberation of information in the absence of descending loops ($a_P = 0$), we defined $R_P$ as follows:

$$R_P = F(a_P L_t, 1 - r_{off}, r_{on}) - F(0{,}1 - r_{off}, r_{on}) \qquad (S16)$$

## From the discrete model to the continuous model

We now consider infinitesimally small time-steps ($dt \to 0$) and the corresponding sensory evidence $dS_t$. As mentioned in the **Main Text**, we assume $r_{on}, r_{off}$ and $a_P$ to be proportional to $dt$. Then equation $(S15)$ reads (after using $(S12)$, $(S14)$ and $(S16)$):

$$L_{t+dt} =$$

$$F\big(w_{int} dS_t + F(a_S w_{int} dS_t, w_S, 1 - w_S) + F(a_P dt L_t, 1 - r_{off} dt, r_{on} dt) - F(0{,}1 - r_{off} dt, r_{on} dt), w_S, 1 - w_S\big) +$$

$$+F\big(L_t + F(a_S w_{int} dS_t, w_S, 1 - w_S) + F(a_P dt L_t, 1 - r_{off} dt, r_{on} dt) - F(0{,}1 - r_{off} dt, r_{on} dt), 1 - r_{off dt}, r_{on} dt\big) =$$

$$= F_S + F_P \qquad (S17)$$

We linearize equation $(S17)$ by taking the Taylor expansion of each term and keeping only the first order terms.

The following general equalities hold (see **Demonstrations** below):

$$F(x, w, 1 - w) = (2w - 1)x + O(x^2), \qquad for\ x \to 0 \qquad (S18)$$

$$F(x, 1 - r_{off} dt, r_{on} dt) = x + dt\left(r_{on}(1 + e^{-x}) - r_{off}(1 + e^{x})\right) + O(dt^2) \qquad (S19)$$

$$\log(1 + x) = x + O(x^2), \qquad for\ x \to 0 \qquad (S20)$$

Using $(S18)$-$(S20)$ we get:

$$F(a_S w_{int} dS_t, w_S, 1 - w_S) = (2w_S - 1)a_S w_{int} dS_t + O(dt^2) \quad (S21)$$

$$F\left(a_P dt L_t, 1 - r_{off} dt, r_{on} dt\right)$$
$$= a_P dt L_t + dt\left(r_{on}(1 + e^{-a_P dt L_t}) - r_{off}(1 + e^{a_P dt L_t})\right)$$
$$+ O(dt^2) \quad (S22)$$

$$F\left(0,1 - r_{off} dt, r_{on} dt\right) = 2dt\left(r_{on} - r_{off}\right) + O(dt^2) \quad (S23)$$

From $(S22)$, using $(S20)$, one gets:

$$F\left(a_P dt L_t, 1 - r_{off} dt, r_{on} dt\right) = a_P dt L_t + 2dt\left(r_{on} - r_{off}\right) + O(dt^2) \quad (S24)$$

Using $(S18)$-$(S24)$, one concludes:

$$F_S = w_{int}(2w_S - 1)\left(1 + a_S(2w_S - 1)\right)dS_t + a_P dt(2w_S - 1)L_t + O(dt^2) \quad (S25)$$

$$F_P = L_t + a_S w_{int}(2w_S - 1)dS_t + a_P dt L_t + dt\left(r_{on}(1 + e^{-L_t}) - r_{off}(1 + e^{L_t})\right)$$
$$+ O(dt^2) \quad (S26)$$

Using $(S25)$ and $(S26)$, $(S17)$ becomes:

$$L_{t+dt} = L_t + 2w_S a_P dt L_t + dt\left(r_{on}(1 + e^{-L_t}) - r_{off}(1 + e^{L_t})\right)$$
$$+ w_{int}(2w_S - 1)(1 + 2w_S a_S)dS_t + O(dt^2) \quad (S27)$$

Then:

$$\frac{dL}{dt} = \frac{L_{t+dt} - L_t}{dt}$$
$$= 2w_S a_P L + (r_{on}e^{-L} - r_{off}e^{L}) + (r_{on} - r_{off}) + w_{int}(2w_S - 1)(1$$
$$+ 2w_S a_S)n_t$$

which is the equation of the **Main text** and $n_t$ is a Gaussian stochastic process.

## Derivation of $(S18)$

Using the definition of $F()$ (eq. $S(6)$) and the Taylor expansion: $e^x = 1 + x + O(x^2)$, one gets:

$$F(x, w, 1-w) = \log\big(w\big(1 + x + O(x^2)\big) + 1 - w\big) - \log\Big((1-w)\big(1 + x + O(x^2)\big) + w\Big)$$
$$= \log\big(1 + wx + O(x^2)\big) - \log\big(1 + x - wx + O(x^2)\big)$$

Using $(S20)$, we get $(S18)$

## Derivation of eq.$(S19)$

Using the definition of $F()$ (eq. $S(6)$):

$$F\big(x, 1 - r_{off}dt, r_{on}dt\big) = x + \log\Big(1 + dt\big(-r_{off} + r_{on}e^{-x}\big)\Big) - \log\Big(1 + dt\big(r_{off}e^{x} - r_{on}\big)\Big)$$

Using $(S20)$, we get $(S19)$

## Bifurcation analysis

### *Pitchfork bifurcation*

We now consider the case when there is no stimulation.

If $r_{on} = r_{off} = r$, one gets:

$$f_L = \frac{df}{dL} = 2w_S a_P - r(e^{-L} + e^{L}) \qquad (S28)$$

A pitchfork bifurcation occurs at the maximum value of $a_P$, for which $f_L < 0, \forall L$ (and consequently $f$ is monotonically descending → there is one stable fixed point) (**Figure 4a,b**)

For $a_P < 0 \rightarrow f_L < 0, \forall L$

For $a_P \geq 0$, the first term in $(S28)$ is always non-negative, while the second term is always negative. Then we need: $2w_S a_P < r(e^{-L} + e^{L}), \forall L$

We know that:

$$\min\big(r(e^{-L} + e^{L})\big) = 2r \qquad (S29)$$

Then:

if $2w_S a_P < 2r \leftrightarrow a_P < \frac{r}{w_S}$, $f_L < 0, \forall L$

Consequently, a pitchfork bifurcation occurs when:

$$a_P^{Pf} = \frac{r}{w_S}$$

For $a_P < \frac{r}{w_S}$, there is one stable fixed point at $L = 0$. For $a_P > \frac{r}{w_S}$, there is an unstable fixed point at $L = 0$ and two stable fixed points given by the following equation:

$$f(L_{fp}) = 2w_S a_P L_{fp} + r(e^{-L_{fp}} - e^{L_{fp}}) = 0 \qquad (S30)$$

The different cases are presented in **Figure S2** (see also **Figure 3c**).

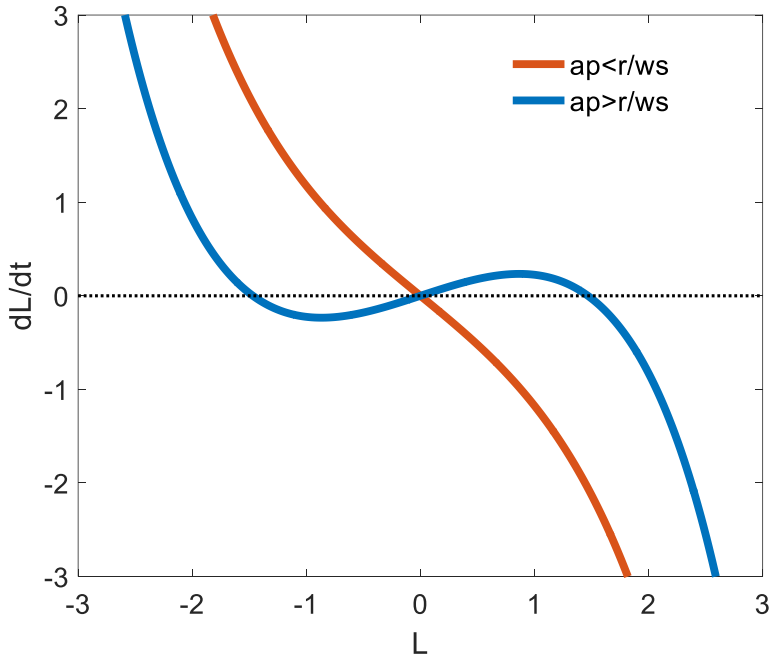


***Figure S2: The Pitchfork Bifurcation***

***Saddle-Node bifurcation***

If $r_{on} \neq r_{off}$, we get:

$$f_L = \frac{df}{dL} = 2w_S a_P - \left(r_{on}e^{-L} + r_{off}e^{L}\right) \qquad (S31)$$

If $a_P = 0$, there is one stable fixed point at $L = \log\left(\frac{r_{on}}{r_{off}}\right)$ (**MainText, Figure S3 (red)**).

If we extend the argument put forward in the case of the pitchfork bifurcation, we find that function $f$ stops being monotonically descending when:

$$a_P = \frac{A}{w_S} \tag{S32}$$

$$\text{where: } A = r_{on}\sqrt{\frac{r_{off}}{r_{on}}} + r_{off}\sqrt{\frac{r_{on}}{r_{off}}} \tag{S33}$$

Because of the asymmetry introduced by the 2 rates, that is not a bifurcation point (**Figure S3 (light blue)**). Instead, a Saddle-Node bifurcation occurs when one of the two local extrema crosses x-axis (**Figure 4c,d**). For the local extrema the following holds:

$$f_L = 0 \rightarrow L_{1,2} = \log\left(\frac{w_S a_P \mp \sqrt{w_S^2 a_P^2 - r_{on} r_{off}}}{r_{off}}\right) \tag{S34}$$

Then, one can calculate the value (of $a_P, r_{on}$ or $r_{off}$) at which SN bifurcation occurs simply by taking: $f(L_{1,2}) = 0$

All the fixed points can be calculated from the following equation (**Figure S3 (blue)**):

$$f(L_{fp}) = 2w_S a_P L_{fp} + r_{on}e^{-L_{fp}} - r_{off}e^{L_{fp}} + (r_{on} - r_{off}) = 0 \tag{S35}$$

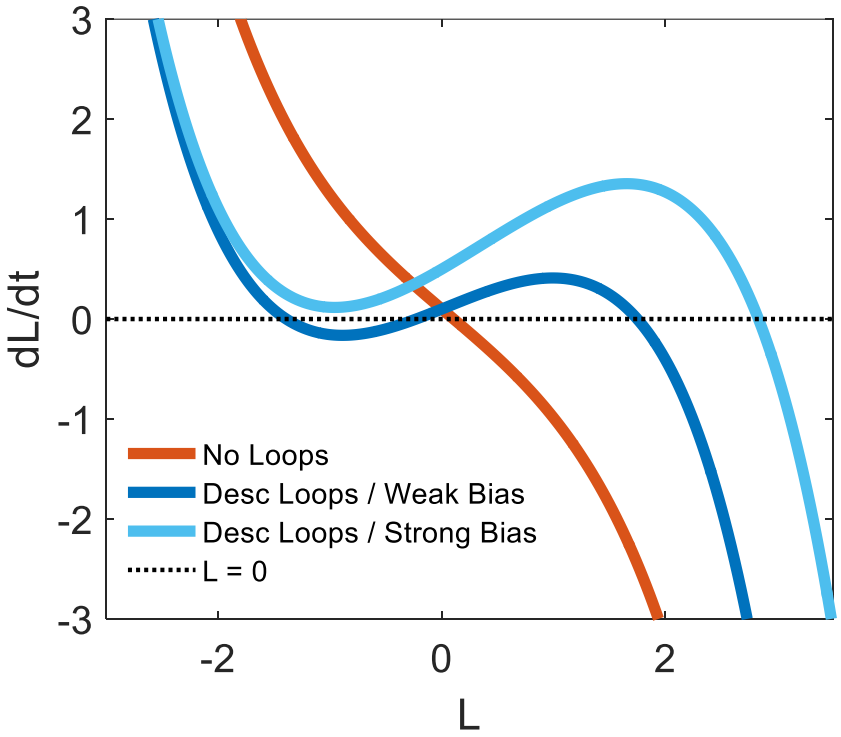


***Figure S3*: The Saddle-Node bifurcation.** *The light blue curve corresponds to an extremely strong implicit bias (the system gets stuck to interpretation 1)*

## MAP decision criterion with hysteresis

In the **Main Text** we present a model whose perceptual decisions are based on MAP decision criterion. For binary variables, that is equivalent to making decisions according to the sign of the Log-Posterior-Ratio. Although it maximises accuracy, in ambiguous contexts MAP decisions are extremely vulnerable to noise, resulting in phase duration histograms which have an exponential shape (see section about **Distribution of phase durations**).

Here, we propose an alternative decision criterion whose threshold changes depending on the current dominant percept (**Figure S4a**). According to this MAP criterion with hysteresis, when the system perceives SFA interpretation (corresponding to positive logits), the threshold is moved to a negative value ($L = -\varepsilon$) and inversely, when it perceives SFB interpretation (negative logits), the threshold becomes slightly positive ($L = \varepsilon$). When none of the 2 interpretations dominates (e.g. when the cube disappears), the threshold is fixed to 0. In short, parameter ε implements some hysteresis of the threshold that makes the system more robust against noise: a switch can only be triggered when there is substantial evidence for the opposite interpretation. In other words, it is equivalent to some form of conservatism, which prevents switching unless it is necessary.

Crucially, the *dynamical circular inference* model with hysteresis predicts positively skewed phase duration histograms, similar to those observed experimentally (**Figure S4b**). Whenever a switch occurs, the threshold jumps to its symmetrical value. Interestingly, even small values of ε are sufficient to prevent noise from causing an instantaneous switch.

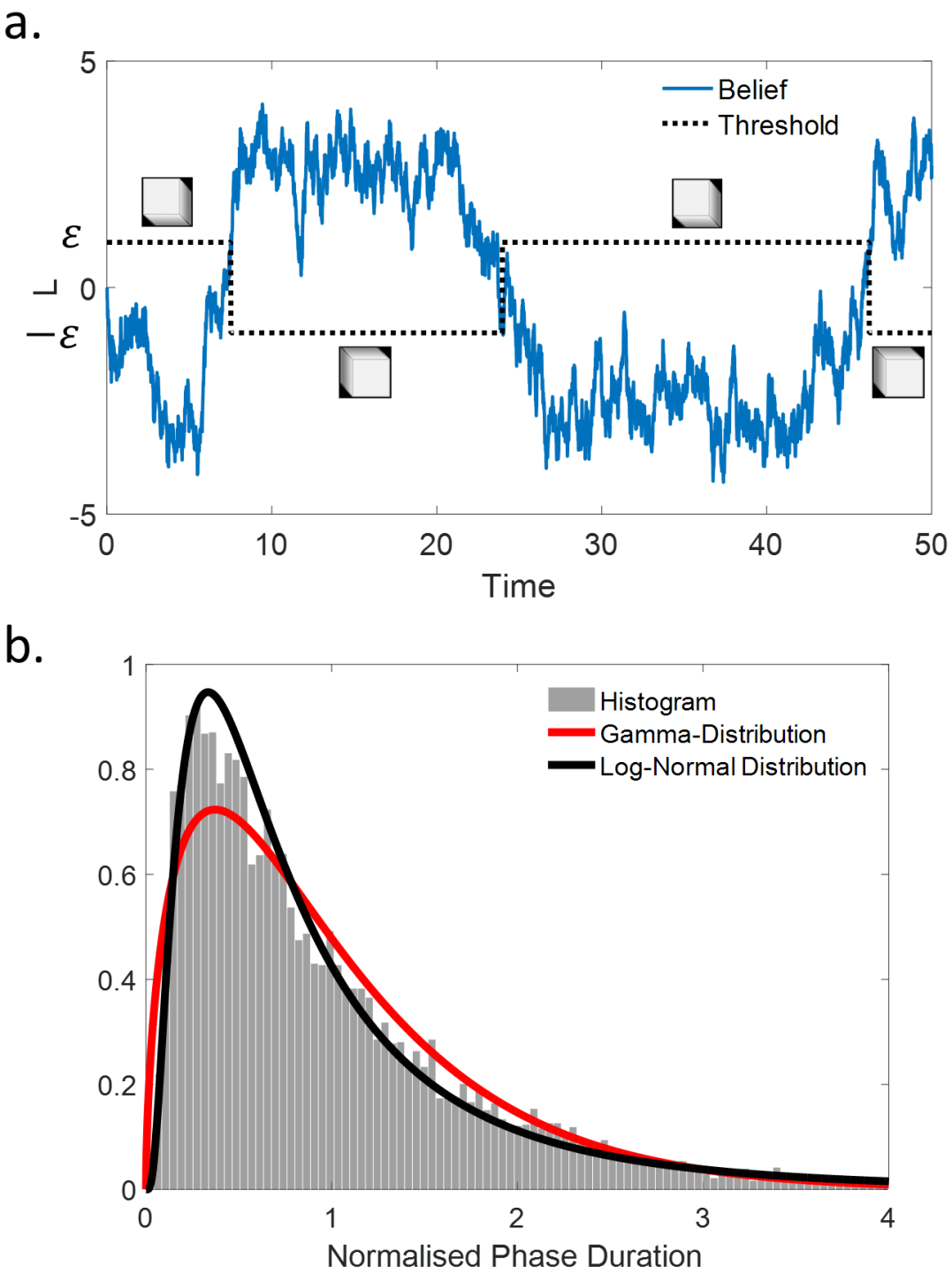


***Figure S4: MAP with hysteresis and histogram.*** ***(a.):*** *A switch occurs only if there is substantial evidence for the opposite interpretation (more conservative decision criterion).* ***(b.)*** *With this additional assumption, the model predicts gamma (or log-normal)-distributed phase duration histograms*

## Model with history-dependent transition rates

In the **Main Text,** we describe a model in which transition rates ($r_{on}, r_{off}$) are fixed and do not change in time (eq. (4)). This model predicts various features of bistable perception, including the stabilisation of perception when the stimulus is presented intermittently (see section on **Intermittent presentation, Main Text**). That is because descending loops generate a memory-like mechanism, which primes perception towards the most recent observation. More particularly, in the absence of any stimulation, the belief $L$ moves towards the closest attractor, depending on the value before the disappearance of the stimulus. If the attractor is far enough from the threshold, the model predicts stabilisation. Because all the parameters are time-independent, the energy landscape is fixed, meaning that the model can only generate monotonic relationships between SP and OFF-Duration (either increase or decrease). As a result, the model cannot predict an initial destabilisation of perception, occurring for short OFF-Durations, which turns into stabilisation for longer disappearances.

Here we show that an extension of the model, which further assumes history-dependent transition rates, can additionally predict a destabilisation of perception when the OFF-Duration is short. As in the original model, in the extended model the belief $L$ is given by the equation eq. (4):

$$\frac{dL}{dt} = 2w_S a_P L + (r_{on}(t)e^{-L} - r_{off}(t)e^{L}) + (r_{on}(t) - r_{off}(t)) + w_{int}(2w_S - 1)(1 + 2w_S a_S)n_t$$
$$= f(L)$$

Contrary to the model presented in the **Main Text**, the extended model posits that transition rates evolve in time, based on a standard leaky-integrator:

$$\tau \frac{dr_{on}}{dt} = -r_{on} + r_{on}^{+} + r_{on,B} \qquad (S35)$$

$$\tau \frac{dr_{off}}{dt} = -r_{off} + r_{off}^{+} + r_{off,B} \qquad (S36)$$

where $\tau$ is the time constant, $r_{on,B}, r_{off,B}$ are the baseline values of the 2 rates, while $r^{+}$ is an additional parameter having the following property:

$$\begin{cases} if\ X = 1\colon\ r_{off}^{+} > 0, r_{on}^{+} = 0 \\ if\ X = 0\colon\ r_{on}^{+} > 0, r_{off}^{+} = 0 \\ if\ none\ is\ dominant\colon r_{on}^{+} = \ r_{off}^{+} = 0 \end{cases} \qquad (S37)$$

($S37$) means that when $X = 1$ is the dominant interpretation, the transition rate from 0 to 1 ($r_{on}$) decreases towards its baseline ($r_{on,B}$), whereas the rate from 1 to 0 ($r_{off}$) increases towards

an upper state ($r^{+}_{off} + r_{off,B}$). The opposite is true if $X = 0$ is the dominant interpretation. Finally, when neither of the two is dominant (e.g. during OFF-Duration), both rates go back to their baseline level. As a result, this mechanism destabilises the currently dominant interpretation, like adaptation in mechanistic models. Crucially, it also affects the energy landscape (the prior, which is given by eq. (7), changes in time), making the attractor that corresponds to the dominant percept shallower and the "non-dominant" attractor deeper. Consequently, it can further generate non-monotonic relationships between SP and OFF-Duration.

**Figure S5** illustrates the predictions of the model ((a) without loops and (b) with loops), in case of $r_{on,B} = r_{off,B}$ (no bias), while **Figure S6** illustrates an explanation based on the phase portraits (**a,b**) and the evolutions of the rates (**c,d**). In both cases, for small values of OFF-Duration (<1s), SP decrease towards chance level (or even below; the fixed point is below 0, because $r_{off} > r_{on}$) while for larger values (>1s) they increase again (both $r_{on}$ and $r_{off}$ have reached their baseline value). Importantly, the loops have the same effect as in the original model, more particularly they affect the convergence point (0.5 if there are no loops, >0.5 if there are loops).

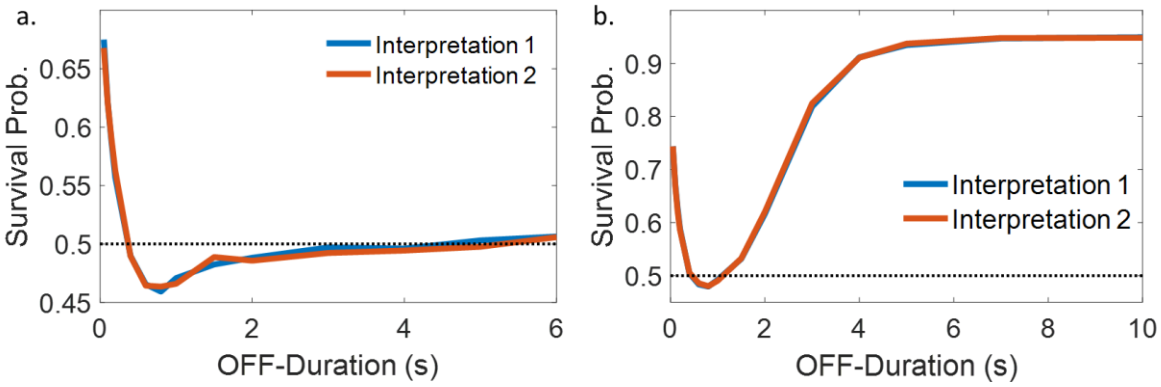


***Figure S5*: Intermittent Presentation with history-dependent transition rates (symmetrical case).** *In agreement with experimental evidence a model with history-dependent transition rates predicts destabilisation for short OFF-durations and stabilisation (if there are descending loops **(b.)**) for longer intervals. As in **Figure 6**, without loops the survival probabilities converge to 0.5 **(a.)**.*

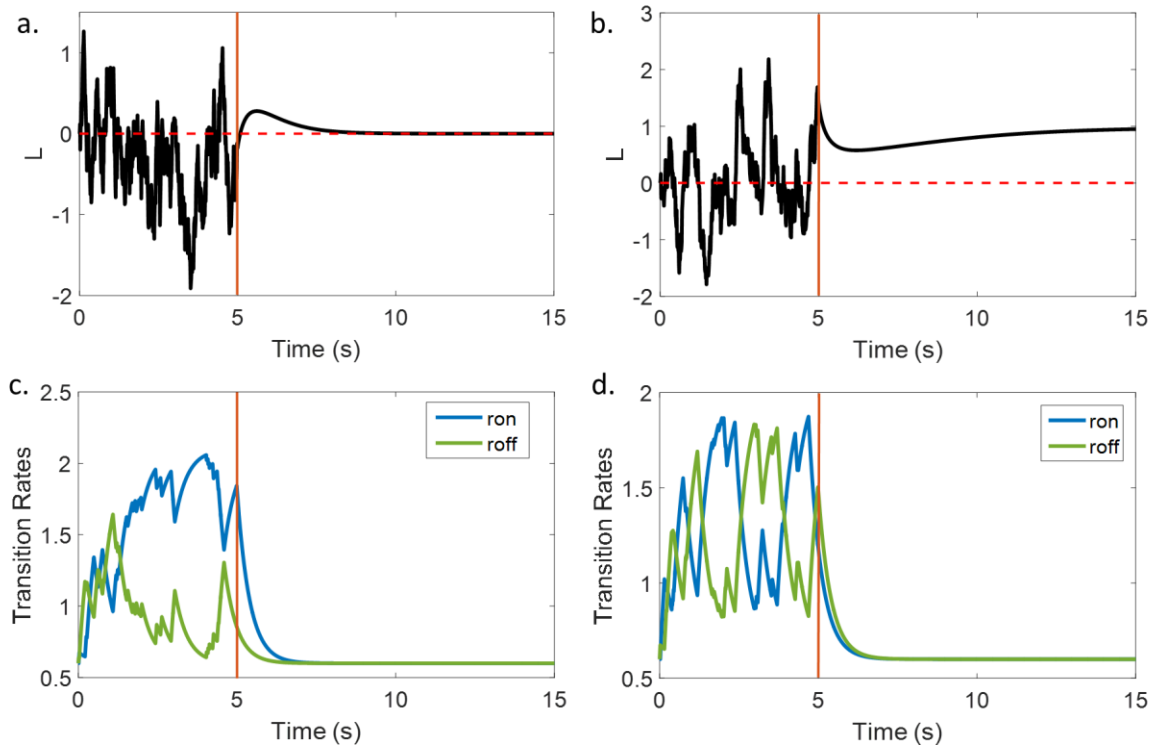


***Figure S6**: **An interpretation of the initial destabilisation in case of history-dependent transition rates**. Panels **(a.)** and **(b.)** illustrate examples of the belief when the stimulus is ON (0:5s) and when it disappears (t>5s) **((a.): No loops; (b.): loops)**, while panels **(c.)** and **(d.)** illustrate the evolution of the transition rates **((c.): No loops; (d.): loops)**. The time-dependent rates generate time-dependent prior / attractors. E.g. in **(a.)**, right after the disappearance of the stimulus (t=5s), $r_{on}$ is larger than $r_{off}$, resulting in a positive prior. In the absence of stimulation, both rates go to their baseline value (the two baseline values are equal – symmetrical case). Consequently, the difference between the rates becomes smaller, resulting in an attractor that moves towards 0.5. That causes a non-monotonic evolution of the belief for t>5, which gives rise to the survival probabilities in **Figure S5 (a.)**.*

## References


1. Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nat Neurosci. 2002;5: 598–604. doi:10.1038/nn858

2. Notredame C-E, Pins D, Denève S, Jardri R. What visual illusions teach us about schizophrenia. Front Integr Neurosci. 2014;8: 1–16. doi:10.3389/fnint.2014.00063

3. Arnold DH. Why is binocular rivalry uncommon ? Discrepant monocular images in the real world. Front Hum Neurosci. 2011;5: 1–7. doi:10.3389/fnhum.2011.00116

4. Blake R, Logothetis NK. Visual Competition. Nat Rev Neurosci. 2002;3: 1–11. doi:10.1038/nrn701

5. Blake R. A Neural Theory of Binocular Rivalry. Psychol Rev. 1989;96: 145–167.

6. Lago-Fernandez LF, Deco G. A model of binocular rivalry based on competition in IT. Neurocomputing. 2002;44–46: 503–507.

7. Laing CR, Chow CC. A Spiking Neuron Model for Binocular Rivalry. J Comput Neurosci. 2002;12: 39–53.

8. Wilson HR. Computational evidence for a rivalry hierarchy in vision. Proc Natl Acad Sci U S A. 2003;100: 14499–503. doi:10.1073/pnas.2333622100

9. Noest AJ, Ee R Van, Nijs MM, Wezel RJA Van. Percept-choice sequences driven by interrupted ambiguous stimuli : A low-level neural model. J Vis. 2007;7: 1–14. doi:10.1167/7.8.10.Introduction

10. Wilson HR. Minimal physiological conditions for binocular rivalry and rivalry memory. Vision Res. 2007;47: 2741–2750. doi:10.1016/j.visres.2007.07.007

11. Moreno-Bote R, Rinzel J, Rubin N. Noise-Induced Alternations in an Attractor Network Model of Perceptual Bistability. J Neurophysiol. 2007;98: 1125–1139. doi:10.1152/jn.00116.2007

12. Shpiro A, Moreno-Bote R, Rubin N, Rinzel J. Balance between noise and adaptation in competition models of perceptual bistability. J Comput Neurosci. 2009;27: 37–54. doi:10.1007/s10827-008-0125-3

13. Panagiotaropoulos TI, Kapoor V, Logothetis NK, Deco G. A Common Neurodynamical Mechanism Could Mediate Externally Induced and Intrinsically Generated Transitions in Visual Awareness. PLoS One. 2013;8. doi:10.1371/journal.pone.0053833

14. Huguet G, Rinzel J, Hupé J. Noise and adaptation in multistable perception: Noise drives when to switch, adaptation determines percept choice. J Vis. 2014;14: 1–24. doi:10.1167/14.3.19.doi

15. Brascamp J, Sterzer P, Blake R, Knapen T. Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. Annu Rev Psychol. 2018;69: 1–27.

16. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: An epistemological review. Cognition. 2008;108: 687–701. doi:10.1016/j.cognition.2008.05.010

17. Weilnhammer V, Stuke H, Hesselmann G, Sterzer P, Schmack K. A predictive coding account of bistable perception - a model-based fMRI study. PLoS Comput Biol. 2017;13: 1–21. doi:10.1371/journal.pcbi.1005536

18. Sundareswara R, Schrater P. Perceptual multistability predicted by search model for Bayesian decisions. J Vis. 2008;8: 1–19. doi:10.1167/8.5.12.Introduction

19. Reichert D, Seriès P, Storkey A. Neuronal adaptation for sampling-based probabilistic inference in perceptual bistability. Adv Neural Inf .... 2011; 1–9. Available: http://papers.nips.cc/paper/4404-neuronal-adaptation-for-sampling-based-probabilistic-inference-in-perceptual-bistability

20. Gershman SJ, Vul E, Tenenbaum JB. Multistability and Perceptual Inference. Neural Comput. 2012;24: 1–24.

21. Bishop C. Pattern Recognition and Machine Learning. Springer; 2006.

22. Jardri R, Denève S. Circular inferences in schizophrenia. Brain. 2013;136: 3227–41. doi:10.1093/brain/awt257

23. Deneve S, Jardri R. Circular inference: Mistaken belief, misplaced trust. Curr Opin Behav Sci. 2016;11: 40–48. doi:10.1016/j.cobeha.2016.04.001

24. Leptourgos P, Denève S, Jardri R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Curr Opin Neurobiol. 2017;46: 154–161. doi:10.1016/j.conb.2017.08.012

25. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference in schizophrenia. Nat Commun. 2017;8: 14218. doi:10.1038/ncomms14218

26. Friston K. Hierarchical Models in the Brain. PLoS Comput Biol. 2008;4. doi:10.1371/journal.pcbi.1000211

27. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci. 2014;8: 1–24. doi:10.3389/fnhum.2014.00825

28. Finlayson NJ, Zhang X, Golomb JD. Differential patterns of 2D location versus depth decoding along the visual hierarchy. Neuroimage. Elsevier; 2017;147: 507–516. doi:10.1016/j.neuroimage.2016.12.039

29. Felleman DJ, Van Essen DC. Distributed hierachical processing in the primate cerebral

cortex. Cereb Cortex. 1991;1: 1–47. doi:10.1093/cercor/1.1.1

30. Lee TS, D.Mumford. Hierarchical bayesian inference in the visual cortex. J Opt Soc Am A. 2003;20: 1434–1448.

31. Braun J, Mattia M. Attractors and noise: Twin drivers of decisions and multistability. Neuroimage. Elsevier Inc.; 2010;52: 740–751. doi:10.1016/j.neuroimage.2009.12.126

32. Moreno-Bote R, Knill DC, Pouget A. Bayesian sampling in visual perception. Proc Natl Acad Sci U S A. 2011;108: 12491–12496. doi:10.1073/pnas.1101430108

33. Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Res. 1998;38: 2817–2832. doi:10.1016/S0042-6989(97)00438-0

34. Douglas RJ, Koch C, Mahowald M, Martin KAC, Suarez HH. Recurrent excitation in neocortical circuits. Science (80- ). 1995;269: 981–985. doi:10.1126/science.7638624

35. Levelt WJM. The Alternation Process in Binocular Rivalry. Br J Psychol. 1966;57: 225–238. doi:10.1111/j.2044-8295.1966.tb01023.x

36. Klink PC, van Ee R, van Wezel RJ a. General validity of Levelt's propositions reveals common computational mechanisms for visual rivalry. PLoS One. 2008;3: e3473. doi:10.1371/journal.pone.0003473

37. Deneve S. Bayesian Spiking Neurons I: Inference. Neural Comput. 2008;20: 91–117. doi:10.1162/neco.2008.20.1.91

38. Shpiro A, Curtu R, Rinzel J, Rubin N. Dynamical Characteristics Common to Neuronal Competition Models. J Neurophysiol. 2007;97: 462–473. doi:10.1152/jn.00604.2006

39. Brascamp JW, Klink PC, Levelt WJM. The "laws" of binocular rivalry: 50 years of Levelt's propositions. Vision Res. 2015;109: 20–37. doi:10.1016/j.visres.2015.02.019

40. Moreno-Bote R, Shpiro A, Rinzel J, Rubin N. Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. J Vis. 2010;10: 1–1. doi:10.1167/10.11.1

41. Walker P. Stochastic properties of binocular rivalry alternations. Percept Psychophys. 1975;18: 467–473.

42. Lehky SR. Binocular rivalry is not chaotic. Proc R Soc London B Biol Sci. 1995;259: 71–76.

43. Nawrot M, Blake R. Neural Integration of Information Specifying Structure from Stereopsis and Motion. Science (80- ). 1989;244: 716–718.

44. Pastukhov A, Braun J. Cumulative history quantifies the role of neural adaptation in multistable perception. J Vis. 2011;11: 12–12. doi:10.1167/11.10.12

45. Orbach J, Ehrlich D, Heath HA. Reversibility of the Necker Cube: I. An examination of

the concept of “satiation of orientation.” Percept Mot Skills. 1963;17: 439–458. doi:10.2466/pms.1963.17.2.439

46. Leopold DA, Wilke M, Maier A, Logothetis NK. Stable perception of visually ambiguous patterns. Nat Neurosci. 2002;5: 605–609. doi:10.1038/nn851

47. Pearson J, Brascamp J. Sensory memory for ambiguous vision. Trends Cogn Sci. 2008;12: 334–41. doi:10.1016/j.tics.2008.05.006

48. Maier A, Wilke M, Logothetis NK, Leopold DA. Perception of Temporally Interleaved Ambiguous Patterns. Curr Biol. 2003;13: 1076–1085. doi:10.1016/S

49. Sterzer P, Rees G. A Neural Basis for Percept Stabilization in Binocular Rivalry. J Cogn Neurosci. 2008;20: 389–399. doi:10.1162/jocn.2008.20039

50. Kornmeier J, Ehm W, Bigalke H, Bach M. Discontinuous presentation of ambiguous figures: How interstimulus-interval durations affect reversal dynamics and ERPs. Psychophysiology. 2007;44: 552–560. doi:10.1111/j.1469-8986.2007.00525.x

51. Levelt WJM. Note on the distribution of dominance times in binocular rivalry. Br J Psychol. 1967;58: 143–145.

52. Zhou YH, Gao JB, White KD, Merk I, Yao K. Perceptual dominance time distributions in multistable visual perception. Biol Cybern. 2004;90: 256–263. doi:10.1007/s00422-004-0472-8

53. Gigante G, Mattia M, Braun J, Del Giudice P. Bistable perception modeled as competing stochastic integrations at two levels. PLoS Comput Biol. 2009;5: 1–9. doi:10.1371/journal.pcbi.1000430

54. Brascamp JW, van Ee R, Pestman WR, van den Berg A V. Distributions of alternation rates in various forms of bistable perception. J Vis. 2005;5: 287–98. doi:10:1167/5.4.1

55. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annu Rev Psychol. 2004;55: 271–304. doi:10.1146/annurev.psych.55.090902.142005

56. Albert S, Schmack K, Sterzer P, Schneider G. A hierarchical stochastic model for bistable perception. PLoS Computational Biology. 2017. doi:10.1371/journal.pcbi.1005856

57. Dayan P. A Hierarchical Model of Binocular Rivalry. Neural Comput. 1998;10: 1119–1135. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0032111193&partnerID=40&md5=2220a1a71a4cfd3e9066c68547e73897

58. Jardri R, Hugdahl K, Hughes M, Brunelin J, Waters F, Alderson-Day B, et al. Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? Schizophr Bull. 2016;42: 1124–1134. doi:10.1093/schbul/sbw075

59. Brascamp J, Sohn H, Lee S-H, Blake R. A monocular contribution to stimulus rivalry. Proc

Natl Acad Sci. 2013;110: 8337–8344. doi:10.1073/pnas.1305393110

60. Brascamp J, Blake R, Knapen T. Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. Nat Neurosci. Nature Publishing Group; 2015;18: 1672–1678. doi:10.1038/nn.4130

61. Lumer ED, Frsiton KJ, Rees G. Neural Correlates of Perceptual Rivalry in the Human Brain. Science (80- ). 1998;280: 1930–1934. doi:10.1126/science.280.5371.1930

62. Sterzer P, Kleinschmidt A. A neural basis for inference in perceptual ambiguity. Proc Natl Acad Sci U S A. 2007;104: 323–8. doi:10.1073/pnas.0609006104

63. Van Loon AM, Knapen T, Scholte HS, St. John-Saaltink E, Donner TH, Lamme VAF. GABA shapes the dynamics of bistable perception. Curr Biol. Elsevier Ltd; 2013;23: 823–827. doi:10.1016/j.cub.2013.03.067

64. Schmack K, Gòmez-Carrillo de Castro A, Rothkirch M, Sekutowicz M, Rössler H, Haynes J-D, et al. Delusions and the role of beliefs in perceptual inference. J Neurosci. 2013;33: 13701–12. doi:10.1523/JNEUROSCI.1778-13.2013

65. Waters F, Blom JD, Dang-Vu TT, Cheyne AJ, Alderson-Day B, Woodruff P, et al. What Is the Link Between Hallucinations, Dreams, and Hypnagogic-Hypnopompic Experiences? Schizophr Bull. 2016;42: 1098–1109. doi:10.1093/schbul/sbw076

66. Alderson-Day B, Lima CF, Evans S, Krishnan S, Shanmugalingam P, Fernyhough C, et al. Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. Brain. 2017;140: 2475–2489. doi:10.1093/brain/awx206

67. Baumeister D, Sedgwick O, Howes O, Peters E. Auditory verbal hallucinations and continuum models of psychosis: A systematic review of the healthy voice-hearer literature. Clin Psychol Rev. Elsevier B.V.; 2017;51: 125–141. doi:10.1016/j.cpr.2016.10.010

68. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science (80- ). 2017;357: 596–600. doi:10.1126/science.aan3458

69. Loh M, Rolls ET, Deco G. A dynamical systems hypothesis of schizophrenia. PLoS Comput Biol. 2007;3: 2255–2265. doi:10.1371/journal.pcbi.0030228

70. Rolls ET, Deco G. A computational neuroscience approach to schizophrenia and its onset. Neurosci Biobehav Rev. Elsevier Ltd; 2011;35: 1644–1653. doi:10.1016/j.neubiorev.2010.09.001

71. Lucas-Meunier E, Monier C, Amar M, Baux G, Frégnac Y, Fossier P. Involvement of nicotinic and muscarinic receptors in the endogenous cholinergic modulation of the balance between excitation and inhibition in the young rat visual cortex. Cereb Cortex. 2009;19: 2411–2427. doi:10.1093/cercor/bhn258

72. Moreau WA, Amar M, Le Roux N, Morel N, Fossier P. Serotoninergic fine-tuning of the excitation-inhibition balance in rat visual cortical networks. Cereb Cortex. 2010;20: 456–467. doi:10.1093/cercor/bhp114

73. Wehr M, Zador AM. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. Nature. 2003;426: 442–446. doi:10.1038/nature02116

74. Okun M, Lampl I. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nat Neurosci. 2008;11: 535–537. doi:10.1038/nn.2105

75. Xue M, Atallah B V., Scanziani M. Equalizing excitation–inhibition ratios across visual cortical neurons. Nature. Nature Publishing Group; 2014;511: 596–600. doi:10.1038/nature13321

# Chapter 4

## Intermittent presentation of ambiguous stimuli: more evidence for circular inference in bistable perception

In preparation for publication as:

Leptourgos P., Eck M., Devignes C., Deneve S. Jardri R. (in prep.). Intermittent presentation of ambiguous stimuli: more evidence for circular inference in bistable perception

## Abstract

Bistable perception is characterized by spontaneous switches of the perceived interpretation of the sensory stimulus, occurring every few seconds and generating uncorrelated sequences of dominance durations. Nevertheless, when bistable stimuli are intermittently presented, a specific pattern occurs with an increased number of switches for short blank durations, followed by a stabilisation phase for longer intervals. Although intermittent bistable perception has been extensively studied, a functional interpretation of the results is still lacking. Here we propose an experimental methodology able to overcome the shortcomings of previous studies. We used different versions of the Necker cube (normal and tilted), accounting for well-known perceptual biases. Sequences of those stimuli were discontinuously presented, separated by randomized, blank intervals of variable duration. We measured the survival probability for each interpretation and for all the blank intervals, reconstructing the entire stabilization curve. In complement to the typical stabilization pattern, we also evidenced a strong effect of the bias. Importantly, our results were well captured by a new computational framework called dynamical circular inference. This model suggests that descending loops in the cortical hierarchy are a fundamental property of the human brain, affecting the way we perceive the world.

## Introduction

Bistable perception constitutes one of the few manifestations of a clear dissociation between the stimulation of a sensory organ and the resulting awareness, offering a unique window into the computational and neural processes underlying perceptual decision making and perceptual consciousness [1]. Continuous display of ambiguous figures generates spontaneous switches of the perceived interpretations, which results in a sequence of random and largely uncorrelated phase durations (intervals of persistence of a single interpretation) [2,3]. This observation led to the consensus that bistability is a memoryless process, mainly driven by noisy evidence [4–6].

Despite the striking evidence for the dominant role of noise in triggering switches, the view that the effect of recent history can be a-priori neglected without further inquiry has been contested [7]. One of the main challenges came from the observation that discontinuous presentation of an ambiguous stimulus profoundly alters participants' behaviour [8]. In their seminal study, Orbach and colleagues found that brief disappearance of the stimulus (up to ~500ms) results in a destabilisation of the perceived interpretation, as denoted by the increase in the alternation rate; on the contrary, longer OFF-Durations led to a gradual decrease in the number of switches, which became almost zero for intervals larger than ~1.4s [9]. Those researchers attributed this dual effect of the blank interval to a form of fatigue ("satiation of orientation"), which builds-up when the stimulus is ON and decays when the stimulus is OFF. Most importantly, this study triggered a lot of discussions regarding the role of adaptation and the implementation of perceptual memory in bistable perception [10–12].

Regardless of the huge impact this work had on the field, it suffered from a number of methodological issues. First, given the discontinuous presentation of the stimulus, the use of the reversal rate as the main variable is questionable: Reversal rate is defined as the number of switches during an interval and it can be affected by the overall time of exposure to the stimulus [8], while it also confounds switches due to disappearance with switches due to prolonged presentation (OFF-Duration vs ON-Duration effects; [13]). Second, reversal rate accounts for both types of switches (interpretation 1 to interpretation 2 and vice-versa), completely ignoring potential differences between the stimulus' interpretations (e.g. eye-dominance in binocular rivalry or "seeing-from-above" bias in the Necker cube; [14,15]; see also **Chapters 2-3** of the present thesis). In the same vein, those studies attributed stabilisation only to acute biases (short-term perceptual memory), completely ignoring the complementary effects of the chronic biases mentioned before (long-term priors) [16]. Finally, another (minor) shortcoming

corresponds to the lack of within-block randomization of the inter-stimulus interval (ISI; OFF-Duration), which results in predictable response timings and could have led to contamination of the results due to expectation or lack of attention [8].

Aside from these methodological issues, intermittent presentation has proved to be a difficult problem to solve from a computational point of view (especially the distinction between short and long intervals as well as the link with continuous presentation; [3,12]). In previous work (**chapter 3** of the present thesis), we proposed a normative (Bayesian) framework for bistable perception based on the notion of *circular inference* [17]. More specifically, we argued that descending loops, a form of aberrant priors' amplification, underlie the rich phenomenology of bistability: priors get trapped and over-counted, giving rise to a positive feedback that pushes beliefs towards more extreme values. From a mechanistic point of view, descending loops form a bistable attractor [5]. Crucially, descending loops (if strong enough) also implement a type of perceptual memory, causing stabilisation during intermittent presentation.

Here we use a novel, improved experimental methodology (briefly introduced in **Chapter 3**) to test the qualitative predictions of a *dynamical circular inference* model regarding intermittent presentation. We present two experiments, based on different versions of the Necker cube (normal and tilted cube). Our results can be well replicated (qualitatively) by a model with positive descending loops and add further support to our claim that circularity (and especially descending loops) makes up a constitutive element of our (perceptual) system, playing an important role in perceptual decision making in healthy populations. Besides, we propose that our method could form a standard technique, further used to understand both normal and pathological perceptual awareness.

## Methods

Our goal was to study the effect of the disappearance of the ambiguous stimulus within a range of OFF-Durations. We ran two different experiments, to further account for well documented chronic biases [16,18].

## Participants

10 participants took part in each experiment. The two samples were independent. Participants were healthy volunteers meeting the following inclusion criteria: age > 18 years, provision of informed consent, normal or corrected-to-normal near visual acuity, no current medical history of neurological or psychiatric disorders, and no current or recent use of psychotropic medication or toxic drugs. **Table 1** provides additional information about the two samples.

| | Normal Cube | Tilted Cube |
|---|---|---|
| Sample Size | 10 | 10 |
| Age (mean (sd)) | 34.5 (9.6) | 20.9 (0.6) |
| Sex (ratio m:f) | 6:4 | 4:6 |
| Reaction Time (s) (mean (sd)) | 0.62 (0.07) | 0.77 (0.17) |

***Table 1: Participants and reaction times.***

## Apparatus

Both experiments took place in a dark room. The stimuli were displayed either on a 15-inch LED computer screen with a resolution of 1920x1080 pixels (experiment 1) or a 17-inch LED computer screen with a resolution of 1280x1024 pixels (experiment 2), both at 60 Hz. Responses were collected using either a keyboard (experiment 1) or a mouse (experiment 2). The background colour of the screen was black. The viewing distance was 60 cm and a headrest secured the position of the head. The experiments were implemented in MATLAB v. 2015b (MathWorks, Natich, MA), using Psychtoolbox v. 3.0.12.

## Stimuli

The stimuli were 200x200 pixels (experiment 1) or 316x316 pixels (experiment 2), light-gray Necker cubes. In each experiment we used a different version of the cube, expected to induce different implicit preferences to the participants [14]. In experiment 1, we used a normal cube (**Figure 1a; left**), known to be an asymmetrical stimulus (people have on average a preference for the "*Seen From Above*" (SFA) interpretation, as compared to the "*Seen From Below*" (SFB) interpretation; see also **Chapters 2 and 3** of the present thesis). Contrariwise, a tilted cube was used in experiment 2 (**Figure 1a; right**), known to be on average symmetrical (no biases, except for potential idiosyncratic preferences). Stimuli were presented in the middle of the screen. A circular fixation point was added in the middle of each cube, to guide participants' gaze.

## Experimental paradigm

As in previous studies, here we used a standard intermittent presentation paradigm (**Figure 1b**) [8,11,19,20]. Both experiments consisted of 12 blocks, each constituting a sequence of ON (stimulus is present) and OFF-Durations (stimulus is absent). Each ON-Duration was followed by an OFF-Duration, except for the last one in each block. A single trial contained an OFF-Duration, as well as the ON-Durations before and after. There were 64 ON-Durations (and 63 OFF-Durations) per block, making up a set of 63 trials per block. Blocks were separated by 10s black-screen breaks, while 2 additional breaks of non-predefined duration were also possible after the fourth and the eighth block.

Participants were instructed to report their dominant percept, as quickly as possible, every time the stimulus reappeared on the screen. Responses were given by pressing the relevant button (Right: Seen From Above/Right; Left: Seen From Below/Left). Contrary to other studies, ON-Durations did not have a standard length, instead the stimulus disappeared right after the first button-press. Consequently, participants could give only one response per ON-Duration (first impression), solving the problem of switches during stimulus' presentation, but also avoiding confounding post-decision effects (e.g. accumulation of evidence after the button-press) and making the task more interactive (increasing the sense of involvement and the levels of attention).

Since we were interested in reconstructing the whole stabilisation curve (see **Figure 2**), we used a set of 9 OFF-Durations, different for each experiment. In experiment 1, blank periods were between 100 ms and 900 ms, with a 100 ms increment, while in experiment 2, they took the following values: {50, 100, 150, 300, 450, 600, 750, 900, 1050 [ms]}. The second set allowed for a higher resolution for short intervals (where an inversion from de-stabilisation to stabilisation usually occurs) but also a wider range of intervals. Crucially, blank intervals were randomized within each block, with 7 repetitions per block per interval. This gave a total number of 84 repetitions per OFF-Duration per experiment.

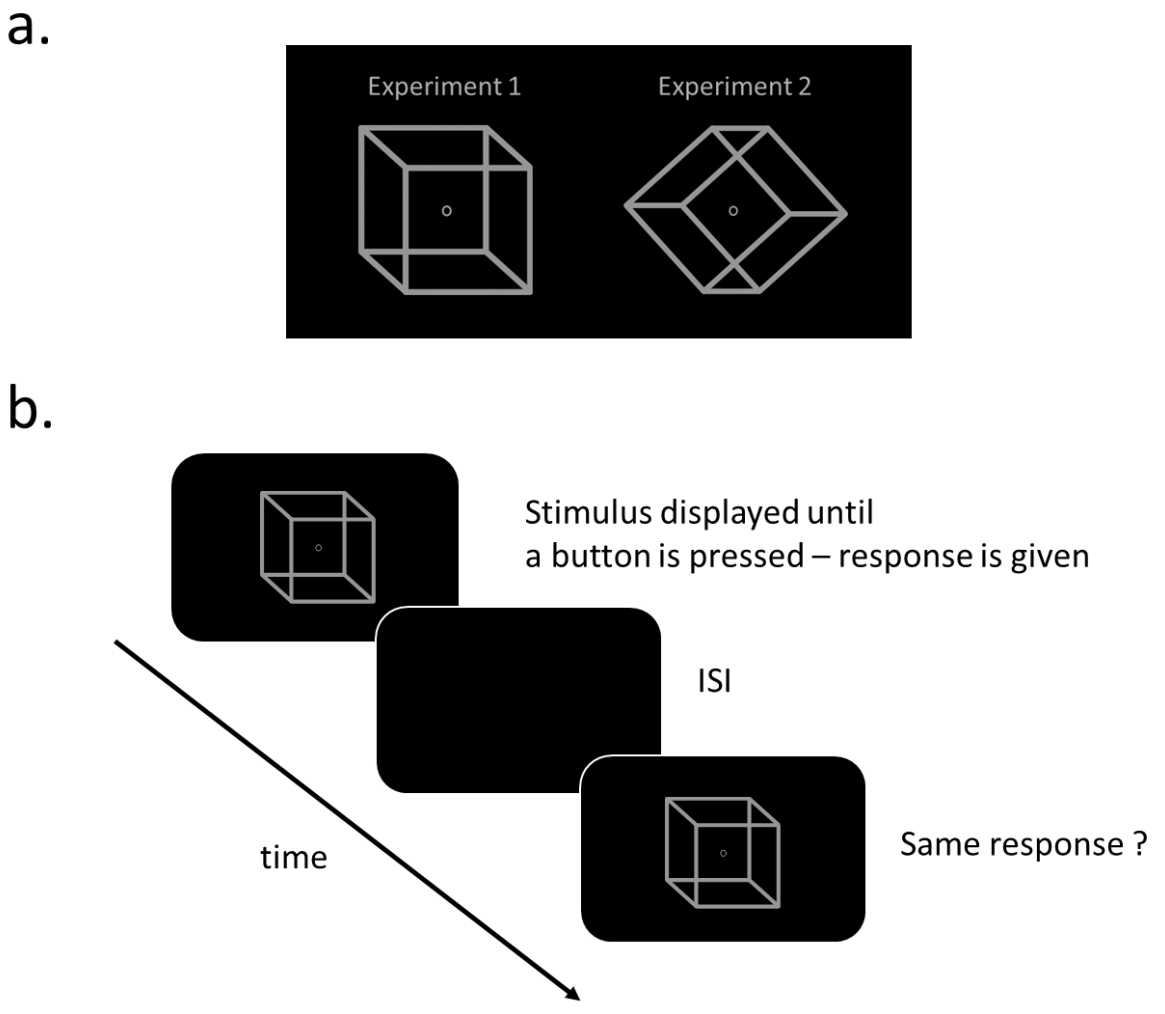


***Figure 1: Stimuli and experimental procedure.*** ***(a.):*** *We used 2 different versions of the Necker cube, a normal cube ("Seen From Above" vs "Seen From Below"; experiment 1) and a tilted cube ("Seen From Right" vs "Seen From Left"; experiment 2). Their main difference is an implicit preference for the SFA interpretation, only present in the normal cube.* ***(b.):*** *Participants were presented with sequences of cubes, separated by*

*blank intervals. Every time the cube was on the screen, participants had to report their dominant percept by pressing a button. Once the button was pressed, the cube disappeared. We measured the survival probabilities for different Inter-Stimulus Intervals (ISI) and used them to reconstruct the stabilisation curves (SP - ISI).*

Before the beginning of the experiment, participants were presented with the stimulus and the two possible interpretations and they had a training session of 5 blocks, 18 trials/block. Those results were not retained in the following analysis.

To avoid the confounding effects of eye-movements when the stimulus is present, participants were instructed to fixate on the central fixation point.

## Model-free analysis

As explained before, reversal rate is not ideal for quantifying stability when the stimulus is intermittently presented. Instead, we chose to use survival probabilities (SP), calculated separately for the two interpretations [21]. A SP can be defined as the probability of perceiving SFA (SFB) after a blank interval, given that SFA (SFB) was also perceived before the blank interval. Importantly, chronic and acute biases [16] have different effects on SP. More specifically, a long-term prior (e.g. SFA bias) generates a difference between the two SP, while a short-term bias makes the sum of the two SP not equal to 1 (see **Figure 2** for an illustration).

Because SP takes values between 0 and 1 and the sample size was small, we used exclusively non-parametric statistics. The effect of the OFF-Duration was tested using a linear mixed-effects model (LMEM) comprising the effects of the blank interval and the different interpretations (SFA vs SFB) as well as their interaction as fixed effects, together with Gaussian random effects for intercepts and slopes (blank interval, interpretation and their interaction). In the tilted cube experiment, the (blank interval x interpretation) interaction term was omitted. Moreover, due to violation of the linearity assumption when considering all OFF-Durations together, in the tilted cube experiment we used two different linear models, one comprising blank intervals between 50 ms and 300 ms (destabilization) and one comprising longer intervals (300 ms $\leq$ OFF $\leq$ 1050 ms; stabilization).

Additionally, because we do not predefine the length of the ON-Duration, our results become vulnerable to longer reaction times (indeed, there is empirical and theoretical evidence suggesting a significant effect of the display duration on SP [3,9]). To control for that, we defined

a threshold (k = 600 ms) and split the data into two groups: a "Low Reaction Time" (LRT) group (RT<k) and a "High Reaction Time" (HRT) group (RT>k). Note that we chose a meaningful threshold, so that both groups contain sufficient amount of data (see **Table 1** for a summary of the RT in the two experiments). Then, we repeated the above analysis (LMEM) in the different subgroups.

Finally, when necessary, we performed post-hoc comparisons using paired rank-sum tests to clarify simple effects in the 2 x 2 design and one-sample Wilcoxon signed rank tests to compare the mean SP with 0.5, i.e., chance level. To test specifically whether two SP ($SP_1$, $SP_0$) are symmetrical or not for a particular condition (if their sum is equal to 1), we used a paired rank-sum test to compare $SP_1$ with (1-$SP_0$) [20]. All significance tests were performed on the two samples of 10 participants each, they were two-tailed and used an alpha value of 0.05 in the statistical toolbox of Matlab v. 2015b (MathWorks, Natich, MA).

## Model and model predictions

Beyond suggesting a novel methodology for testing intermittent presentation in bistable perception, our primary goal was to test the predictions of a *dynamical circular inference* (dCI) model, described elsewhere (see **Chapter 3**), and in particular, to what extent descending loops play an important role in perception.

Here, we consider the more general version of the dCI model, in which the transition rates (the leak) are not stable, but change over time, depending on the current dominant percept, based on equations of leaky integration. As a result, we are able to capture not only the stabilisation trends (SP for longer OFF-Durations; captured also by dCI model with fixed transition rates) but also the initial destabilisation of the percepts (SP for shorter OFF-Durations; not predicted by the simpler model). Note however that, since descending loops have important qualitative effects only on the convergence point of the SP (see **Figure 2** and **chapter 3** of the current thesis), our conclusions would be similar if the transition rates were fixed.

In brief, the dCI model with changing rates can be formalized in the following way:

$$\frac{dL}{dt} = 2w_S a_P L + (r_{on}(t)e^{-L} - r_{off}(t)e^{L}) + (r_{on}(t) - r_{off}(t)) + w_{int}(2w_S - 1)(1 + 2w_S a_S)n_t \quad (1)$$

$$\tau \frac{dr_{on}}{dt} = -r_{on} + r_{on}^{+} + r_{on,B} \quad (2)$$

$$\tau \frac{dr_{off}}{dt} = -r_{off} + r_{off}^{+} + r_{off,B} \quad (3)$$

Eq. 1 describes how the belief about the 3D interpretation of the cube changes in time, under the influence of the descending loops. Eq. 2 and 3 represent the evolution of the rates (note that $r_{on}^{+}$ and $r_{off}^{+}$ are not fixed, but depend on the dominant percept) (for more details, see **Supplementary Material** of **Chapter 3**). We highlight that the noise term in (1) (last term) disappears during the OFF-Duration.

The predictions of the model (with and without descending loops; symmetrical and asymmetrical stimulus) are presented in **Figure 2**. When the stimulus is symmetrical (as in the tilted cube experiment; **Figure 2(a.,b.)**), the two SP overlap, as a result of the lack of a chronic bias. When there are no loops (**Figure 2a**), SP converges to 0.5 (the only fixed point of the system). In the presence of loops however (**Figure 2b**), SP reaches a different value (close to 1 if loops are strong enough), potentially causing stabilisation of both interpretations (note that in this case the sum of the SP is above 1, an indication of a persistent acute bias, generated by the loops).

When the stimulus is asymmetrical (as in the case of the normal cube experiment; **Figure 2(c.,d.)**), SP converge to different values. Without loops, they converge to symmetrical values above and below chance (**Figure 2c**). On the contrary, the descending loops generate a bistable attractor which forces SP to non-symmetrical values (**Figure 2d**; the location of the attractors depends on the strength of the loops and the baseline values of the two rates).

It's interesting to highlight that in all four cases, the model is able to predict non-monotonic stabilisation curves, in which an initial destabilisation is followed by a prolonged increase of the SP (which can lead to almost complete stability). We note however that the properties of the initial deepening depend on the choice of the parameters (mainly the parameters in eq. 2 and 3).

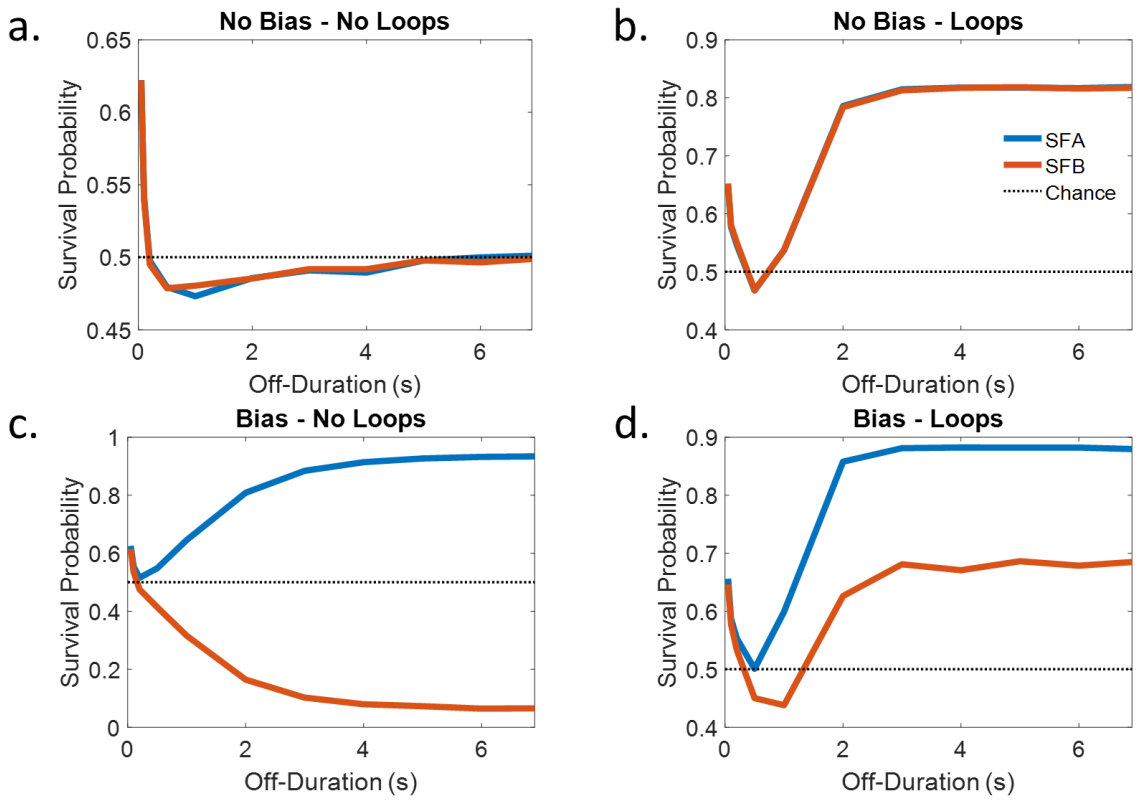


***Figure 2: Model predictions for stabilization curves (SP – OFF-Durations).*** *The model we used was the dynamical circular inference model, presented elsewhere (see* ***Chapter 3****).* ***(a.):*** *No bias (e.g. tilted cube), no (descending) loops. The 2 SP overlap and converge to 0.5.* ***(b.):*** *No bias, loops. SP overlap and converge to a value not equal to 0.5.* ***(c.):*** *Bias (e.g. normal cube), no loops. SP do not overlap and converge to symmetrical values.* ***(d.):*** *Bias, loops. SP do not overlap and converge to non-symmetrical values. [Bias: ron_b = 0.35, roff_b = 0.3; No Bias: ron_b = roff_b = 0.3; No loops: ap = ac = 0; Loops: ap = 1.5, ac = 0].*

## Results

### Experiment 1 (Normal cube)

The stabilisation curves from experiment 1 (normal cube) are presented in **Figure 3a**. Simple visual inspection unveils a lot about the stimulus and the underlying processes. The first thing that we observe is that the two curves (blue: SFA; red: SFB) do not overlap, as we would expect from an asymmetrical stimulus. Additionally, the values of the two SP for long OFF-Durations are not symmetrical (do not sum to 1): the blue curve increases monotonically, reaching a value close to 0.8, while the red curve, after a steep initial decrease, remains stable around chance.

The model free-analysis (LMEM) revealed a significant effect of the OFF-Duration (**$\beta$ = 0.0002, p < 0.001**) while interpretation had no significant effect (**p = 0.72**). Instead, a significant

interaction between the OFF-Duration and the interpretation (**β = -0.0004, p < 0.001**) accounted for the difference between the blue and the red curve. As a control, we repeated the analysis after removing the shortest blank duration (OFF = 100 ms), which added a non-linear effect to the red curve (SFB). The results were not affected by the change.

To test whether the convergence points are symmetrical or not, we focused on the longest OFF-Duration (OFF = 900 ms). The comparison between SP(SFA; OFF = 900 ms) and (1 – SP(SFB; OFF = 900 ms)) revealed a significant difference (**p = 0.009**), meaning that an acute bias is present, as we would expect from a system containing descending loops (**Figure 2d**). In addition to that and for the sake of completeness, we verified that SP(SFA; OFF = 900 ms) was significantly larger than chance (**p = 0.002**), on the contrary SP(SFB; OFF = 900 ms) was not found different from 0.5 (**p = 1**).

Why is there no initial decrease in SP(SFA). Despite the monotonic increase of the blue curve, it's worth noting that the curve starts from very low (SP(SFA; OFF = 100 ms) ~= 0.55). A value so close to chance for such a short OFF-Duration means that there is an initial destabilization for short blank durations, probably characterized by a very short time-scale, thus not captured by our experiment.

Is the result a consequence of the averaging across participants? **Figure S1** illustrates the stabilization curves of the 10 individual participants. Interestingly, although they are noisier, we observe that the average pattern (blue above red; initial destabilization followed by stabilization; non-symmetrical convergence points) is also present for most subjects, suggesting that descending loops might be at work in most of our healthy participants.

Finally, to control for a potential contamination of the results by very long reaction times, we split the results into LRT and HRT and repeated the analysis for each category separately. Those stabilization curves are presented in **Figure S3(a.,b.)**. As expected (longer reaction times imply more accumulation of evidence, hence SP being closer to chance), we found different profiles for the two subgroups: HRT curves (**Figure S3a**) were almost flat and very close to 0.5; more importantly, LRT curves (**Figure S3b**) did not differ qualitatively from the ones presented before. A similar model-free analysis gave similar results as for the entire sample (significant effect of OFF-Duration, non-significant effect of interpretation, significant interaction).

In summary, the results from experiment 1 are not compatible with a system that does exact inference (leaky integration of noisy evidence; **Figure 2c**); conversely they suggest that a

bistable attractor, potentially due to the positive feedback generated by descending loops, is a constitutive mechanism of our perceptual system.

## Experiment 2 (Tilted cube)

The stabilisation curves from experiment 2 (tilted cube) are presented in **Figure 3b**. Compared to the previous results, the most striking difference is that the 2 curves (blue: SFA; red: SFB) now overlap, given the symmetry (no general preference) of the tilted cube. Furthermore, both SP show an initial deepening (we remind that contrary to experiment 1, here the shortest OFF-Duration is 50 ms) while it's noteworthy that for long blank durations, the two SP are not far from chance.

Different LMEM (to avoid violation of linearity) were used for the destabilisation (50 ms:300 ms) and the stabilization (300 ms: 1050 ms) parts of the curves. The first LMEM gave a significant effect of the OFF-Duration (**$\beta$ = -0.0015, p < 0.001**) and a trend for the effect of the interpretation (**$\beta$ = -0.047, p = 0.07**), while the second one revealed again a significant (but this time positive - stabilisation) effect of the OFF-Duration (**$\beta$ = 0.0003, p < 0.001**) and a non-significant effect of the interpretation (**p = 0.56**), as we expected from the tilted version of the Necker cube.

Interestingly, there is no significant difference between SP(SFA; OFF = 1050 ms) and (1 – SP(SFB; OFF = 1050 ms)) (**p = 0.28**), meaning that (at least for this blank interval) there is no acute bias, contrary to what we would expect from a system with descending loops (and contrary to experiment 1). Is this conclusion meaningful? Apart from the fact that a negative finding always need to be interpreted with caution (especially when the sample is small), we see an increasing trend for OFF-Durations between 300 ms and 1050 ms. Would this trend persist for longer intervals or the curves would converge at chance level? To definitely answer this question, we would need a new experiment, testing longer durations. Nevertheless, many previous studies have suggested high levels of stability for longer OFF-Durations, in agreement with the predictions of the dCI model with descending loops (**Figure 2b**; [8,9].

**Figure S2** presents the stabilization curves of the 10 individual participants. As in experiment 1, individuals exhibit the same pattern as the averaged data. It's interesting to note however that we observe consistent biases within participants, suggesting the presence of idiosyncratic preferences, counterbalanced in the entire sample and thus not seen in the averaged data. Additionally, we see that few stabilization curves converge to chance, supporting

our claim that a bistable attractor is still at play, despite the averaged data suggesting the opposite.

Finally, **FigureS3(c.,d.)** presents the results for the HRT and the LRT subgroups. As before, no qualitative difference exists between the averaged data and the LRT data (same for the model-free analysis), while HRT results are closer to chance levels.

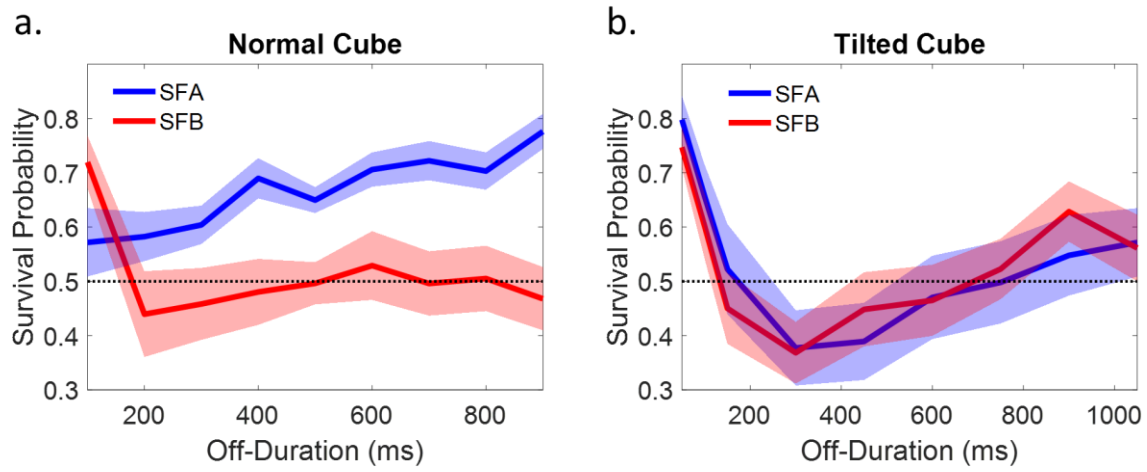


***Figure 3: Experimental stabilization curves. (a.):*** *Normal cube (experiment 1). The SP do not overlap and converge to non-symmetrical values.* ***(b.):*** *Tilted cube (experiment 2). The SP overlap and for OFF=1050 ms, their values are not significantly different from chance. It is not clear whether for longer blank durations the SP would increase or if they would converge to 0.5*

## Discussion

Bistable perception is a unique phenomenon [10]. The visual competition generated by ambiguous stimuli offers a unique opportunity to explore the computational anatomy of perception [1,22,23]. Despite previous efforts, a holistic interpretation of the phenomenon, especially at the functional level, is still missing.

Here, we probed one of the less well-understood aspects of bistable perception, namely the stabilization profile when the stimulus is discontinuously presented [8]. Paradoxically, although the phenomenon is known for more than 50 years [9], it has been largely ignored by functional studies [14,22,24–26]. We investigated intermittent presentation of a bistable stimulus in healthy participants, using a novel methodology that overcomes various shortcomings of previous studies. We ran two experiments with different versions of the Necker cube, each having unique properties. First, we used a normal Necker cube, a common

ambiguous figure known to induce a strong and persistent bias (''Seen From Above'' bias). This was contrasted with results from a tilted cube, in which the implicit preference has been neutralized. Both stimuli were tested in a variety of randomized blank-durations, ranging between a few tens of milliseconds to more than a second, giving us the opportunity to reconstruct the whole stabilisation curve. Importantly, stability was quantified by the two survival probabilities, two measures well suited for discontinuous stimulation [21], which clearly distinguish between chronic and acute biases [16] and ignore confounding factors such as the exposure time to stimulus [8].

Our results reproduced the well-known pattern of intermittent presentation [9]: an initial destabilization (for OFF-Durations below ~= 400 ms) is followed by a long lasting stabilization (in our experiment the longest interval tested was ~= 1 s, but evidence suggests that the same percept persists even after tens of seconds [8]). Interestingly, our results illustrate a striking effect of the priors on this pattern: when there is an asymmetry between the two interpretations (e.g. normal cube), only the strong interpretation is clearly stabilised (**Figure 3a; blue curve**). On the contrary, the SP of the weak interpretation remains close to chance, even for long OFF-Durations (**Figure 3a, red curve**). This result is also crucial from a methodological point of view, since it demonstrates the limitations of measures which do not separate the two interpretations, such as the reversal rate.

We interpreted those results in the context of the recently proposed *dynamical circular inference* framework (see **Chapter 3**). According to dCI, bistability and its rich phenomenology is generated by descending loops in brain circuits. Descending loops are a type of information loops, which cause reverberation and uncontrolled amplification of priors descending the cortical hierarchy [17,27]. Our previous work showed that when combined with dynamics (Hidden Markov Model), descending loops transform the system which is normally a leaky integrator of ambiguous sensory evidence into a bistable attractor, with two stable belief states. Among other things, this affects the behaviour of the system when the stimulus is presented intermittently (**Figure 2**). Crucially, our current experimental results are not compatible with a simple integrator, instead they are qualitatively well captured by a system with (descending) loops.

From this point of view, this study can be added to the accumulating evidence suggesting that loops (especially descending loops) are a fundamental mechanism of normal (non-pathological) brain circuits. A recent study found that the behaviour of healthy participants (and of schizophrenia patients) in a probabilistic reasoning task was best accounted

by a circular inference model with prominent descending loops (although the main difference between patients and controls was found in the amount of climbing loops) [28]. In the same vein, our previous work showed that the combination of priors and sensory evidence in bistable perception follows the rules dictated by circularity (see **Chapter 2** of the present thesis)

Is circular inference the only possible interpretation of the current results? The true answer is certainly not. As argued before, descending loops generate a system that is (mechanistically speaking) equivalent to a bistable attractor. In principle, any mechanism / function with similar dynamics (which continues affecting behaviour in the absence of stimulation) could have produced those results (see for example [12], for a mechanistic interpretation of the effects of intermittent presentation of the stimulus). However, combined with the more restrictive evidence mentioned above ([28]; **Chapter 2** of the thesis), circular inference seems a very plausible candidate.

Although the results presented here strongly point to the presence of an acute bias (bistable attractor), they are not conclusive. This is particularly true for the tilted presentation experiment, in which SP were not significantly different from chance (for OFF = 1050 ms), despite the increasing trend. To further clarify this point, an additional experiment is necessary, testing longer OFF-Durations.

A step further could also be reached by moving from qualitative comparisons to more quantitative approaches, including model fitting and model comparisons. This would allow us to give a more detailed account of participants' behaviour, not only at the level of the group, but also at the level of the individuals, e.g. by detecting subgroups with different characteristics (no loops, more prominent descending loops, climbing loops etc.).

Finally, another interesting extension of this study would be to include other groups known to present prediction impairments, such as schizophrenia patients or even healthy populations experiencing non-clinical hallucinations [29]. The circular inference framework was initially introduced as a model of psychosis (schizophrenia), consequently the present methodology could shed more light to the mechanisms underlying symptoms such as hallucinations and delusions.

In summary, we introduced a new methodology to study intermittent presentation of an ambiguous stimulus, also testing the relevant qualitative predictions of our *dynamical circular inference* model. Our results are compatible with the idea that descending loops shape

the way we perceive the world, playing a role far beyond schizophrenia and the related psychosis.

## Supplementary Material

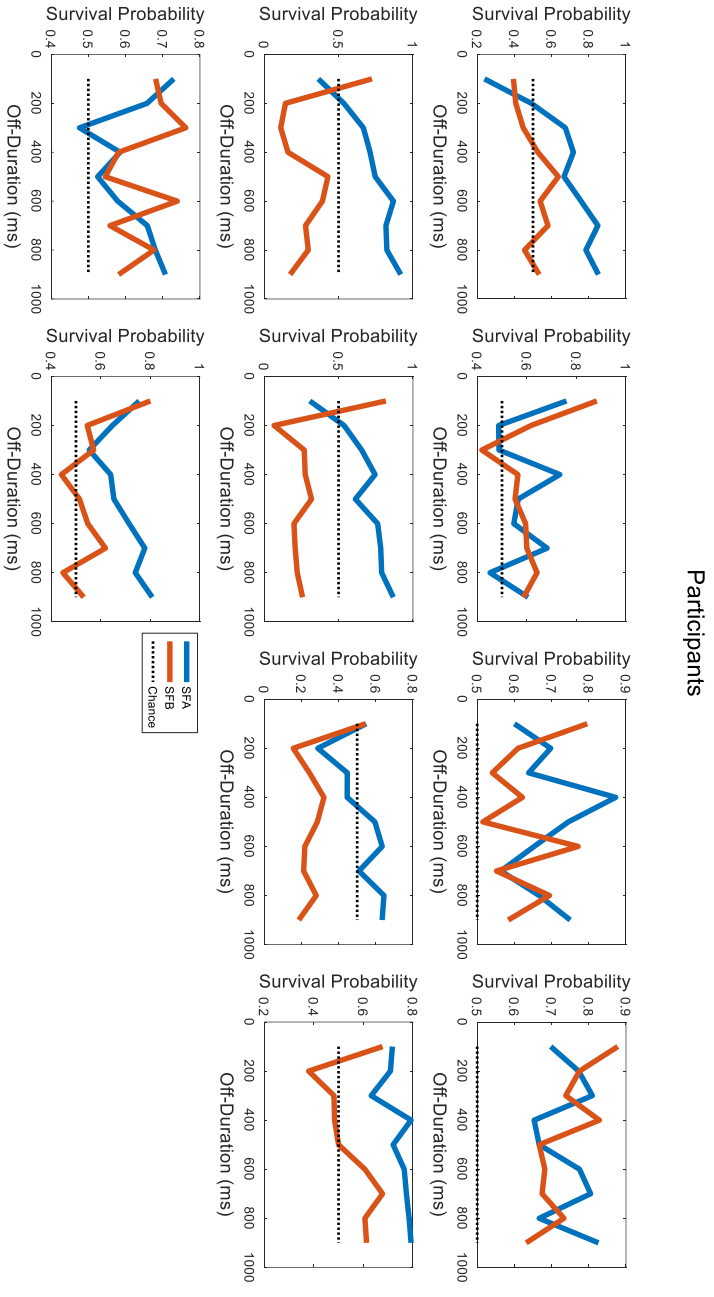


***Figure S1: Results of individuals in experiment 1 (Normal cube).*** *In almost all the cases, the blue curve (SFA) is above the red (SFB) and the SP converge to non-symmetrical values.*

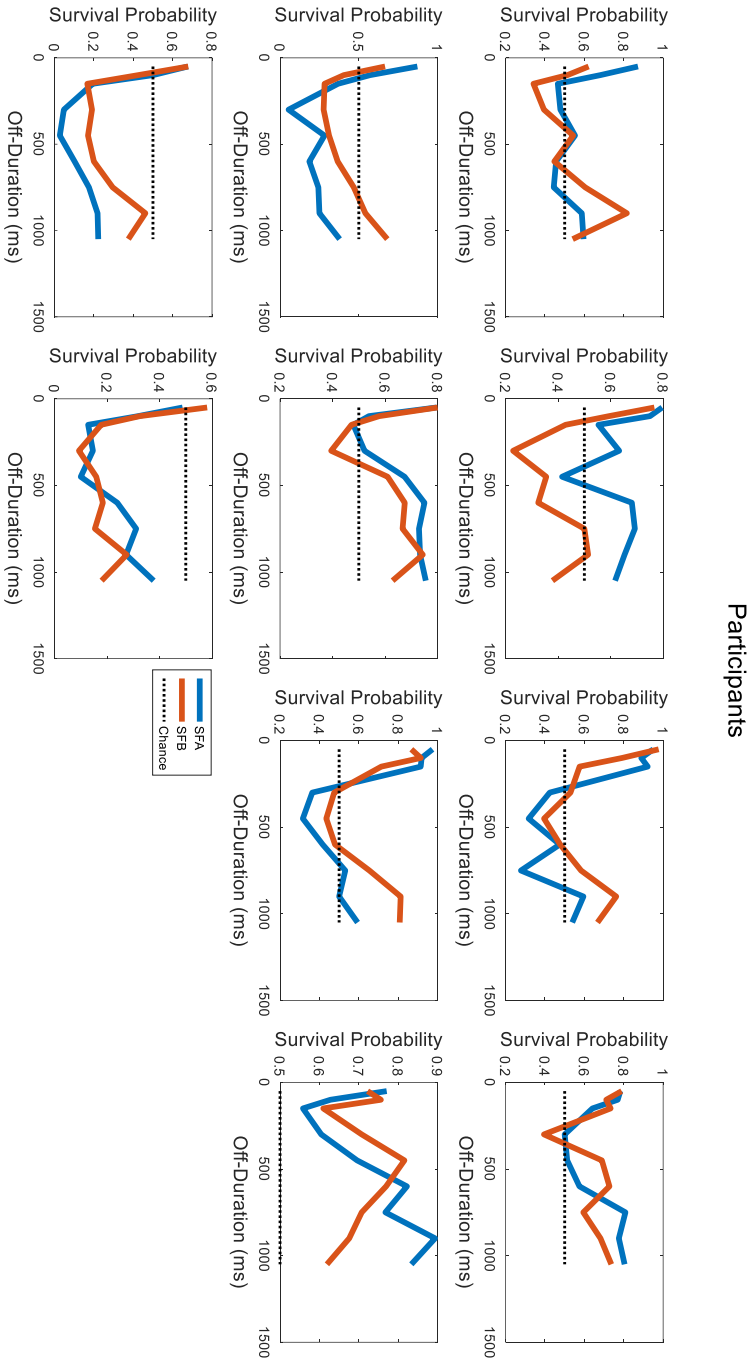


***Figure S2: Results of individuals in experiment 2 (Tilted cube).*** *Idiosyncratic biases are present in most participants.*

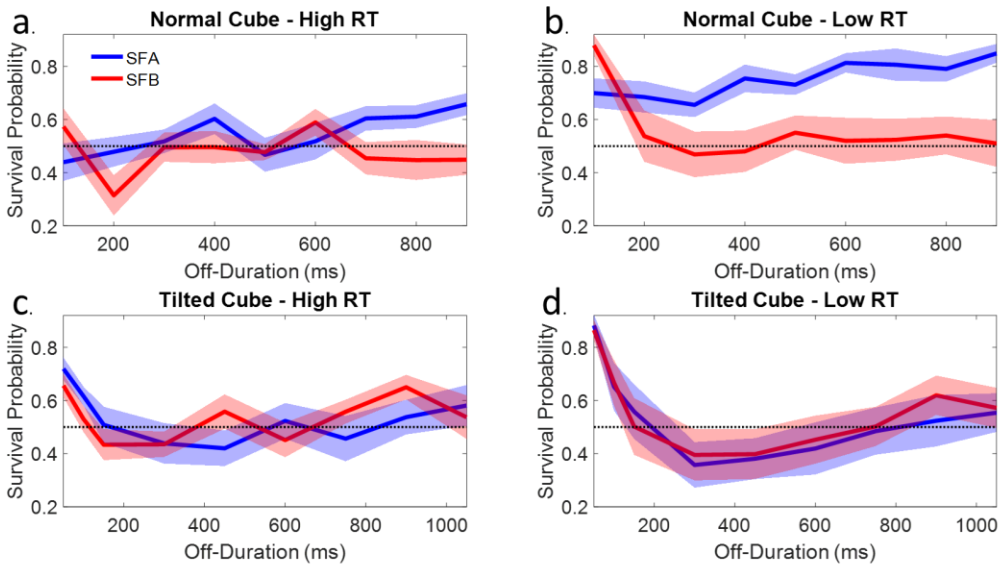


***Figure S3: Experimental results separately for trials with low and high reaction times.*** *High reaction time makes the stabilization curves flatter. For low reaction times, the curves are qualitatively similar to those in the main text.*

## References


1. Blake R, Brascamp J, Heeger DJ. Can binocular rivalry reveal neural correlates of consciousness ? Philos Trans R Soc B. 2014;369. doi:http://dx.doi.org/10.1098/rstb.2013.0211

2. Walker P. Stochastic properties of binocular rivalry alternations. Percept Psychophys. 1975;18: 467–473.

3. Gigante G, Mattia M, Braun J, Del Giudice P. Bistable perception modeled as competing stochastic integrations at two levels. PLoS Comput Biol. 2009;5: 1–9. doi:10.1371/journal.pcbi.1000430

4. Lehky SR. Binocular rivalry is not chaotic. Proc R Soc London B Biol Sci. 1995;259: 71–76.

5. Moreno-Bote R, Rinzel J, Rubin N. Noise-Induced Alternations in an Attractor Network Model of Perceptual Bistability. J Neurophysiol. 2007;98: 1125–1139. doi:10.1152/jn.00116.2007

6. Panagiotaropoulos TI, Kapoor V, Logothetis NK, Deco G. A Common Neurodynamical Mechanism Could Mediate Externally Induced and Intrinsically Generated Transitions in Visual Awareness. PLoS One. 2013;8. doi:10.1371/journal.pone.0053833

7. Pastukhov A, Braun J. Cumulative history quantifies the role of neural adaptation in multistable perception. J Vis. 2011;11: 12–12. doi:10.1167/11.10.12

8. Leopold DA, Wilke M, Maier A, Logothetis NK. Stable perception of visually ambiguous patterns. Nat Neurosci. 2002;5: 605–609. doi:10.1038/nn851

9. Orbach J, Ehrlich D, Heath HA. Reversibility of the Necker Cube: I. An examination of the concept of "satiation of orientation." Percept Mot Skills. 1963;17: 439–458. doi:10.2466/pms.1963.17.2.439

10. Blake R, Logothetis NK. Visual Competition. Nat Rev Neurosci. 2002;3: 1–11. doi:10.1038/nrn701

11. Maier A, Wilke M, Logothetis NK, Leopold DA. Perception of Temporally Interleaved Ambiguous Patterns. Curr Biol. 2003;13: 1076–1085. doi:10.1016/S

12. Noest AJ, Ee R Van, Nijs MM, Wezel RJA Van. Percept-choice sequences driven by interrupted ambiguous stimuli : A low-level neural model. J Vis. 2007;7: 1–14. doi:10.1167/7.8.10.Introduction

13. Brascamp JW, Knapen THJ, Kanai R, Noest AJ, van Ee R, van den Berg A V. Multi-timescale perceptual history resolves visual. PLoS One. 2008;3. doi:10.1371/journal.pone.0001497

14. Sundareswara R, Schrater P. Perceptual multistability predicted by search model for

Bayesian decisions. J Vis. 2008;8: 1–19. doi:10.1167/8.5.12.Introduction

15. Wexler M, Duyck M, Mamassian P. Persistent states in vision break universality and time invariance. Proc Natl Acad Sci. 2015;112: 14990–14995. doi:10.1073/pnas.1508847112

16. Al-dossari M, Blake R, Brascamp JW, Freeman AW. Chronic and acute biases in perceptual stabilization. J Vis. 2015;15: 1–11. doi:10.1167/15.16.4.doi

17. Jardri R, Denève S. Circular inferences in schizophrenia. Brain. 2013;136: 3227–41. doi:10.1093/brain/awt257

18. Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Res. 1998;38: 2817–2832. doi:10.1016/S0042-6989(97)00438-0

19. Kornmeier J, Ehm W, Bigalke H, Bach M. Discontinuous presentation of ambiguous figures: How interstimulus-interval durations affect reversal dynamics and ERPs. Psychophysiology. 2007;44: 552–560. doi:10.1111/j.1469-8986.2007.00525.x

20. Kogo N, Hermans L, Stuer D, van Ee R, Wagemans J. Temporal dynamics of different cases of bi-stable figure-ground perception. Vision Res. 2015;106: 7–19. doi:10.1016/j.visres.2014.10.029

21. Mamassian P, Goutcher R. Temporal dynamics in bistable perception. J Vis. 2005;5: 361–75. doi:10:1167/5.4.7

22. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: An epistemological review. Cognition. 2008;108: 687–701. doi:10.1016/j.cognition.2008.05.010

23. Brascamp J, Sterzer P, Blake R, Knapen T. Multistable Perception and the Role of the Frontoparietal Cortex in Perceptual Inference. Annu Rev Psychol. 2018;69: 1–27.

24. Dayan P. A Hierarchical Model of Binocular Rivalry. Neural Comput. 1998;10: 1119–1135. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-0032111193&partnerID=40&md5=2220a1a71a4cfd3e9066c68547e73897

25. Moreno-Bote R, Knill DC, Pouget A. Bayesian sampling in visual perception. Proc Natl Acad Sci U S A. 2011;108: 12491–12496. doi:10.1073/pnas.1101430108

26. Gershman SJ, Vul E, Tenenbaum JB. Multistability and Perceptual Inference. Neural Comput. 2012;24: 1–24.

27. Leptourgos P, Denève S, Jardri R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Curr Opin Neurobiol. 2017;46: 154–161. doi:10.1016/j.conb.2017.08.012

28. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference in schizophrenia. Nat Commun. 2017;8: 14218. doi:10.1038/ncomms14218

29. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science (80- ). 2017;357: 596–600. doi:10.1126/science.aan3458

# Part II

## Circular inference in the psychosis spectrum

# Chapter 5

## Can circular inference relate the neuropathological and behavioural aspects of schizophrenia?

Published as:

Leptourgos P., Deneve S., Jardri R. Can circular inference relate the neuropathological and behavioural aspects of schizophrenia? Curr. Opin. Neurobiol. 2017; 46: 154–161.

## Abstract

Schizophrenia is a complex and heterogeneous mental disorder, and researchers have only recently begun to understand its neuropathology. However, since the time of Kraepelin and Bleuler, much information has been accumulated regarding the behavioral abnormalities usually encountered in patients suffering from schizophrenia. Despite recent progress, how the latter are caused by the former is still debated. Here, we argue that Circular Inference, a computational framework proposed as a potential explanation for various schizophrenia symptoms, could help end this debate. Based on Marr's three levels of analysis, we discuss how impairments in local and more global neural circuits could generate aberrant beliefs, with far-ranging consequences from probabilistic decision making to high-level visual perception in conditions of ambiguity. Interestingly, the Circular Inference Framework appears to be compatible with a variety of pathophysiological theories of schizophrenia while simulating the behavioral symptoms.

## Introduction

We live in an ambiguous and constantly evolving environment. Being able to make sense and act in such an uncertain world is fundamental for our survival. Consequently, one would expect our brain to be equipped with mechanisms capable of representing and using this uncertainty to draw valid conclusions. Indeed, today there is substantial evidence that various cognitive and motor tasks are probabilistic in nature [1-3], and many of these tasks are performed by humans almost optimally [4,5]. At the same time, scientists have become more and more interested in tasks in which human performance is suboptimal [6,7], which could be due to the use of wrong information or the use of approximations. More recently, this type of impaired inference has been theorized to be at the roots of various neurological or mental disorders, including schizophrenia (Cf. **Box 1**) [8-10].

In this review, we focus on a particular framework for schizophrenia, called *Circular Inference* [10,17,18]. In the first part, we discuss important computational and algorithmic aspects of the framework and its relevance to perception and cognition. In the second part, we propose potential neural and anatomical implementations of the framework and draw connections with other well established neurobiological models of schizophrenia [19,20].

## Box 1: The schizophrenia spectrum

Schizophrenia is a common mental disorder (approximately 1% lifetime prevalence), with a heterogeneous genetic and neurobiological background, that may clinically result in some combination of positive symptoms (i.e., features that are not normally present, such as *hallucinations, delusions or disorganized thinking*), negative symptoms (i.e., characterized by the absence of normal functions, such as *social withdrawal or affective flattening*) and a broad set of cognitive dysfunctions [11]. A unique molecular process/cognitive domain appears unlikely to be involved in schizophrenia, and among the various pathophysiological models proposed to account for this complex phenotype, a widespread change in the neural balance of excitation/inhibition has received multiscale support [12]. The main findings in schizophrenia are: (i) the reduction in the GABA-synthesizing enzyme GAD-67 measured in post-mortem tissue [13]; (ii) abnormalities in Delta/Gamma/Theta band oscillations [14]; (iii) the effectiveness of $D2_R$ antagonists on psychotic symptoms [15], suggesting a dopamine hyperfunction (at least in the mesolimbic pathway); and (iv) the similarity in clinical manifestations after administering $NMDA_R$ antagonists to healthy volunteers [16], suggesting $NMDA_R$ hypofunction.

## The computational level: The Bayesian formalism

When we look at the face of a person, we instantly perceive it as three dimensional since we have depth perception. Although that might seem like a trivial task, which is executed by the brain in a few msec with amazing accuracy, the truth is very different. The 3D shape of a face has to be inferred from the ambiguous 2D retinal projection, using only the inconclusive visual information and prior knowledge accumulated from the past. The optimal integration of such ambiguous information can be formalized using the Bayes theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

where $Y$ and $X$ are random variables (continuous or discrete) representing the 3D interpretation and the 2D retinal image, respectively; $P(Y|X)$ is the posterior probability representing our subjective belief about the 3D interpretation after receiving the new sensory evidence; $P(Y)$ is the prior or our subjective belief before the new evidence; $P(X|Y)$ is the likelihood function that formalizes the dependence of the sensory evidence on the 3D interpretation; and finally $P(X)$ is a normalization term that ensures that the posterior is a probability distribution summing to 1.

From such a perspective, visual perception can be seen as the process of guessing the most probable cause (e.g., 3D object) of the sensory evidence (e.g., 2D retinal image) [21]. For the guess to be optimal (at least for Gaussian variables), likelihood and prior knowledge have to be weighted by their precision, which corresponds to the inverse of the variance of the respective probability distributions. If the information is very precise, then its relative contribution becomes larger.

## The algorithmic level: Belief propagation and circularity

In real life, most of the decision-making problems, perceptual or not, that we have to solve depend on many variables. In many cases, finding the posterior probability of those variables is not an easy task, as it might need calculation of intractable integrals or simply a huge number of summations, which increases exponentially with the number of variables. This problem can be solved by using a generative model, which is a hierarchical representation of the causal structure of the world. A generative model consists of nodes, representing variables, and edges, representing conditional dependencies. Nodes can be arranged in a hierarchical way

such that variables in one layer are potential causes of the variables in the layer below (Cf. **Figure 1a**).

A very general, powerful and efficient algorithm to perform inference in such a generative model is belief propagation (i.e., the sum-product algorithm, [23]). In belief propagation, sensory information $S$ (in **Figure 1a**, this corresponds to the probability of a leaf being present based only on sensory information) climbs the hierarchy in a feedforward way (bottom-up processing) and at the same time, prior information $P$ (e.g., probability of being in a forest, before receiving any sensory information) moves downwards as feedback (top-down processing). Then, each node calculates a belief for the underlying variable (equivalent to the posterior, e.g. $P(X_{tree}|S)$) and sends local messages (e.g., $M_{tree\rightarrow leaf} = P(X_{leaf}|X_{tree})$) to all the neighboring nodes. As a result, information, in the form of beliefs, is propagated throughout the whole system.

If we assume binary variables and use the log-ratios of the probabilities, then beliefs and messages can be calculated by the following recursive equations [10]:

$$M_{ij}^{t+1} = W_{ij}\left(B_i^t - M_{ji}^t\right) \quad (2)$$

$$B_i^{t+1} = \sum_j M_{ji}^{t+1} \quad (3)$$

where:

$M_{ij}^t$ is the message from node i to node j in time t

$B_i^t$ is the belief of node i at time t

$W_{ij}(B)$ is a sigmoid function of $B$

The second equation simply means that each node calculates a belief by summing the messages coming from all its neighbors (e.g., the belief about the presence of a tree is equal to the sum of the messages from the forest and the leaf nodes). The first equation, on the other hand, means that the message travelling from node (i) to node (j) (e.g., from forest to tree) is a function of the belief of the sending node (i) (in our example, the forest), after we subtract the effect that the receiving node (j) has on the sending node (i) (e.g., message from tree to forest).

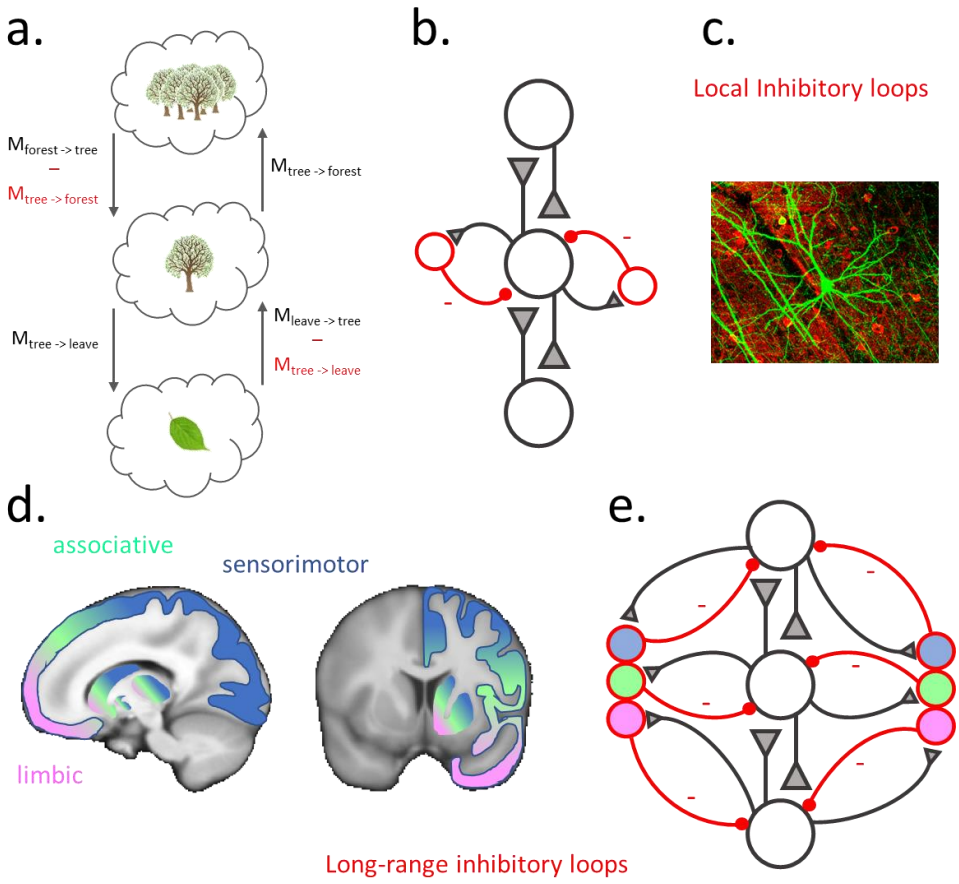


***Figure 1. Circular Belief Propagation: principles and possible neural implementations. a)*** *Belief propagation in a generative model with 3 nodes (the « leaf » is caused by a « tree », which is caused by a « forest »). Messages are locally propagated between nodes (in black), while redundant information is removed (in red). For any 2 neighboring nodes, when we consider message passing in one direction, the redundant information consists of the message sent in the opposite direction;* ***b)*** *A first possible neural implementation of the inhibitory control mechanisms, consists in local inhibitory loops (in red). Different pools of inhibitory interneurons could be responsible for removing redundant feedforward sensory evidence or top-down prior information. The presented network corresponds to the generative model introduced in figure* ***(1a)*** *and the gray connections represent excitatory synapses (between pyramidal cells in different cortical areas or between pyramidal cells in one cortical area and the corresponding inhibitory interneurons);* ***c)*** *This is compatible with microscopic findings showing that pyramidal neurons (in green) are surrounded by different types of GABAergic interneurons (in red, taken from Wei-Chund Allen et al., PLoS Biol 2006);* ***e)*** *A second possible neural implementation of the inhibitory control mechanism consists in long-range inhibitory connections (red) between subcortical structures (e.g., different parts of the striatum, caudate or thalami) and the cortical mantle. Colored nodes correspond to the 3 projection areas presented in* ***(d)****; functional subcortical-cortical territories are color-coded and presented on brain sagittal (left) and coronal views (right). Blue is for sensory-motor circuits, green for associative circuits, and pink for limbic circuits).*

This correction is crucial. Without it, the algorithm would produce loops, i.e., reverberations of bottom-up and/or top-down information. In such “loopy” belief propagation,

the consequences are treated as causes and vice-versa, and the information in the upward and the downward stream can be mixed and overcounted. As a result, beliefs can take extreme values (e.g., absolute certainty) and the system becomes overconfident (**Figure 3**, see also the section on **Behavioral Correlates**). In other cases, beliefs can be reversed (believe that something is present when there is nothing, i.e., having an aberrant perceptual belief or an hallucination) or start oscillating (i.e., a phenomenon called *frustrated network*). Recently, this kind of circular propagation of information in cortical and subcortical networks of the brain has been suggested to underlie the positive and possibly the negative and disorganized symptoms of schizophrenia [18]. In the next section, we will describe the possible neural and anatomical implementations of belief propagation in the brain, and we will discuss how *circular inference* might be associated with well-known physiological and anatomical impairments in schizophrenia.

## The neural level: Implementing inhibitory loops

Currently, the brain is commonly considered a hierarchical system [24,25]. Many algorithms could be used by such a system to make probabilistic inferences [23]. Again, among the many suggestions, belief propagation is certainly one of the most biologically plausible since it is analogous to the integration and propagation of activity in neurons and neural microcircuits. Neural models implementing different versions of belief propagation can be very efficient and robust but also perform a variety of cognitive processes [26-31]. Finally, belief propagation is flexible and can be implemented with both discrete and continuous variables while at the same time representing the whole probability distributions and not just estimates.

Because belief propagation works by propagating top-down messages from high-level representations to low-level sensory features and bottom-up messages in the other direction, precisely controlling message-passing in the hierarchical network is crucial to avoid circularity, which would result in an overcounting of ascending or descending information [10]. At the neural level, this operation can be ensured by inhibition, which is in charge of suppressing the information just sent in the cortical hierarchy from any feedback message. We recently proposed that two types of inhibitory connections could carry out this mission [17], i.e. one type removes redundant feedback messages from the bottom-up stream and symmetrically the other type removes redundant feedforward messages from the top-down stream (Cf. **Figure 1b, d**). This implementation is notably supported by recent findings showing that influences along feedforward projections and those along feedback projections, synchronize in different frequency bands in human visual cortex [32]. Such a microcircuit would optimally balance

excitation and inhibition (i.e., maintain the E/I ratio at approximately 1). If the E/I ratio is impaired in favor of excitation, sensory evidence (bottom-up information) and prior expectations (top-down information) reverberate in the network, causing *Circular Inference*.

Such inhibitory control could rely on *local inhibitory* loops, i.e., inhibitory interneurons controlling the feedforward and feedback flows (Cf. **Figure 1b,c**). A nonspecific global disruption in inhibition is notably supported by the examination of arrays of genes related to GABA neurotransmission (GAD-67 being only one) that showed abnormal expression in association, limbic, motor and sensory cortices [33]. In the same vein, GABA concentration deficits in the occipital cortex of schizophrenia patients have been found to be correlated with impaired behavioral measures of visual inhibition [34]. Despite a widespread inhibitory deficit (i.e., local impairments at each level of the cortical hierarchy), this model predicts an inhomogeneous spatial pattern of aberrant beliefs [17], which are compatible with brain imaging findings in schizophrenia patients [19].

Beyond the local circuit, a second possible implementation relies on *long-range inhibitory loops* (Cf. **Figure 1c, d**). Two prominent long-range extracortical inputs, limbic and thalamic, drive neocortical inhibition [35]. For example, spiral projections between the striatum, a key structure involved in psychosis in general [36] and hallucinations in particular [37], and the neocortex would be a good candidate for inhibitory control of feedback signals (**Figure 1d, left, e**). Interestingly, a similar implementation has been proposed for sending reward prediction errors from the frontal to sensorimotor areas as required for hierarchical reinforcement learning [38]. In the same vein, thalamocortical dysconnectivity is a replicated finding in schizophrenia [20], which may result in long-range inhibitory dysfunctions. Thalamocortical loops (**Figure 1d, right,e)** would constitute a good candidate for the inhibitory control of feedforward signals since it forms multiple parallel pathways to relay information to and from the cortex [39]. The Inflow and outflow of causal information between the visual cortex and the thalamus has been recently shown to be affected in schizophrenia [40]. Interestingly, using a computational model of the neural dynamics in resting-state fMRI, Yang et al. have found that increasing the effective strength of connectivity at either the local or long-range level resulted in an elevated E/I ratio that was able to capture both the local and global neural variability observed in schizophrenia [41]. Again, such impairments have been found to be maximum in association networks [42], which appear compatible with fMRI capture studies showing increased signal in modality-dependent association cortices during hallucinations [43].

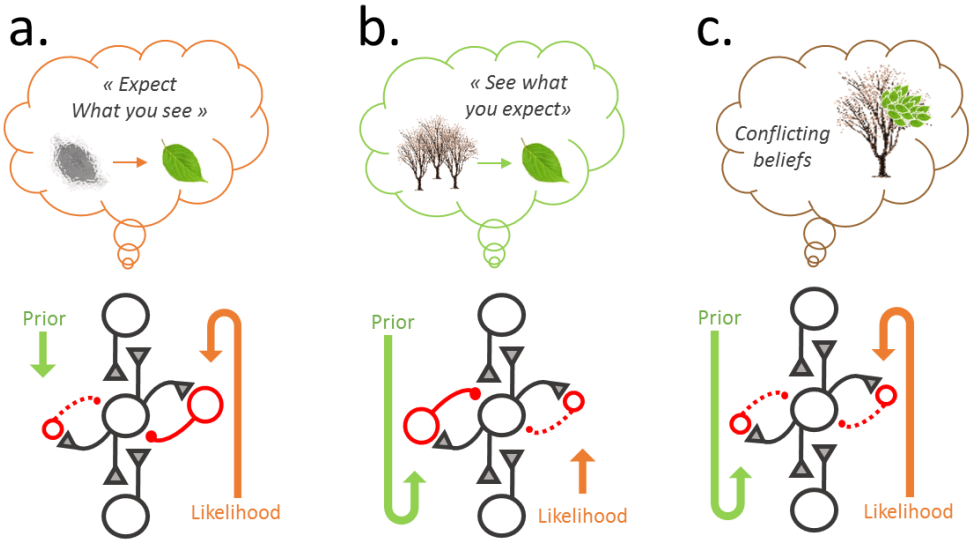


***Figure 2. Behavioral consequences of Circular Inference. a)*** *In a system with climbing loops, sensory evidence is reverberated and corrupts the prior knowledge (orange arrow). As a result, such a system expects what it perceives (for example, even very weak evidence for leaves will result in expectations for leaves, which in turn reinforces the initial input);* ***b)*** *In the opposite case of a system with descending loops, the prior is reverberated and thus corrupts sensory evidence (green arrow). Consequently, this system perceives what it expects (for example, expecting that you are in a forest will automatically result in seeing leaves, even during winter);* ***c)*** *If both climbing loops and descending loops are present, in cases where sensory evidence contradicts the prior, the system (also called « frustrated network ») might start believing two mutually exclusive facts simultaneously; for example, leaves are present and absent at the same time, on the same tree.*

## Behavioral correlates of circular inference

*Circular inference* may result from either ~~a~~ form of inhibitory loop implementations (i.e., local or long-range), producing serious impairments in perception and decisions under uncertainty. For instance, *circular inference* can result in strange coincidence detection, a percept in the absence of corresponding sensory input (i.e., hallucinations, see **Figure 2a,b**), making decisions on the basis of limited evidence (jumping to conclusions), or the learning of unshakable aberrant beliefs (i.e., delusions), as observed during the prodromal and acute psychotic phase of schizophrenia. *Circular inference* can also result in strong, mutually incompatible representations at different levels of the hierarchy (i.e., dissociative thoughts, Cf. **Figure 2c**).

In order to validate the *circular inference* framework at the behavioural level, we used an adapted version of the beads-task paradigm, the “fisher task”, where a fish, either red or black, is fished from one of 2 lakes, each with different percentages of red and black fishes. We

systematically manipulated the prior probabilities ($P(X_{lake})$, representing the preference of the fisherman) and likelihoods ($P(X_{fish}|X_{lake})$, i.e. the real percentages of red and black fishes in each lake) and asked participants to report their confidence (($P(X_{lake}|X_{fish})$, i.e. the posterior probability). We showed that the behavior of schizophrenia patients and controls was best explained by a parametric circular inference model [18]. The signature of circularity included a *sigmoidal shape* (rather than a linear curve) for confidence (log-posterior ratio) as a function of log-likelihood and log-prior ratio, *slopes larger than 1* for weakly informative likelihood and priors (instead of slope equal to one) and *nonlinear interactions* between likelihood and priors (e.g., larger slopes for non-informative prior or likelihood), as illustrated in **Figure 3**. Using this framework, we have recently shown that positive symptoms correlated with the strength of the ascending loops (i.e., an impaired inhibition of redundant bottom-up inputs), negative symptoms correlated with the strength of descending loops (i.e., an impaired inhibition of redundant top-down inputs), and disorganized symptoms correlated with both impairments. Beyond the support of an association between psychotic symptoms and sensory overcounting, this finding is compatible with the paradoxical finding that patients with schizophrenia, notably those with psychotic features, are less vulnerable to sensory illusions [22,44]. *Circular Inference* thus appears to be able to model high-level but also low-level impairments in the schizophrenia spectrum. Interestingly, we found that circular inference was also present in control subjects, especially in the case of descending loops. A tendency to "perceive what we expect" (misinterpret top-down predictions as if they were external sensory signals) may be widely present in the general population. This may account in turn for the strength of certain perceptual illusions as well as for bistable perception, a phenomenon that only requires a small amount of descending loops (Cf. **Box 2**).

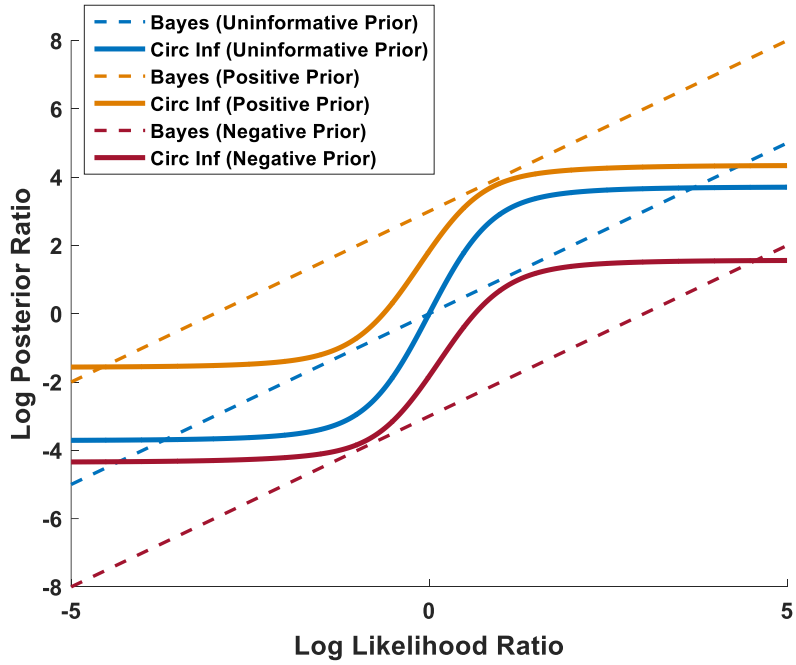


***Figure 3. Bayes and Circular Inference predictions.*** *The Log-Posterior-Ratio is plotted as a function of the Log-Likelihood-Ratio (for different values of the Prior). Dashed lines represent simple Bayes predictions, while plain lines correspond to Circular Inference model predictions. The orange curve is for the cases where priors agree with sensory evidence. The red curve is for the case where priors disagree with sensory evidence. The blue curve is for uninformative prior. Two important features of Circular Inference are illustrated. First, there is an interaction between sensory evidence and priors, i.e., the slope of the blue sigmoid curve, corresponding to ambiguous prior, is larger than the slope of the other 2 sigmoids. This relationship is not captured by simple Bayes. Second, in all 3 cases presented, the slopes of the sigmoids are larger than the slopes of the corresponding dashed lines (equal to 1), suggesting overconfidence caused by Circular Inference. This is compatible with the « jumping-to-conclusion » phenomenon, which was previously shown to be associated with delusional beliefs.*

## Box 2. Circular inference and bistability

Beyond confidence in behavioral decisions, Circular Inference may also account for bistable perception due to the presence of descending loops (e.g., an incomplete cancellation of redundant top-down messages). The problem of bistable perception can be formalized using a Hidden Markov Model: The current percept (e.g., one of the two configurations of a "Necker Cube") depends on the previous percept ("Markov") and is updated by an indirect, noisy sensory observation ("Hidden"). Without circularity, there is only one stable belief corresponding to the prior probability of each configuration (in that case, the figure would always be perceived as ambiguous, Cf. **Figure 4a, b, c, f**). Noisy and ambiguous sensory inputs can make the posterior hover around the prior but can never give rise to a strong unambiguous percept (e.g., a period

when the probability is consistently high for one interpretation, i.e., orange and green gradients in **Figure 4f, g**). Moreover, it cannot capture many other aspects of bistable perception (such as gamma distribution of the percept durations). However, in the presence of a limited degree of descending loops, the system becomes bistable (Cf. **Figure 4d, e, g**). This system can thus generate strong and persistent percepts. Such a parametric model applied to bistable perception data could be a powerful tool to disentangle the contribution of ascending and descending inference loops. Bistable perception is also known to be affected in schizophrenia [22].

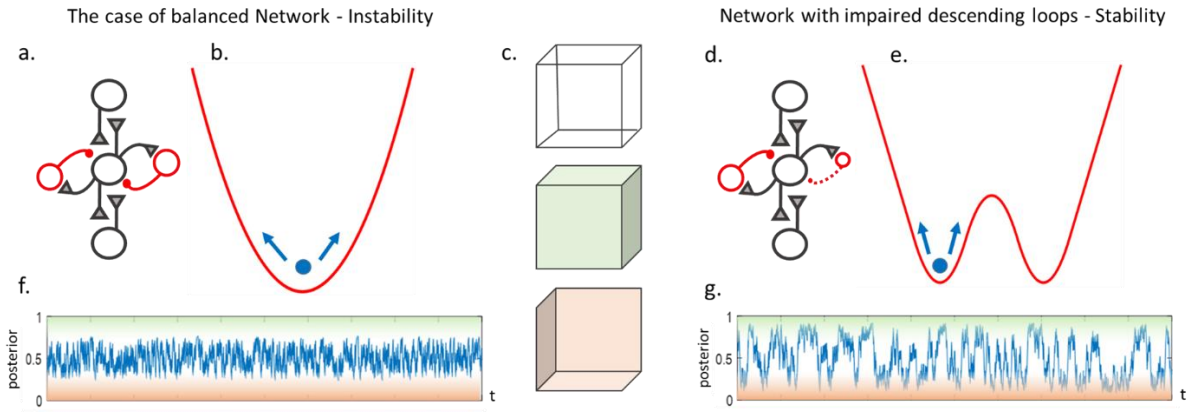


***Figure 4. Circular Inference and the dynamics of bistable perception.*** *The Necker Cube is a bistable percept with 2 possible interpretations* ***(c)****, i.e., a 3D cube as seen from above (in green) or a 3D cube as seen from below (in orange). The following two belief-propagation networks are considered: a balanced network without loops* ***(a)*** *and a network with a small amount of impaired descending loops (descending circular inference)* ***(d)****. Perception of the Necker Cube in a balanced network looks like a « random walk » around chance level* ***(f)****, which appears unable to generate strong and persistent representations* ***(b)****. In contrast, overcounting of priors in a network with descending loops* ***(d)*** *leads to bistability by increasing the strength and persistence of percepts* ***(g)****, i.e., through a bistable attractor* ***(e)****. These simulations appear in agreement with recent findings showing that healthy participants may exhibit small amounts of descending loops [18].*

## Conclusion and perspectives

This brief review has highlighted areas of recent progress in computational approaches to schizophrenia. In such a context, the understanding of how neural networks interact to produce mental events in an uncertain world is an important goal. We chose to focus on a particular hierarchical Bayesian framework, i.e. the *Circular Inference model.* However, it should be highlighted that *Computational Psychiatry* [45,46] is a rapidly growing field, and various other ideas have also been suggested to account for the schizophrenia spectrum, including impaired *predictive coding* [9, see also [18] for a comprehensive comparison with Circular

Inference], or defective outcome prediction, due to problematic error-likelihood predictions in the prefrontal cortex [47].

We noted that a tightly calibrated balance between excitation and inhibition is crucial to ensure an efficient functioning of the nervous system. Among the proposed pathophysiological models of schizophrenia, an impaired E/I ratio is able to capture several features of the disorder (positive symptoms, cognitive deficits, etc.). Various hypotheses about the neural impairment involved in schizophrenia have been proposed and are still debated, e.g., are we facing a deficit that could be widespread across the cortical hierarchy? Or would the disorder result more from focal dysfunctions in a few "hotspots", such as the thalamic or limbic loops?

Significantly, the *Circular Inference* model, is compatible with these apparently competing hypotheses and can begin to associate different scales of understanding. Even if *Circular Inference* first received direct experimental behavioral support in schizophrenia[18], there is an urgent need to close current knowledge gaps. In particular, this framework needs to be validated at the neurophysiological level (e.g., using a multiscale approach combining computational modeling with electroencephalography, to detect the dynamical footprints of the reverberated messages, or with Magnetic Resonance Spectroscopy to confirm an E/I imbalance in the thalami or the striatum of patients suffering from schizophrenia), and its neural implementation needs to be further characterized.

## References with marks


1. Kording KP, Wolpert DM: Bayesian integration in sensorimotor learning. *Nature* 2004, 427:244-247.

2. Kersten D, Yuille A: Bayesian models of object perception. *Curr Opin Neurobiol* 2003, 13:150-158.

3. Kersten D, Mamassian P, Yuille A: Object perception as Bayesian inference. *Annu Rev Psychol* 2004, 55:271-304.

4. Ernst MO, Banks MS: Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 2002, 415:429-433.

5. Weiss Y, Simoncelli EP, Adelson EH: Motion illusions as optimal percepts. *Nat Neurosci* 2002, 5:598-604.

6. Acerbi L, Vijayakumar S, Wolpert DM: On the origins of suboptimality in human probabilistic inference. *PLoS Comput Biol* 2014, 10:e1003661.

7. Drugowitsch J, Wyart V, Devauchelle AD, Koechlin E: Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* 2016, 92:1398-1411.

8. Fletcher PC, Frith CD: Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci* 2009, 10:48-58.
   An in-depth review of findings supporting Bayesian approaches of psychotic symptoms in schizophrenia.

9. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ: The computational anatomy of psychosis. *Front Psychiatry* 2013, 4:47.

10. Jardri R, Deneve S: Circular inferences in schizophrenia. *Brain* 2013, 136:3227-3241.

11. Kahn RS, Sommer IE, Murray RM, Meyer-Lindenberg A, Weinberger DR, Cannon TD, O'Donovan M, Correll CU, Kane JM, Van Os J, et al.: Schizophrenia. *Nat Rev Dis Prim* 2015, 1:1-23.

12. Lisman J: Excitation, inhibition, local oscillations, or large-scale loops: what causes the symptoms of schizophrenia? *Curr Opin Neurobiol* 2012, 22:537-544.
   A comprehensive synthesis on the possible micro- and macro-circuits affected by excitation/inhibition imbalance in schizophrenia.

13. Curley AA, Arion D, Volk DW, Asafu-Adjei JK, Sampson AR, Fish KN, Lewis DA: Cortical deficits of glutamic acid decarboxylase 67 expression in schizophrenia: clinical, protein, and cell type-specific features. *Am J Psychiatry* 2011, 168:921-929.

14. Uhlhaas PJ, Singer W: Abnormal neural oscillations and synchrony in schizophrenia. *Nat Rev Neurosci* 2010, 11:100-113.

15. Zhang JP, Lencz T, Malhotra AK: D2 receptor genetic variation and clinical response to antipsychotic drug treatment: a meta-analysis. *Am J Psychiatry* 2010, 167:763-772.

16. Corlett PR, Honey GD, Krystal JH, Fletcher PC: Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology* 2011, 36:294-315.

17. Denève S, Jardri R: Circular inference: mistaken belief, misplaced trust. *Curr Opin Behav Sci* 2016, 11:40-48.

18. Jardri R, Duverne S, Litvinova AS, Deneve S: Experimental evidence for circular inference in schizophrenia. *Nat Commun* 2017, 8:14218.
First demonstration that Circular Inference outperforms simple and weighted Bayesian models in fitting participant-per-participant's behavior during a probabilistic reasoning task. Patients with schizophrenia exhibited a larger amount of ascending loops compared with matched healthy controls.

19. Jardri R, Hugdahl K, Hugues M, Brunelin J, Waters F, Alderson-Day B, Smailes D, Sterzer P, Corlett PR, Leptourgos P, et al.: Are hallucinations due to an imbalance between excitatory and inhibitory influences on the brain? *Schizophr Bull* 2016, 42:1124-1134.

20. Murray JD, Anticevic A: Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. *Schizophr Res* 2017, 180:70-77.

21. Knill DC, Richards W: *Perception as bayesian inference*. Cambridge, MA: Cambridge University Press; 1996.

22. Notredame CE, Pins D, Deneve S, Jardri R: What visual illusions teach us about schizophrenia. *Front Integr Neurosci* 2014, 8:63.

23. Bishop CM: *Pattern Recognition and Machine Learning*. New York: Springer; 2007.

24. Friston K: Hierarchical models in the brain. *PLoS Comput Biol* 2008, 4:e1000211.

25. Markov NT, Kennedy H: The importance of being hierarchical. *Curr Opin Neurobiol* 2013, 23:187-194.

26. Lochmann T, Deneve S: Neural processing as causal inference. *Curr Opin Neurobiol* 2011, 21:774-781.

27. Litvak S, Ullman S: Cortical circuitry implementing graphical models. *Neural Comput* 2009, 21:3010-3056.

28. George D, Hawkins J: Towards a mathematical theory of cortical micro-circuits. *PLoS Comput Biol* 2009, 5:e1000532.

29. Steimer A, Maass W, Douglas R: Belief propagation in networks of spiking neurons. *Neural Comput* 2009, 21:2502-2523.

30. Ott T, Stoop R: Benefits and pitfalls of belief-propagation-mediated superparamagnetic clustering. *Phys Rev E Stat Nonlin Soft Matter Phys* 2006, 74:042103.

31. Rao RPN: Neural Models of Bayesian Belief Propagation. In *The Bayesian Brain: Probabilistic Approaches to Neural Coding*. Edited by Doya K, Ishii S, Pouget A, Rao RPN: MIT Press; 2007:235-260.

32. Michalareas G, Vezoli J, van Pelt S, Schoffelen JM, Kennedy H, Fries P: Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* 2016, 89:384-397.
Excellent demonstration of how influences along bottom-up and top-down projections in visual cortex predominate in selective frequency bands. These findings support the idea of different inhibitory loops for feedforward and feedback flows as proposed in the Circular Inference framework.

33. Hashimoto T, Bazmi HH, Mirnics K, Wu Q, Sampson AR, Lewis DA: Conserved regional patterns of GABA-related transcript expression in the neocortex of subjects with schizophrenia. *Am J Psychiatry* 2008, 165:479-489.

34. Yoon JH, Maddock RJ, Rokem A, Silver MA, Minzenberg MJ, Ragland JD, Carter CS: GABA Concentration Is Reduced in Visual Cortex in Schizophrenia and Correlates with Orientation-Specific Surround Suppression. *Journal of Neuroscience* 2010, 30:3777-3781.

35. Maffei A: Fifty shades of inhibition. *Curr Opin Neurobiol* 2016, 43:43-47.

36. Howes O, Bose S, Turkheimer F, Valli I, Egerton A, Stahl D, Valmaggia L, Allen P, Murray R, McGuire P: Progressive increase in striatal dopamine synthesis capacity as patients develop psychosis: a PET study. *Mol Psychiatry* 2011, 16:885-886.

37. Rolland B, Amad A, Poulet E, Bordet R, Vignaud A, Bation R, Delmaire C, Thomas P, Cottencin O, Jardri R: Resting-state functional connectivity of the nucleus accumbens in auditory and visual hallucinations in schizophrenia. *Schizophrenia Bulletin* 2015, 41:291-299.

38. Haruno M, Kawato M: Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI in stimulus-action-reward association learning. *Neural Netw* 2006, 19:1242-1254.
A thoughtful reinforcement learning model in the motor domain based on striatal-cortical loops, supporting the possible implementation of long-range control loops in neural networks. While not directly related to schizophrenia, this paper suggests that spiral cortical-subcortical connections might be involved in the backpropagation of reward prediction errors. A similar mechanism could be involved in the control of descending loops in the circular inference framework.

39. Behrens TE, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CA, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, et al.: Non-invasive mapping of

connections between human thalamus and cortex using diffusion imaging. *Nat Neurosci* 2003, 6:750-757.

40. Iwabuchi SJ, Palaniyappan L: Abnormalities in the effective connectivity of visuothalamic circuitry in schizophrenia. *Psychol Med* 2017:1-11.

41. Yang GJ, Murray JD, Repovs G, Cole MW, Savic A, Glasser MF, Pittenger C, Krystal JH, Wang XJ, Pearlson GD, et al.: Altered global brain signal in schizophrenia. *Proc Natl Acad Sci U S A* 2014, 111:7438-7443.

42. Yang GJ, Murray JD, Wang XJ, Glahn DC, Pearlson GD, Repovs G, Krystal JH, Anticevic A: Functional hierarchy underlies preferential connectivity disturbances in schizophrenia. *Proc Natl Acad Sci U S A* 2015:in press.

43. Jardri R, Thomas P, Delmaire C, Delion P, Pins D: Neurodynamical organization of modality-dependent hallucinations. *Cereb Cortex* 2013, 23:1108-1117.

44. Dakin S, Carlin P, Hemsley D: Weak suppression of visual context in chronic schizophrenia. *Curr Biol* 2005, 15:R822-824.

45. Montague PR, Dolan RJ, Friston KJ, Dayan P: Computational psychiatry. *Trends Cogn Sci* 2012, 16:72-80.

46. Huys QJ, Maia TV, Frank MJ: Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 2016, 19:404-413.

47. Krawitz A, Braver TS, Barch DM, Brown JW: Impaired error-likelihood prediction in medial prefrontal cortex in schizophrenia. *Neuroimage* 2011, 54:1506-1517.

# Chapter 6

## Bistable perception in schizophrenia: a functional approach based on circular inference

In preparation for publication as:

Leptourgos P., Eck M., Tiberghien M., Deneve S., Jardri R., (in prep.). Bistable perception in schizophrenia: A functional approach based on circular inference

## Abstract

Schizophrenia is a devastating psychiatric disorder characterized by heterogeneous symptoms, including aberrant percepts (hallucinations) and fixed, idiosyncratic beliefs (delusions), grouped under the term "positive symptoms". Previous work has shown that the positive dimension of schizophrenia is linked to impairments in inference mechanisms, in particular the aberrant amplification of sensory information due to circular inference. Additionally, common perceptual phenomena such as bistable perception, a persistent oscillation between two mutually exclusive interpretations under conditions of high ambiguity, have also been associated with moderate forms of circularity in non-clinical populations. In this study, we sought to better understand the computational (inferential) mechanisms associated with schizophrenia, using bistability as a tool. Two groups of schizophrenia patients, with prominent positive symptoms, were compared with matched groups of healthy controls in two tasks including bistable visual stimuli (i.e., the Necker cube, NC). First, participants were continuously exposed to different versions of the NC (ambiguous or with additional visual cues) and reported their dominant percept every time a sound signal was given. In the second experiment, participants were discontinuously exposed to ambiguous cubes and responded as quickly as possible every time the cube appeared on the screen. In the continuous-presentation experiment, we found that patients were less affected by visual cues and less stable than controls, while we also observed a (group x cue) interaction. Interestingly, we also found that patients exhibit a tendency to ignore evidence that contradicted their implicit preference to see cubes from above, which, together with the reduced effect of supporting cues, correlated with non-clinical psychotic traits and the severity of positive symptoms. In the discontinuous-presentation experiment, we found that patients exhibited an enhanced destabilization (for short blank intervals) and a reduced stabilization of the weak ("Seen-From-Below") interpretation (for longer intervals). In-silico simulations based on *dynamical circular inference* can account for these specific patterns in a model combining enhanced climbing loops with an overestimation of the environmental volatility. Altogether, our results provide additional evidence for the involvement of circularity in the generation of psychotic experiences, pointing towards an enhancement of bottom-up, rather than top-down processing.

## Introduction

Schizophrenia is a devastating and heterogeneous disorder with a high lifetime prevalence (approximately 0.5% - 1% [1]). Patients diagnosed with schizophrenia exhibit a variety of symptoms, which are usually subdivided into three major (and largely independent) clusters [2,3]: the positive dimension, including psychotic symptoms such as hallucinations (aberrant percepts in the absence of sensory input) and delusions (bizarre and fixed beliefs); the negative dimension, defined as the absence of normal behaviors and comprising symptoms such as lack of volition, reduced speech, social withdrawal and emotional blunting; finally a set of cognitive impairments including deficits in working memory and attention but also irrational speech and disorganization.

Although the exact causes of schizophrenia remain a subject of debate, there is a growing consensus that impairments in predictive processing of the brain might underlie the positive (and potentially the other) dimension(s) of the disorder [4–7]. In the same vein, we suggested a few years ago that hallucinations and delusions might be the result of uncontrolled propagation of probabilistic messages between cortical areas representing a hierarchical internal model of the external world [8]. At the neurophysiological level, this *circular inference* (CI) was proposed to be a consequence of an imbalance between neural excitation and inhibition [9,10], a property that has been extensively linked with schizophrenia and psychosis [11,12]. More recently, the pertinence of the framework was demonstrated using a variant of the beads task (named the ''Fisher task''). In this experiment, the behavior of schizophrenia patients was best captured by a CI model (compared to purely Bayesian models), while the most striking difference between patients and healthy controls was in the amount of climbing loops in a hierarchical neural model (i.e., amplification of sensory evidence) [13]. In addition to that, climbing loops correlated with the severity of the positive symptoms whereas descending loops (i.e., amplification of priors) correlated with the severity of the negative symptoms, suggesting that different dimensions might be related to different model parameters and thus, underlying mechanisms.

Strikingly, CI has also been associated with normal brain functioning. In the same study, Jardri and colleagues found that a CI model best accounted not only for patients' responses, but also for the behavior of healthy participants [13]. Furthermore, in the first part of the thesis (**Chapters 2,3 and 4**), we presented a detailed computational account of bistable perception (switching between mutually exclusive interpretations, while the sensory organ is stimulated by an unchanged, ambiguous stimulus), based on the notion of circularity (**Chapter 3**), which was

also validated by experimental evidence (**Chapters 2 and 4**). Interestingly, both studies highlighted the importance of descending loops in explaining cognitive processing (probabilistic reasoning in the former, high-level perceptual processing in the latter) in healthy populations.

In the light of the aforementioned results, we would like to advance the following statement: mild circularity (and mostly descending loops) is a general property of the brain, affecting the way we interpret and interact with the world. An increased amount of loops, following genetic or neurodevelopmental abnormalities, could lead to important deviations from Bayesian optimality, resulting in serious cognitive, perceptual or even motor impairments and the manifestation of pathological symptoms. For example, too strong climbing loops could result in a system that ''expects what it sees'', causing sensory-driven hallucinations and reduced vulnerability to visual illusions, both observed in schizophrenia.

The current work combines previous results and methodologies in order to test the above hypothesis and bring new insights regarding the mechanisms underlying the various clinical dimensions of schizophrenia. Having established links between *circular inference* and bistable perception on one hand and *circular inference* and schizophrenia on the other, we used bistable perception as a tool to probe the computational (inferential) abnormalities related to schizophrenia. We compared the performance of schizophrenia patients with that of matched healthy controls in two bistable perception tasks. First, we used a task in which the stimulus (Necker cube) was continuously displayed and responses were collected in a discontinuous fashion [14]. Additionally, the strength of the sensory evidence was manipulated by adding visual cues (see also **Chapter 2**). Second, we used a task in which the cubes were discontinuously presented, using the methodology introduced in **Chapter 4** [15,16]**.** Importantly, we referred to a *dynamical circular inference* model (see **Chapter 3**) to make qualitative predictions about patients' behavior, for different scenarios (increased climbing loops, increased descending loops etc.).

The preliminary results presented in this chapter show significant differences between patients with schizophrenia and healthy controls, supporting the idea that the machinery underlying bistable perception is impaired in schizophrenia. Although a system with stronger climbing loops could explain most of the observed patterns, it fails to capture certain trends, suggesting the presence of secondary (or more complex) deficits. We critically discuss what those deficits could be, along with more general alternative interpretations of the data.

## Methods

We conducted two bistable perception experiments (continuous presentation of the Necker cube: later called CNC; and discontinuous presentation of the Necker cube: latter called DNC). For each of them, two groups of participants were enrolled: a group of schizophrenia patients with prominent positive symptoms and a group of healthy controls We would like to highlight that the findings presented in this chapter are preliminary (i.e., from a relatively small sample), and that recruitment is still on-going (expected final samples are of 30 patients and 30 controls per experiment).

### Participants

12 schizophrenia patients and 17 healthy control subjects participated in the first experiment (CNC), while 9 patients and 14 healthy participants took part in the second one (DNC). Among them, 8 patients and 4 controls participated in both experiments (first in CNC, then in DNC). The groups were matched in age and sex. **Table 1** provides additional information about the four samples. All the participants were recruited in the Lille city area and were tested in the same experimental conditions on the CURE research platform. Patients all met the ICD-10 criteria for schizophrenia [17]. The main inclusion criteria, for both groups, were the following: age > 18 years, provision of informed consent, normal or corrected-to-normal visual acuity, no past or current medical history of neurological, sensory or psychiatric disorders (for patients, that included the absence of an Axis-II or secondary Axis-I diagnosis), and no current or recent use of psychotropic medication or toxic drugs. A senior psychiatrist confirmed the absence of psychiatric symptoms in the control groups using the Mini International Neuropsychiatric Interview (MINI-ICD-10) [18]. The study was approved by an ethics committee (CPP Sud-Ouest Outre-Mer I).

| | Experiment 1 (CNC) | | | Experiment 2 (DNC) | | | CNC vs DNC |
|---|---|---|---|---|---|---|---|
| | CTR | SCZ | P | CTR | SCZ | p | p (CTR / SCZ) |
| ***Demographics*** | | | | | | | |
| Sample Size (n) | 17 | 12 | - | 14 | 9 | - | - |
| Age (y. o.) | 35.8 (13.6) | 39.5 (8.4) | 0.1 | 38.6 (13.3) | 41.1 (7.8) | 0.33 | 0.40/ 0.59 |
| Gender (m:f) | 11:6 | 10:2 | 0.41 | 11:3 | 8:1 | 1 | 0.46 / 1 |
| Education | 16.1 (2.2) | 13.9 (3) | 0.03* | 18 (3) | 13.7 (3.6) | 0.008 ** | 0.09 / 0.86 |
| ***Neuropsychological Evaluation*** | | | | | | | |
| Spatial attention (Bells task) | 1.4 (1.8) | 2.6 (1.7) | 0.05* | 1.4 (1.9) | 2 (1.7) | 0.27 | 0.88 / 0.45 |
| Cognitive inhibition (Stroop task) | 0.6 (1) | 2.4 (5.4) | 0.98 | 0.4 (1) | 1.4 (3.1) | 0.57 | 0.53 / 1 |
| Working memory (backward digit span) | 5.3 (1.9) | 5.2 (1.7) | 0.93 | 5.9 (2.3) | 5.2 (1.7) | 0.5 | 0.50 / 0.97 |
| Non-clinical beliefs (PDI-21) | 10.8 (9.5) | 50.1 (38.1) | *** | 7.9 (5.5) | 42 (23.5) | *** | 0.51 / 0.85 |
| Non-clinical hallucinations (LSHS) | 9.9 (7.7) | 30.3 (11.9) | *** | 10.9 (7.5) | 27.3 (9.1) | *** | 0.68 / 0.75 |
| ***Clinical dimensions (PANSS)*** | | | | | | | |
| PANSS Total | - | 41.1 (11.8) | - | - | 40 (12.1) | - | 0.91 |
| PANSS Positive | - | 11 (4.9) | - | - | 10.1 (5.5) | - | 0.85 |
| PANSS Negative | - | 10.8 (4.9) | - | - | 11.8 (5.3) | - | 0.70 |
| PANSS Disorganized | - | 6.6 (3) | - | - | 6 (3.5) | - | 0.41 |
| PANSS Excited | - | 5.2 (1.6) | - | - | 4.8 (1) | - | 0.80 |
| PANSS Depressed | - | 7.5 (3.6) | - | - | 7.4 (2.6) | - | 0.88 |
| ***Medication Status*** | | | | | | | |
| Antipsychotics eq. dose (OLZ in mg) | - | 24.5 (24.3) | - | - | 28.1 (29.3) | - | 0.82 |
| Minor tranquilizers eq. dose (DZP in mg) | - | 12.5 (21.6) | - | - | 6.4 (8) | - | 0.77 |

***Table 1: Characteristics of the recruited samples.***

## Symptoms' (patients) and psychotic traits' (patients, controls) assessment

Symptoms' severity was assessed in schizophrenia patients using the *Positive and Negative Syndrome Scale* (PANSS) [19], with items clustered in reference to a 5-factor model, i.e., positive, negative, disorganized, excited and depressed factors [20]. In addition to that, we measured non-clinical psychotic traits in all the samples using two different scales: the ''Peters et al. Delusions Inventory – 21'' (PDI-21), measuring non-clinical beliefs [21] and the revised ''Launay Slade Hallucinations Scale'' (LSHS-R) measuring non-clinical hallucinations [22]. Scores obtained for these different scales are presented in **Table 1**.

## Neuropsychological tests

All the participants also passed 3 neuropsychological tests, to make sure that their cognitive abilities were not severely impaired: a Stroop interference task (for cognitive

inhibition), the Bells cancellation task (for spatial attention) and the digit span task (for working memory) (see **Table 1**).

## Patients' medications

All the recruited patients exhibited a partial remission. They notably received antipsychotic drugs (typical and/or atypical antipsychotics), and in some cases minor tranquilizers. In order to account for potential confounding effects, we referred to Olanzapine-equivalency (OLZ-eq) (antipsychotics; [23]) and Diazepam-equivalency (DZP-eq) (minor tranquilizers; [24]) (see **Table 1**).

## Apparatus

Both experiments took place in the same dark room. The stimuli were displayed on a 17-inch LED computer screen with a resolution of 1280x1024 pixels, at 60 Hz. Responses were collected using a keyboard. The background colour of the screen was black. The viewing distance was 60 cm and a head-rest secured the position of the head. The experiments were implemented in MATLAB v. 2015b (MathWorks, Natich, MA), using Psychtoolbox v. 3.0.12.

## Stimuli

The stimuli were 200x200-pixels, light-gray (or contrasted; see **Experimental paradigm**) Necker cubes. In both experiments we used a normal cube (**Figure 1a**), expected to generate an implicit preference for the "*Seen From Above*" (SFA) interpretation, as compared to the "*Seen From Below*" (SFB) interpretation ([25]; see also **Chapters 2,3 and 4** of the current thesis). Stimuli were presented in the middle of the screen. A circular fixation point was added in the middle of each cube, to guide participants' gaze.

## Experimental paradigms

The first experiment (CNC, see **Figure 1b**) was inspired by the Mamassian and Goutcher's protocol [14] and consisted of 6 blocks of 5 consecutive runs. Using a forced-choice

method, we asked participants to report their ongoing interpretation as soon as they heard a warning sound, which occurred 25 times per run in a pseudo-regular manner (mean inter-sound interval = 1.5 s, uniformly distributed between 1 and 2 s). Each response corresponded to a trial, providing a discontinuous sampling of the task's perceptual dynamics. Runs were separated from each other by a 10 s black screen to minimize between-run influences. The experiment was also interspersed with 5 between-block breaks of non-predefined duration.

We manipulated sensory evidence either by making the cubes homogeneously gray (i.e., perfectly ambiguous) or cuing them using shadows (**Figure 1a**). This additional depth information was intended to bias perception toward one interpretation or the other. It was specified by two parameters:

- First the orientation, which was defined in relation to the implicit prior. A shadow falling on the top left corner was expected to emphasize the SFA preference (classified as a supporting cue). Conversely, a shadow that fell on the bottom right corner was characterized as a contradictory cue, as it went against implicit bias.
- Second, the strength of the cue (which can also be conceived in terms of the amount of sensory information), controlled by the shadowing contrast level. Weak and strong cues corresponded to 20% and 30% contrast, respectively.

The 1st and 4th blocks always consisted of presentation of an ambiguous cube. The other blocks were randomly allocated a different type of cue, defined by the 2 x 2 factorial combination of 2 possible orientations (contradicting or supporting) and 2 possible strengths (weak or strong).

The second experiment (DNC) was an ''intermittent presentation'' experiment, similar to the ones described in **Chapter 4** (**Figure 1c**) [15,16]**.** It consisted of 12 blocks, each constituting a sequence of ON- (stimulus is present) and OFF-Durations (stimulus is absent). A single trial contained an OFF-Duration, as well as the ON-Durations before and after. There were 64 ON-Durations (and 63 OFF-Durations) per block, making up a set of 63 trials per block. Blocks were separated by 10 s black-screen breaks, while 2 additional breaks of non-predefined duration were also possible after the fourth and the eighth block.

Likewise CNC, DNC was also a forced choice task. Participants were instructed to report their dominant percept, as quickly as possible, every time the stimulus reappeared on the screen. Responses were given by pressing the relevant button (Right: Seen From Above/Right; Left: Seen From Below/Left). ON-Durations did not have a standard length, instead the stimulus

disappeared right after the first button-press. Consequently, participants could give only one response per ON-Duration (i.e., their first impression), solving the problem of switches during stimulus' presentation, but also avoiding confounding post-decision effects (e.g. accumulation of evidence after the button-press) and making the task more interactive (increasing the sense of involvement and the levels of attention).

As in **Chapter 4**, we were interested in reconstructing the whole stabilisation curve. Consequently, we used a set of 9 OFF-Durations: {50, 100, 150, 300, 450, 600, 750, 900, 1050 [ms]}, the same as in the tilted cube experiment presented before. This set, different from the one used in the normal cube experiment in healthy controls (see **Chapter 4**) allowed for a higher resolution for short intervals (where an inversion from de-stabilisation to stabilisation usually occurs) but also a wider range of intervals. Crucially, blank intervals were randomized within each block, with 7 repetitions per block per interval. This gave a total number of 84 repetitions per OFF-Duration per experiment.

Before both experiments, participants were presented with the stimulus and the two possible interpretations and they had a training session (CNC: 3 runs, 25 trials/run, only ambiguous cube; DNC: 5 blocks, 18 trials/block). Those results were not retained in the following analysis. The basic instruction was to passively view the cubes without trying to constrain perception.

To avoid the confounding effects of eye-movements when the stimulus is present, participants were instructed to fixate on the central fixation point.

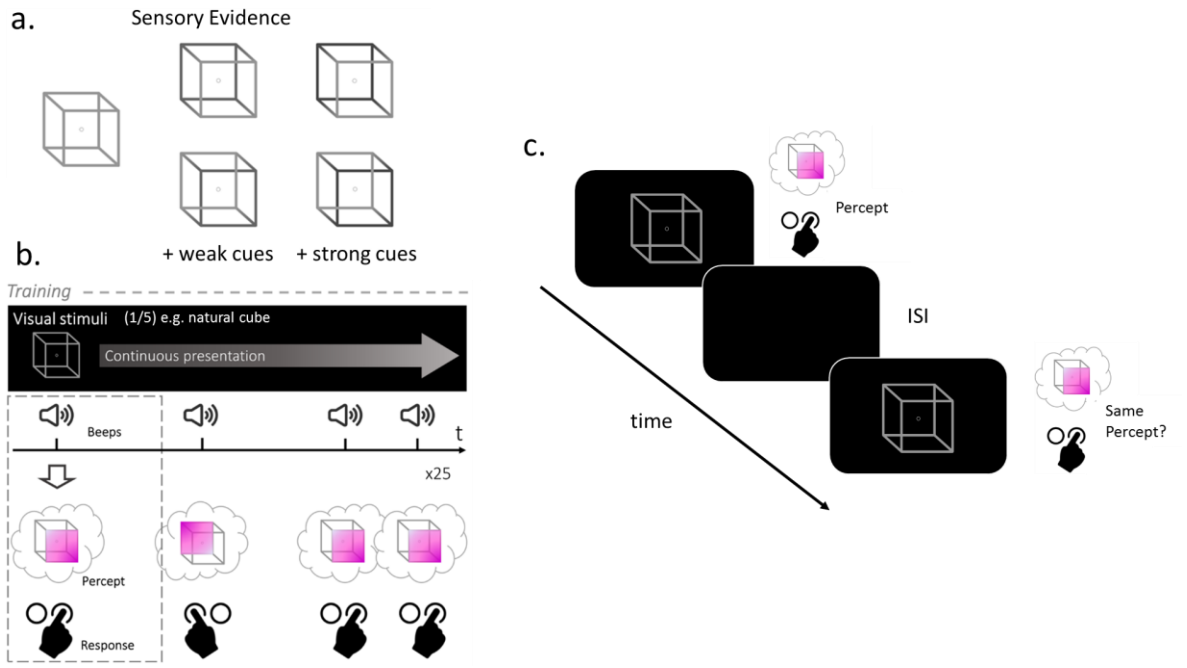


***Figure 1: Stimuli and experimental paradigms.*** ***(a.):*** *The stimuli were Necker cubes, completely ambiguous (CNC and DNC) or disambiguated with the addition of contrast (strong or weak; supporting the implicit preference or not; only in CNC).* ***(b.):*** *The CNC paradigm is inspired by (Mamassian and Goutcher, 2005) and is described in depth in* ***Chapter 2****. In brief, a cube was continuously displayed and participants reported the perceived interpretation (by pressing a button) every time they heard a warning sound.* ***(c.):*** *The DNC protocol was introduced in* ***Chapter 4****. The cubes were presented intermittently, separated by a black screen that lasted from 50 to 1050 ms (9 randomized blank intervals). Every time the cube reappeared on the screen, participants reported their first impression by pressing the relevant button. The cube disappeared right after the participants' response. We measured the survival probabilities for both interpretations and for all the blank intervals.*

## Model-free analysis

In experiment 1 (CNC experiment), the main variable was the relative predominance (RP), which corresponds to the probability of perceiving the SFA interpretation. Additionally, we measured stability by calculating the survival probabilities (SP) (separately for each interpretation), which is the probability that a percept persists between two sounds, conditioned upon the previous percept [14]. As in our previous work (**Chapter 4**), the main variable in experiment 2 (DNC) was the survival probability (SP) computed for each interpretation.

Because both RP and SP take values between 0 and 1 and the sample size was small, we used exclusively non-parametric statistics. To account for the different effects while avoiding multiple comparisons problems, we used linear mixed-effects models (LMEM). In particular, in

CNC, we used different models for the effects on RP and SP: For the former (RP), we used a LMEM comprising the effects of the group (patients vs healthy controls) and the effect of the contrast as well as their interaction as fixed effects, together with Gaussian random effects for intercepts and slopes. For the latter (SP), we used the same model adding the effect of the response (SFA vs SFB) and all the 2-way ((cue x group), (cue x response), (group x response)) and 3-way (cue x group x response) interactions as fixed effects, while as random effects we considered, apart from the random intercepts, the effects of the response, the contrast and their interaction per participant (random slopes). In the case of patients, in order to further investigate the possibility that supporting and contradicting evidence do not have the same effect on RP and SP (as well as to test potential violation of Levelt's 2nd law [26]), we repeated this analysis, using Friedman's test and focusing on this group while considering separately the conditions in which the visual cue supports / contradicts the implicit SFA-preference (for SP, the analysis was performed separately for the two SP as well).

In the DNC experiment, to avoid the problems posed by the violation of linearity, we used distinct linear models for the destabilisation period (OFF-Durations between 50 and 150 ms) and the stabilisation period (OFF-Durations between 300 and 1050 ms, see also **Chapter 4**). They both comprised the following terms: group, contrast, response, all the 2-way interactions and a 3 way interaction as fixed effects and random intercept and slopes for contrast, response and their interaction. Additionally, because we do not predefine the length of the ON-Duration, our results were vulnerable to longer reaction times (apart from the known effect of the presentation time on intermittent bistable perception [15,27], we expected patients to be slower in giving their responses [28]). To control for that, we re-analyzed the data after splitting them into two subgroups, based on a threshold (k = 700 ms): a "Low Reaction Time" (LRT) subgroup (RT<k) and a "High Reaction Time" (HRT) subgroup (RT>k). Note that we chose a meaningful threshold, so that both groups (both for patients and controls) contain sufficient amount of data.

Furthermore, when necessary, we performed post-hoc comparisons for both experiments using paired (or unpaired) rank-sum tests to clarify simple effects in the 2 x 2 design and one-sample Wilcoxon signed rank tests to compare the results in different conditions with the chance level (0.5). In the DNC experiment, to test specifically whether two SP ($SP_1$, $SP_0$) are symmetrical or not for a particular condition (if their sum is equal to 1), we used a paired rank-sum test to compare $SP_1$ with ($1-SP_0$) [29]. All significance tests were performed on the entire samples, they were two-tailed and used an alpha value of 0.05 in the statistical toolbox of Matlab v. 2015b (MathWorks, Natich, MA).

Finally, to account for the effects of medication and symptoms (in both experiments), we added the equivalent doses of OLZ and DZP as well as the PDI, LSHS and PANSS scores to the linear models as covariates. Note that since medication and PANSS scores are defined only for patients, the corresponding linear models only included this group. In addition to that, to further probe the links between medication/ symptoms and bistable perception, we looked directly for associations using the Spearman's rank correlation coefficient r. Again, we would like to highlight that due to the small samples and the large number of potential correlations, we only explored associations with a particular meaning or for which we had preliminary evidence (e.g. positive effects in the LMEM).

## Model and model predictions

The results of the two experiments were interpreted in the context of a *dynamical circular inference* (dCI) model, which is described in detail in **Chapter 3**.

As we did in **Chapter 4** and in order to account also for the initial destabilization when the stimulus is intermittently presented (DNC experiment), we considered the more general version of the dCI model, in which the transition rates (the leak) are not stable, but change over time, depending on the current dominant percept, based on equations of leaky integration (see **Supplementary Material** in **Chapter 3**).

In brief, the dCI model with changing rates can be formalized in the following way:

$$\frac{dL}{dt} = 2w_S a_P L + (r_{on}(t)e^{-L} - r_{off}(t)e^{L}) + (r_{on}(t) - r_{off}(t)) + w_{int}(2w_S - 1)(1 + 2w_S a_S)n_t \quad (1)$$

$$\tau \frac{dr_{on}}{dt} = -r_{on} + r_{on}^{+} + r_{on,B} \quad (2)$$

$$\tau \frac{dr_{off}}{dt} = -r_{off} + r_{off}^{+} + r_{off,B} \quad (3)$$

The predictions of the model (for an asymmetrical stimulus) for different parameters (descending loops, climbing loops and baseline transition rates, vaguely corresponding to an estimate of the environmental volatility) are presented in **Figure 2**. In the **Panels g,h** and **i**, the difference between the two rates is kept constant. When the stimulus is continuously presented, descending loops and climbing loops have opposite results on RP and SP. More specifically, descending loops increase both bias (due to the implicit preference (ambiguous condition) and to visual cues; **Figure 2a**) and persistence (**Figure 2b**), while climbing loops decrease them (**Figures 2(d.,e.)**). Note that increasing the baseline rates has the same effect as increasing the

climbing loops (in a CNC experiment) (**Figure 2(g.,h.)**). Crucially, all those alterations predict a change in the slope of the curves, which corresponds to an expected interaction between the cue and the group. This prediction differentiates those interpretations from others, such as a change in one of the two rates, which would only cause a shift of the psychometric curve to the left or to the right (**Figure S1**). It's important to highlight that the predictions described here (dCI model with changing rates) are qualitatively the same as the ones presented in **Chapter 3** for the simpler version of the model (see also **Figure 8** in **Chapter 3**).

Descending loops and climbing loops also have opposite results on the stabilisation curves, when the ambiguous stimulus is discontinuously displayed. **Panels c, f and i** in **Figure 2** demonstrate the effect of the blank period on the SP, for different values of descending loops (**Figure 2c**), climbing loops (**Figure 2f**) and baseline rates (**Figure 2i**). It's convenient to consider the two parts of the curve separately (initial destabilization followed by a more persistent stabilization). For the stabilization part of the curve, the more complex model with the changing rates makes the same predictions as the dCI model with a constant prior (rates) (**Chapter 3**). In particular, stronger descending loops (which push the two stable fixed points of the system further away from 0.5 – chance level) result in more stabilization (convergence of both SP to larger values). On the contrary, both stronger climbing loops and higher rates cause a decrease in the stabilisation (at least for as long as there is a bistable attractor), but for different reasons: the higher rates bring the attractors closer to 0.5, while the stronger climbing loops increase the effect of the noisy evidence (the sensory gain) when the stimulus is present (and until a decision is made).

Interestingly, in the destabilisation phase there is a dissociation between climbing loops and rates: while stronger loops (climbing and descending) cause less destabilisation (SP decreases less), higher rates have the opposite effect, enhancing destabilisation of both percepts. Overall, we see a triple dissociation between climbing loops, descending loops and transition rates, which could guide the interpretation of our results in the context of circular inferences. We note however that, in the absence of analytical results, our predictions regarding the destabilisation phase in intermittent bistable perception are entirely based on simulations. Although those predictions are true for a wide range of parameters, a formal proof of the generality of the result still needs to be presented.

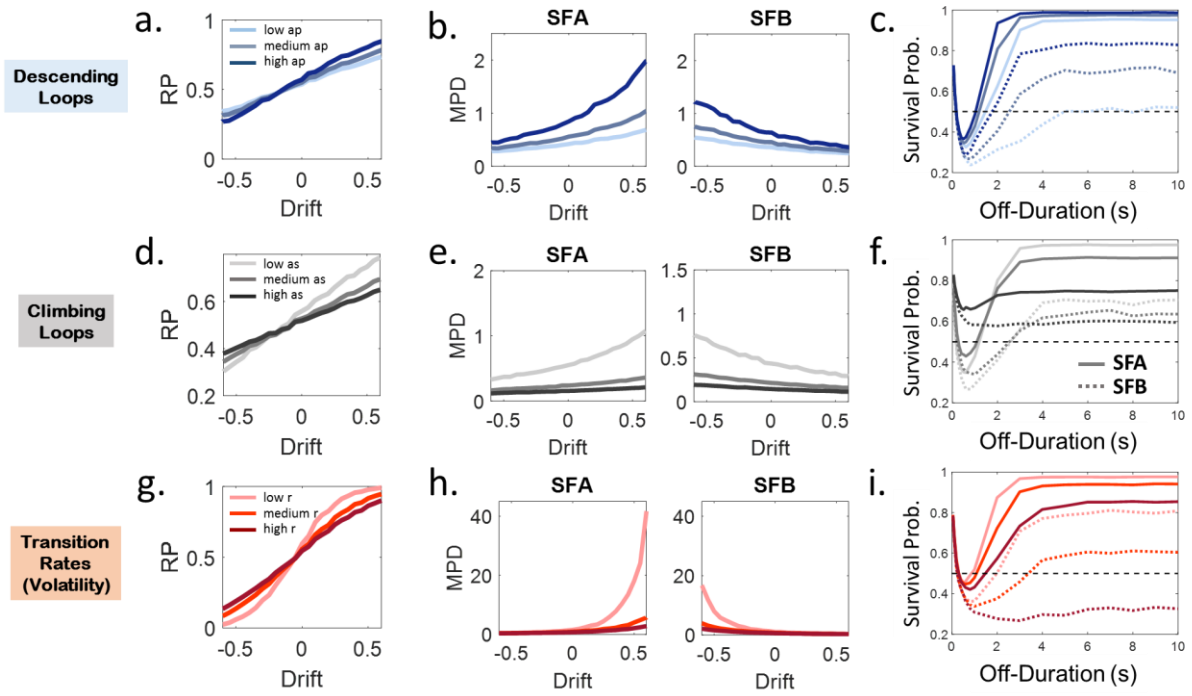


***Figure 2: Model Predictions for CNC and DNC experiments.*** ***Panels a, b, d, e, g*** *and* ***h*** *correspond to predictions for the CNC experiment while* ***panels c, f*** *and* ***i*** *illustrate the predictions for the DNC experiment. Three possible alterations are depicted. Lighter colours illustrate smaller values of the parameters.* ***(a.,b.,c.):*** *Stronger descending loops result in enhanced effects of cues on RP (a.; drift ≠ 0), more biased responses in the ambiguous condition (****a.****; drift = 0), more stability* ***(b.)*** *while they also generate more stabilisation and less destabilisation when the stimulus is intermittently presented* ***(c.)****.* ***(d.,e.,f.):*** *The opposite effects are caused by a strengthening of the climbing loops. In that case, we observe a weaker effect of cues on RP, less biased responses in the ambiguous condition, less stability, less stabilisation and less destabilisation (the last prediction is the same between climbing and descending loops).* ***(g.,h.,i.):*** *Higher baseline transition rates produce qualitatively the same behaviour as the stronger climbing loops, apart from an enhanced destabilisation for short blank durations in intermittent presentation tasks. Note that all 3 manipulations produce a non-linear multiplicative effect on RP, instead of a simple shift of the psychometric curve.*

## Results

Patients and healthy controls were matched between the two experiments in terms of demographic characteristics, medication, and symptoms' severity (**Table 1**).

### Experiment 1 (CNC)

The results of the CNC experiment are presented in **Figure 3**. In terms of bias (Realtive Predominance (RP)), there is an important difference in the slopes of the two curves (**Figure**

**3a**), suggesting a difference in the effect of the visual cues between the two groups. Additionally, we observe higher stability for controls (Survival Probabilities (SP); **Figure 3b**), at least for the interpretation that is supported by the visual cue.

*Relative Predominance*

The linear mixed models for the RP revealed a significant effect of the visual cues (**$\beta$ = 0.45, p < 0.001**). There was no significant effect of the group (**p = 0.7**), instead there was a strong (cue x group) interaction (**$\beta$ = -0.38, p < 0.001**), resulting in a smaller slope and weaker effects of the visual cues in patients (**Figure 3a**; red curve). Both groups exhibited an implicit preference, as implied by the significant difference between the RP in the ambiguous condition and chance (**p = 0.011** for controls; **p = 0.027** for patients). Although controls were on average more biased than patients when there was no cue, the difference didn't reach statistical significance (**p = 0.32**).

In order to explore potential asymmetries in the effect of the cues in patients, we repeated the analysis, considering only this group and separating supporting and contradictory cues. We found that the supporting cues (strong and weak) had a significant (albeit weak) effect on RP (non-parametric Friedman's test: **$\chi^2(2) = 6.5$, p = 0.039**). Interestingly, the effect of the contradictory cues (strong and weak) was not found significant (non-parametric Friedman's test: **p = 0.47**), suggesting that patients might be ignoring dis-confirmatory evidence [30]. A direct comparison between the effect of the strong supporting cue and the strong contradictory cue didn't reach statistical significance (Comparison: [RP(Str Supp) – RP(Amb)] with [RP(Amb) – RP(Contr Cue)]; **p = 0.15**). As a control, this analysis was repeated in healthy subjects: we found that both types of cues significantly affected RP (supporting evidence: **$\chi^2(2) = 19.01$, p < 0.001**; contradicting evidence: **$\chi^2(2) = 14.94$, p < 0.001**) while a direct comparison between the effects of strong opposite cues didn't give significant results (**p = 0.46**).

*Survival Probability*

A similar analysis was repeated for stability (SP), revealing significant effects of the visual cues (**$\beta$ = 0.22, p < 0.001**) and of the response (**$\beta$ = -0.11, p < 0.001**) (participants were more stable when perceiving SFA than when perceiving SFB) but no significant effects of the group (**p = 0.29**). Additionally, we found significant two-way interactions (group x cue: **$\beta$ = -0.19, p =**

**0.003**; response x cue: **β = -0.49, p < 0.001**; group x response interaction was not found significant (**p = 0.58**)), meaning that patients (**Figure 3b**; red curves) were less affected by cues than controls (**Figure 3b**; blue curves) and that cues had opposite effects on the two SP. Finally, we also evidenced a significant cue x group x response 3-way interaction (**β = 0.38, p < 0.001**). As with RP, there was no significant difference in the stability of the two groups in the ambiguous condition, for any of the two responses (although SP(SFA) was larger for controls, compared to patients; SFA: **p = 0.11**; SFB: **p = 0.77**).

Likewise RP, in order to test whether there were asymmetries in the effects of the cues on stability, we used Friedman's test separately for supporting and contradictory cues, individually for patients and healthy controls. For the "*Seen From Above*" interpretation (SFA), we found significant effects both for the supporting cues ($\chi^2(2) = 11.76$, **p = 0.003**) and for the contradictory cues ($\chi^2(2) = 7.18$, **p = 0.03**) for controls, but also for patients (Supp: $\chi^2(2) = 6.17$, **p = 0.045**; Contr: $\chi^2(2) = 6$, **p = 0.049**). Note however that for patients, the effect of the contradictory cue was in the opposite direction, denoting again a potential bias against disconfirmatory evidence. For the "*Seen From Below*" interpretation, there was a significant effect of the contradictory cues ($x^2(2) = 11.41$, **p = 0.003**) but no significant effect for the supporting cues (**p = 0.49**) for healthy controls. For patients on the other hand, neither the supporting cues (**p = 0.37**) nor the contradictory cues (**p = 0.26**) changed significantly the persistence of SFB.

### *2nd Levelt's law*

Overall, in healthy participants, we observed a stronger effect of the cues on the stability of the interpretation that is supported by the contrast, compared with the opposite interpretation. That is in agreement with the revised 2nd Levelt's law, suggesting that "Manipulations of stimulus strength of one perceptual interpretation of a bistable stimulus will mainly influence the average dominance duration of the perceptual interpretation corresponding to the strongest stimulus" [26,31]. Interestingly, this proposition seems to be violated in patients with schizophrenia, potentially due to a bias against disconfirmatory evidence [30] (see also **Figure S2**).

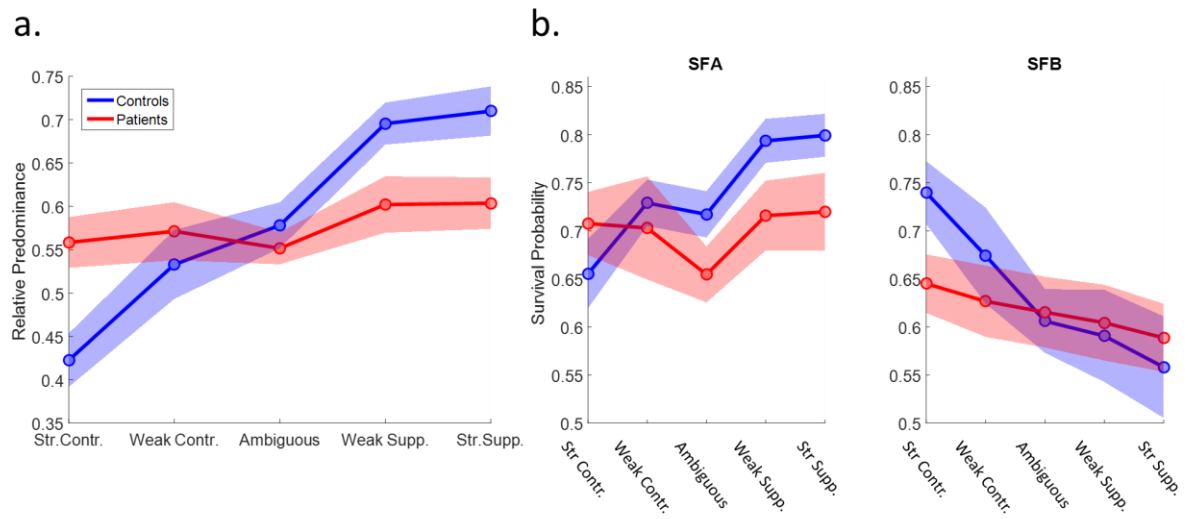


***Figure 3: Results of the CNC experiment.*** ***(a.):*** *RP as function of the cue condition. Patients were less biased by visual cues, compared to healthy controls. A (group x cue) interaction was observed. Patients were not affected by contradictory evidence.* ***(b.):*** *SP (calculated separately for the two interpretations) as a function of the cue condition. Patients were less stable than healthy participants when perceiving the interpretation that was supported by the cue.*

### *Medication*

In order to control for a potential confounding effect of medication on the previous results, OLZ-eq and DZP-eq were added to the model as fixed effects (separately and all together). None were significant (or affected the rest of the results), except for a trend for OLZ (**$\beta$ = -0.0017, p = 0.052**). It suggests that an increase in antipsychotic dosage decreases RP. In order to further understand this effect, we looked for correlations between the OLZ-eq and the RP for the different cue conditions, using Spearman's rank correlation coefficient. We found no significant correlations. For the sake of completeness, we also tried Pearson's correlation. Interestingly, we found a significant negative correlation between OLZ and the RP for the strong contradictory cue (**$r$ = -0.77, p = 0.003**), suggesting that medication might reverse the bias against dis-confirmatory evidence, mentioned before. In any case, a larger sample could help us decide whether those results are meaningful, or an artefact due to the small sample.

### *Severity of symptoms / non-clinical psychotic traits*

A similar method was used to assess the relationship that may exist between the symptoms' severity (and the non-clinical traits) and bistable perception. Neither PDI, nor LSHS

yielded significant results when added as covariates to the linear models. Conversely, we found a significant positive effect of the depressed factor of PANSS on RP (**β = 0.012, p = 0.042**), as well as a trend for the positive factor (**β = 0.008, p = 0.068**), meaning that more severe symptoms (positive or affective) are related with higher RP. In addition to that, we found that when combining patients and controls, both PDI and LSHS were positively correlated with the RP for strong contradictory cue (PDI: $r_s$ **= 0.46, p = 0.013**; LSHS: $r_s$ **= 0.37, p = 0.048**), suggesting an enhancement of the bias against dis-confirmatory bias in participants with stronger psychotic traits (**Figure 4(a.,b.)**). Moreover, LSHS also correlated negatively with the RP for the weak supporting cue ($r_s$ **= 0.41, p = 0.028**) (**Figure 4d**), offering evidence for a link between hallucinations and a weakened effect of visual cues (see **Figure 2**). Finally, we found a positive correlation between the positive factor of PANSS and RP for strong contradictory cues ($r_s$ **= 0.68, p = 0.015**) (**Figure 4c**), in agreement with our previous observations.

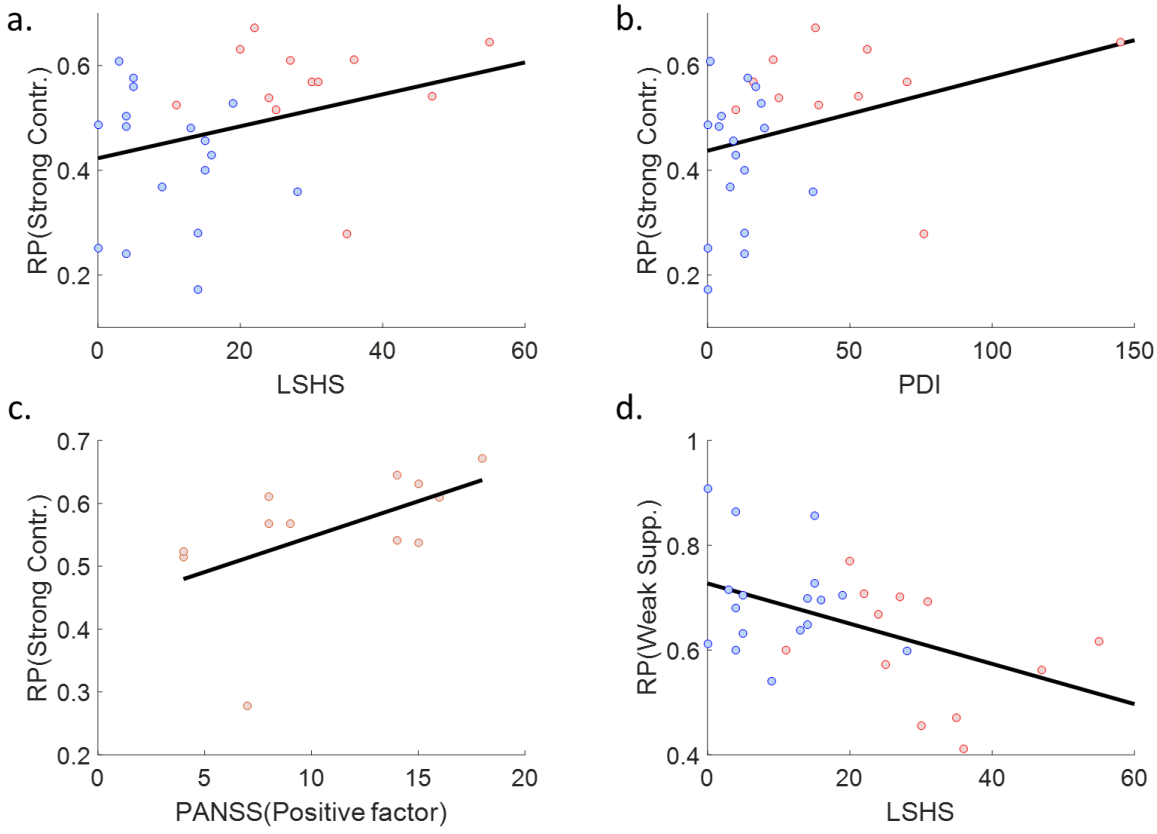


***Figure 4: Correlations in CNC experiment. (a.,b.,c.):*** *Non-clinical psychotic traits (**(a.):** non-clinical hallucinations (LSHS); **(b.)**: non-clinical beliefs (PDI)) and the severity of the positive symptoms in patients **(c.)** were associated with a larger RP (closer or above chance) in the presence of strong contradictory cues, potentially denoting an enhanced bias against dis-confirmatory evidence. **(d.):** Non-clinical hallucinations (LSHS) were also positively correlated with RP in the presence of a weak supporting cue, linking the false*

*percepts with the slope of the RP curve. Healthy controls are represented by blue dots and patients by red dots.*

### *Model-based interpretation*

Our results are compatible with our first hypothesis that psychotic symptoms in schizophrenia are caused by an aberrant amplification of sensory evidence in the cortical hierarchy (climbing loops; **Figure 2(d.,e.)**) [13]. Interestingly they could also be explained by dysregulated temporal statistics, in particular by a system with pathologically high transition rates, for example a system that over-estimates the environmental volatility (**Figure 2(g.,h.)**). On the contrary, the present findings rule out the alternative hypothesis that schizophrenia's symptoms may result from over-counted priors, due to the presence of descending loops (**Figure 2(a.,b.)**). Similarly, they rule out the possibility that patients are simply less (or more) biased, as a result of smaller (or larger) difference between the two transition rates (**Figure S1**).

## Experiment 2 (DNC)

**Figure 5** illustrates the stabilization curves for SFA (left panel) and SFB (right panel). We observe two main differences between schizophrenia patients (red curves) and healthy subjects (blue curves): First, patients' SFB curve converges to a lower value, closer to chance. Second, patients get more destabilized than controls in the first (destabilization) part, for both interpretations. Importantly, our results in healthy participants largely replicated our previous results reported in **Chapter 4**.

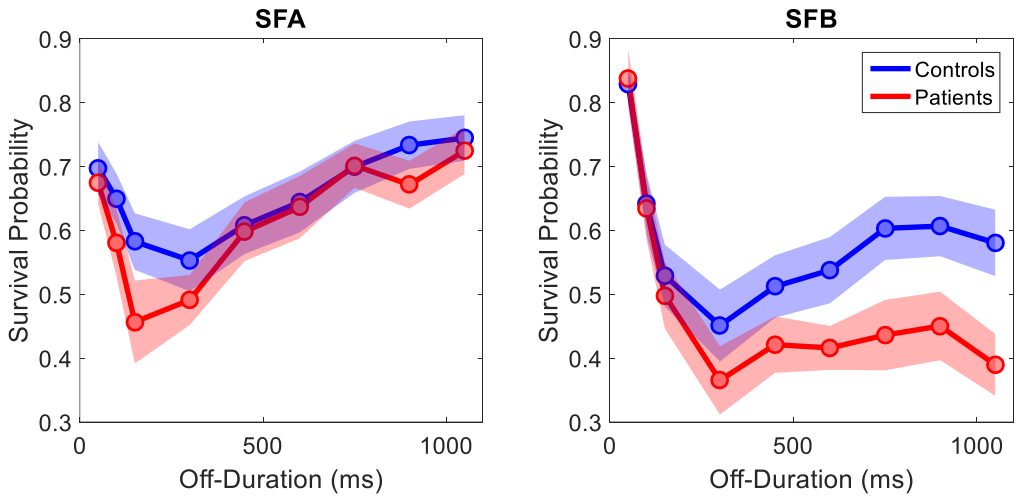

***Figure 5: Stabilisation curves from DNC experiment.*** *The curves were separately plotted for the two interpretations. Patients exhibited a trend for enhanced destabilisation for short OFF-Durations (both interpretations) as well as less stability for longer durations (SFB / Weak interpretation). We observed no difference between the groups in the stabilisation of the strong interpretation (SFA).*

*Stabilization curves – Destabilization part*

As explained in the **Methods**, since blank duration has a non-linear effect on SP, we used different linear mixed models for the destabilisation part (50 ms:150 ms) and the stabilization part (300 ms:1050 ms)of the curves. The LMEM for the destabilisation part gave a significant effect of the OFF-Duration (**β = -0.001, p = 0.003**; the negative effect denotes destabilisation) and of the response (**β = 0.208, p < 0.001**; SFB is more stable, possibly due to a slower destabilisation), on the other hand there was no significant effect of the group (**p = 0.62**). In terms of 2-way interactions, we found a significant (OFF-Duration x response) interaction (**β = -0.002, p < 0.001**; SFB decreases more steeply) and a trend for the (OFF-Duration x group) interaction (**β = -0.001, p = 0.089**; patients get destabilised more quickly than controls), while (group x response) interaction didn't reach statistical significance (**p = 0.99**). Finally, there was no 3-way interaction either (**p = 0.37**).

In order to verify whether the amount of destabilization was different between the two groups, we compared their SP in the relevant conditions (OFF = 150 ms and OFF = 300 ms), in which maximum destabilization occurs. Although none of those comparisons reached statistical significance (SFA, OFF = 150ms: **p = 0.18**; SFB, OFF = 150ms: **p = 0.75**; SFA, OFF = 300ms: **p = 0.43**; SFB, OFF = 300ms: **p = 0.28**), we highlight that patients were always more destabilised than controls.

*Stabilization curves – Stabilization part*

Interestingly, the LMEM for the stabilisation part revealed a significant effect of the OFF-Duration (**β = 0.0003, p < 0.001**; the positive effect denotes stabilisation) while there was also a trend for a 3-way interaction (**β = -0.0002, p = 0.076**; the (Off-Duration x group) interaction is more pronounced for SFB than for SFA). All the other effects were found not significant (group: **p = 0.59**; response: **p = 0.18**; (Off-Duration x group): **p = 0.84**; (Off-Duration x response): **p = 0.18**; (group x response): **p = 0.98**).

We also tested whether the convergence points were symmetrical, separately for patients and controls (comparison: SP(SFA; OFF = 1050 ms) with (1 – SP(SFB; OFF = 1050 ms))) [29]. Asymmetrical convergence denotes the presence of an acute bias, a characteristic of the descending loops (see **Chapter 4**; [32]). The result was significant for controls (**p = 0.002**), while for patients we only found a trend (**p = 0.07**), potentially due to the small sample. Additionally, we found that both groups' SP for SFB and OFF = 1050 ms were not significantly different from chance (controls: **p = 0.24**; patients: **p = 0.055**), on the contrary SP for SFA were significantly above chance (controls: **p < 0.001**; patients: **p = 0.008**). Finally, the convergence point of the SFB curve, but not that of the SFA curve was significantly different between the two groups (SP(SFB; OFF = 1050 ms): **p = 0.02**; SP(SFA; OFF = 1050 ms): **p = 0.68**).

*Individuals*

**Figures S3** and **S4** illustrate the stabilization curves of the 14 controls and the 9 patients respectively, individually for each participant. As in our previous experiments (**Chapter 4**), although results of individuals are noisier, we observe that the average pattern (blue above red; initial destabilization followed by stabilization; non-symmetrical convergence points) is also present for most subjects, suggesting that the results presented above are indeed meaningful. However, an interesting observation is that almost half of the patients' curves converge to symmetrical points, indicating the lack of memorisation / acute bias, which in the circular inference framework can be interpreted as lack of descending loops. This between patients variability could have important theoretical and clinical implications (see **Chapter 7**) and needs to be further investigated in our future work (participant per participant fitting, categorization of patients in subgroups, based on (the phenomenology of) their symptoms etc.).

*Reaction time*

Furthermore, we ensured that between-groups differences in reaction times (RT) did not affected our results. **Figure S5** presents the average reaction times for the two groups and for the different OFF-Durations. A comparison between the two groups, after collapsing all the OFF-Durations, revealed only a statistical trend (**p = 0.055**) (note that an analysis taking into account within-group differences (different OFF-Durations) gave the same result regarding the group effect). To make sure that our results were not contaminated by very long reaction times (the distribution of RT in patients has a longer tail), we split the results into LRT and HRT

(threshold: RT = 0.7 s) and repeated the analysis for each category separately. Those stabilization curves are presented in **Figure S6**. As expected (longer reaction times imply more accumulation of evidence, hence SP being flatter and closer to chance), we found different profiles for the two subgroups: HRT curves (**Figure S6(a.,c.)**) were almost flat and close to 0.5; more importantly, LRT curves (**Figure S6(b.,d.)**) did not differ qualitatively from the ones presented before. A similar model-free analysis gave similar results as for the entire sample for the destabilization part, while for the stabilization part the effects of the response (**β = -0.14, p = 0.03**; SFA is more stable than SFB) and the 3-way interaction (**β = -0.0003, p = 0.004**) became significant.

*Medication, severity of symptoms and non-clinical psychotic traits*

We found no significant effects of medication, for any type of equivalent dosage (OLZ, DZP). Additionally, results were not significant, when we added PDI or LSHS in the LMEM (overall or LRT) as covariates. Nevertheless, we found a significant positive effect of the positive factor of PANSS in the destabilization phase (**β = 0.014, p = 0.03**; this effect did not survive in LRT), but also significant positive effects of the positive factor (**β = 0.016, p < 0.001**) and the negative factor (**β = 0.014, p = 0.03**; this effect also did not survive in LRT) of PANSS in the stabilization phase. In the correlation analysis that followed, we found a significant negative correlation between LSHS (controls and patients taken together) and the stabilization point of the SFB curve (SP(SFB; OFF=1050ms)) (**$r_s$ = -0.51, p = 0.01**), substantiating a potential link between hallucinations and the decreased stabilisation of the weak (SFB) interpretation (**Figure 6**).

*Model-based interpretation*

The results of the DNC experiment are largely in agreement with our previous results (CNC experiment; [13]), suggesting that schizophrenia (and psychotic symptoms in particular) is a manifestation of over-counted sensory information (climbing loops) Note that this interpretation does not explain why patients were more destabilized for short blank intervals. This result could signify the presence of a secondary impairment in the representation of the temporal statistics of the environment (increased rates).

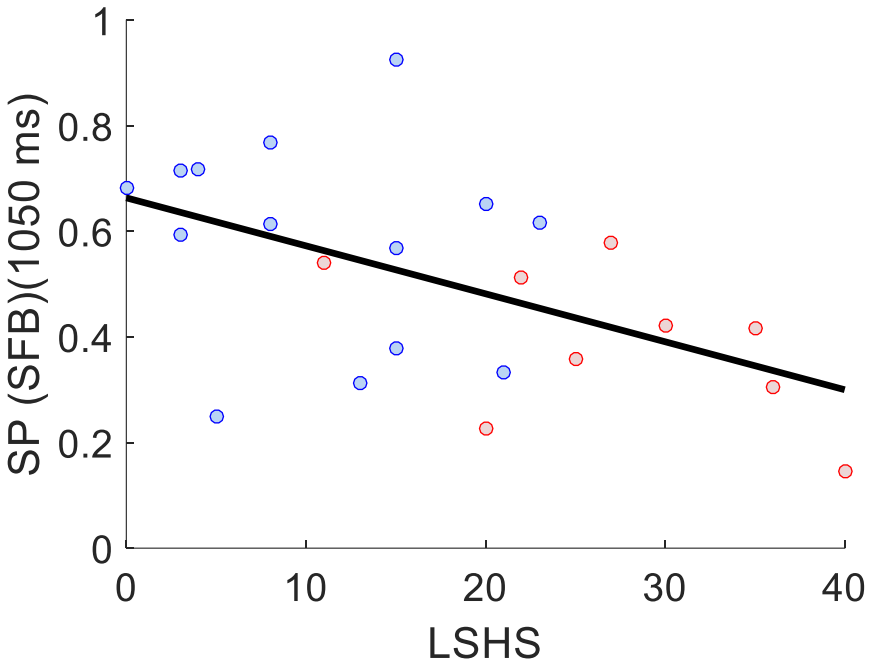


***Figure 6: Correlations for DNC experiment.*** *The reduced stabilisation of the SFB interpretation for long blank intervals (OFF = 1050 ms) (that corresponds to SP being closer to 0) correlated with the non-clinical psychotic traits in the entire sample (non-clinical hallucinations (LSHS)). Healthy controls are represented by blue dots and patients by red dots.*

## Discussion

In the present work, we compared schizophrenia patients with matched healthy controls in two bistable perception tasks: a continuous-presentation task (CNC), inspired by [14] and an intermittent-presentation task (DNC), introduced earlier in this thesis (**Chapter 4**). The two studies were largely based on the circular inference framework, a form of probabilistic inference in which information can be amplified because of the presence of loops [8,10]. Having established theoretical (see **Chapter 3** but also [8,33]) and empirical links (see **Chapters 2,4** and [13]) between circularity and both schizophrenia and bistability, here we sought to use perception of ambiguous stimuli as a tool to probe the mechanisms that generate aberrant inference in the disorder. More particularly, we asked to what extent sensory amplification (rather than prior amplification) can explain psychotic symptoms and whether additional mechanisms might be at play. Additionally, we could address the role of dynamics, largely neglected in previous CI work, in the generation of the symptomatology.

Despite limited samples, we evidenced significant differences between the groups in the way they perceive ambiguity, suggesting that the perceptual mechanisms, responsible for

bistability, are seriously impaired in schizophrenia. In the CNC experiment, we discovered that patients are less biased by visual cues, while they are also less stable than healthy participants. Crucially, instead of a simple additive group effect on RP (like the one presented in **Figure S1**), we observed a multiplicative effect (interaction; see also **Figure 2**), which rules out the possibility that patients simply exhibit a weaker implicit preference (**Figure S1**).

In addition to that, we found preliminary results for a "bias against disconfirmatory evidence" in patients [30], although our limited statistical power did not allow for a clear rejection of the possibility that such an asymmetry is a general trend in both groups. Interestingly, we found that this bias was exacerbated in participants with more severe psychotic traits / symptoms, while it was also found reduced with higher dosage of medication. Our discovery of a negative correlation between non-clinical hallucinations' traits (LSHS) and the RP in a supporting-cue condition strengthened our claim for a link between the reduced slope of the RP and the psychotic symptoms.

Additionally, the DNC experiment revealed differences between the two groups in both parts of the stabilization curves. In the early destabilization part, we found hints (the difference didn't reach statistical significance) of faster and stronger destabilization (reduction of the SP) in patients, as compared with healthy participants. On top of that, we observed a weaker stabilization of the SFB interpretation in patients, which also correlated with the non-clinical hallucinations' traits (LSHS).

The present results are largely in agreement with previous experiments, which used bistable perception to study different psychopathologies such as schizophrenia and bipolar disorder. Most studies found increased reversal rates and reduced cognitive control both in schizophrenia and in bipolar disorder ([34,35]; but see also [36]), which has been linked to increased effects of noise [37]. More recently, Schmack and colleagues created a bridge between psychotic symptoms, bistable perception and impaired predictive processing [38]: they associated non-clinical bizarre beliefs (PDI score) in healthy subjects with reduced stability in an intermittent presentation task on one hand and with an enhanced belief-induced bias on the other. They also suggested a two-level explanation including weak low-level predictions and strong high-level predictions. In two follow-up studies, they extended their discontinuous-presentation result in a group of schizophrenia patients [39] but failed to do the same for the high-level predictions [40].

Our results replicated the findings regarding stability in continuous- and discontinuous-presentation bistable tasks and extended them, first by considering more conditions (e.g. visual

cues in CNC and multiple blank durations (instead of one single interval) in DNC) and second by suggesting methodological improvements. As explained in **Chapter 2**, our continuous-presentation method [14] minimizes the role of attention which has been shown to play a crucial role in bistable perception [41,42], a function known to be impaired in schizophrenia [43]. Second, the present method solves the problems posed by potential motor deficits in patients: This procedure is not affected by differences in reaction time, as one could use the time of the sound as a proxy for the time of the decision. Regarding the methodological advantages of our discontinuous-presentation experiment, we briefly mention the use of SP instead of reversal rates, the consideration of potential chronic biases and asymmetries between the two interpretations, the reconstruction of the entire stabilization curve and the randomization of the blank intervals. For more details, please refer to the corresponding section in **Chapter 4.**

Today, there is an abundance of evidence suggesting that the brain is a probabilistic machine, constantly guessing what is out there, what could be the result of a decision or what is the most probable outcome of an action [44–47]. To optimize predictions, it must combine sensory information with constantly updated priors, which can be mapped on feed-forward and feedback processing respectively [48]. Taken together, our results speak to a profound alteration in predictive mechanisms in schizophrenia. In a previous study, Jardri and colleagues used a variation of the beads task to probe those alterations [13]. They concluded that schizophrenia is mainly caused by an aberrant amplification of the sensory evidence (i.e. climbing loops), which results in suboptimal inferences and consequently in biased decisions (e.g. "jumping to conclusions", common in patients with prominent delusions [49,50]), false percepts but also bizarre and unshakable beliefs [8]. The results presented here are compatible with such an explanation, whose qualitative predictions are presented in **Figure 2(d.,e.,f.)**. On the contrary, our results contradict the hypothesis that schizophrenia is due to prior-amplification (**Figure 2(a.,b.,c.)**). It's worth noting though that descending loops can also generate hallucinations and other false percepts and beliefs and might underlie other types of psychotic experiences (see **Chapter 7**). As a result, the possibility that patients have less descending loops (reduced over-counting of priors) seems less plausible.

In **Chapter 3**, we showed that in the context of *dynamical circular inference*, climbing loops essentially increase the gain of the noisy sensory evidence. Interestingly, this gain depends not only on the climbing loops, but also on the properties of the internal model (mean and variance of the likelihood function, feed-forward weight etc.). We highlight that an impairment in any of those parameters would have the same effect on the behaviour of the patients, consequently the present study cannot vote for one of those alternatives.

Despite the good qualitative agreement, we must highlight that certain patterns in the data do not fit in the above interpretation. First, in the DNC experiment we did not observe a reduced stabilization in patients in the SFA interpretation (**Figure 5**; left panel), as we would expect from a system with stronger climbing loops. Although this could be an important deviation from the model's predictions, we note that for oFF=1050ms, both SP still increase, consequently a difference might appear later. A follow-up experiment, testing longer intervals, would be necessary to clarify this point.

A second discrepancy is found in the destabilization profiles of the same curves. More particularly, we observed enhanced destabilization for patients, despite the climbing loops predicting the opposite (**Figure 2f**). Although this result did not reach significance in our preliminary sample, it's highly probable that it will become significant in a larger sample. What could have caused this effect? One possibility is that patients, apart from over-counting their sensory information, also overestimate the environmental volatility (increased baseline rates; **Figure 2i**). A system with more leak (in which the difference between the rates is kept constant), would accumulate less information, ultimately exhibiting a stabilization profile similar to the observed one. Note that apart from the difference in the destabilization, a system with increased rates behaves qualitatively in exactly the same as a system with strong climbing loops.

Could a single impairment in the rates underlie all the observed differences? Although the present results seem consistent with such an interpretation, we argue that this is highly improbable. In the work by Jardri and colleagues, dynamics was irrelevant, as a result a difference in the rates wouldn't have produced the observed deviation between patients and controls [13].

The idea that patients have increased baseline transition rates might seem at odds with some recent results, suggesting that psychotic patients underestimate environmental volatility [51]. This contradiction might be due to differences in the experimental design. More particularly, in the Powers et al study, the contingencies were actually changing, as a result participants had to learn online how quickly they change. Conversely, in the present study switches do not correspond to real events, hence the rates are not learnt during the task, but vaguely reflect an a priori estimation of the temporal statistics of the environment.

Apart from those two major disagreements, we also found preliminary proof for a "bias against dis-confirmatory evidence" (BADE), which is also not directly predicted by our dCI model. A BADE has been shown a prominent characteristic of delusional patients [30,52] and healthy individuals with high delusions-proneness [53] in reasoning tasks and has been linked

to the generation and maintenance of fixed beliefs [54,55]. To the best of our knowledge, the presence of such a bias in perception has been largely neglected. If the bias survives in the final sample (and is shown to be a unique characteristic of patients), it could signify a secondary impairment or potentially a compensatory mechanism, reducing the constant update of the high-level beliefs by the amplified sensory information, at the cost of throwing away useful information. Alternatively, it could be a by-product of climbing loops, in particular a consequence of the aberrant learning caused by the reverberated sensory input [8].

A few limitations must be mentioned. First, as mentioned earlier, there is a problem of statistical power, due to the preliminary character of the results. We expect to have 30 participants per group / per experiment in the final sample, which is large enough (according to sample size estimations) to address the open questions mentioned above.

Second, all our interpretations are based on qualitative comparisons between the data and simulations. Although this approach gives important (albeit rough) intuitions regarding the underlying mechanisms, a model fitting procedure (at the level of individuals) is necessary in order to arbitrate between the different possible interpretations. This quantitative account could demonstrate the respective roles of loops and rates in the observed behaviour and account for differences between different subgroups of patients (e.g. patients with auditory hallucinations vs patients with multi-modal, audio-visual hallucinations). Such an analysis is scheduled for when we reach the final sample-size.

Finally, an impaired contrast sensitivity in patients [56] could have contributed to the observed difference in the CNC experiment. Such an impairment has been linked to a more general gain control problem in schizophrenia, potentially mediated by a hypo-function of the NMDA receptors [57,58]. In order to control for that, an additional test will have to be added in the battery of neuropsychological tests already applied, explicitly testing the capacity of participants to perceive the contrast between the lines.

In summary, this study brings new evidence to the long lasting debate of whether inferences in schizophrenia are mainly prior- or sensory-driven [4–7,13,39,51,59,60]. Beyond the algorithmic differences (belief propagation vs predictive coding etc.) and the nature of the impairment (loops vs impairment in precision-weighting etc.), our preliminary evidence support the view that psychosis is the result of aberrantly strong bottom-up processing. Further work is necessary in order to associate the phenomenology of the symptoms (macro-scale) with brain computations on one hand (meso-scale) and neural processing on the other hand (micro-scale), resulting in a holistic, multi-scale account of psychosis.

## Supplementary Material

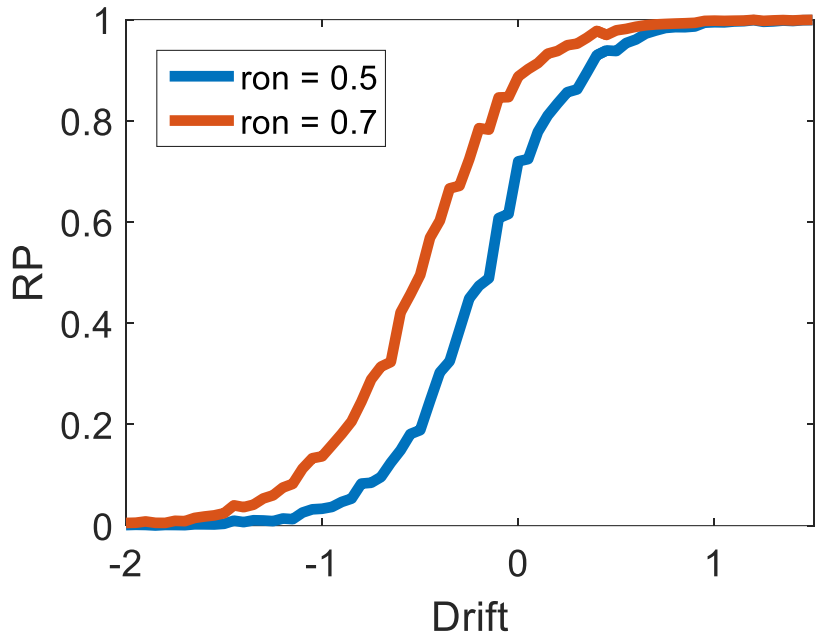


***Figure S1: Predictions for different ron.*** *Unlike changing the strength of loops (climbing and descending) or the value of both baseline rates (**Figure 2**), altering only one of the rates (e.g. baseline ron) results in a shift of the psychometric curve to the left or to the right (additive effect) but not in a change of the slope (multiplicative effect).*

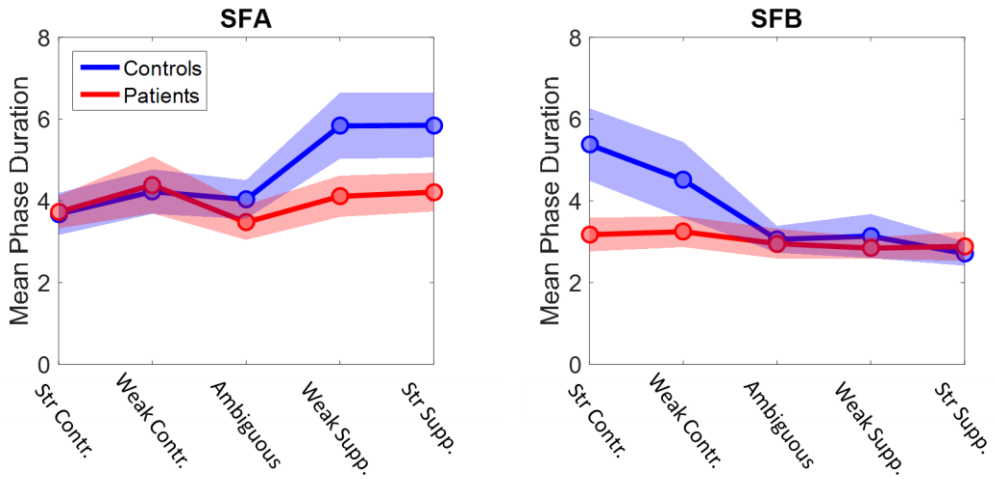


***Figure S2: Mean phase durations (MPD) in the CNC experiment.*** *This figure complements **Figure 3** in the main text, showing that patients violate Levelt's revised second proposition, which dictates that a change in the stimulus strength (e.g. by adding a visual cue) affects mainly the MPD of the currently stronger interpretation. Patients are much less affected by cues compared to controls, while they seem to completely ignore contradictory evidence.*

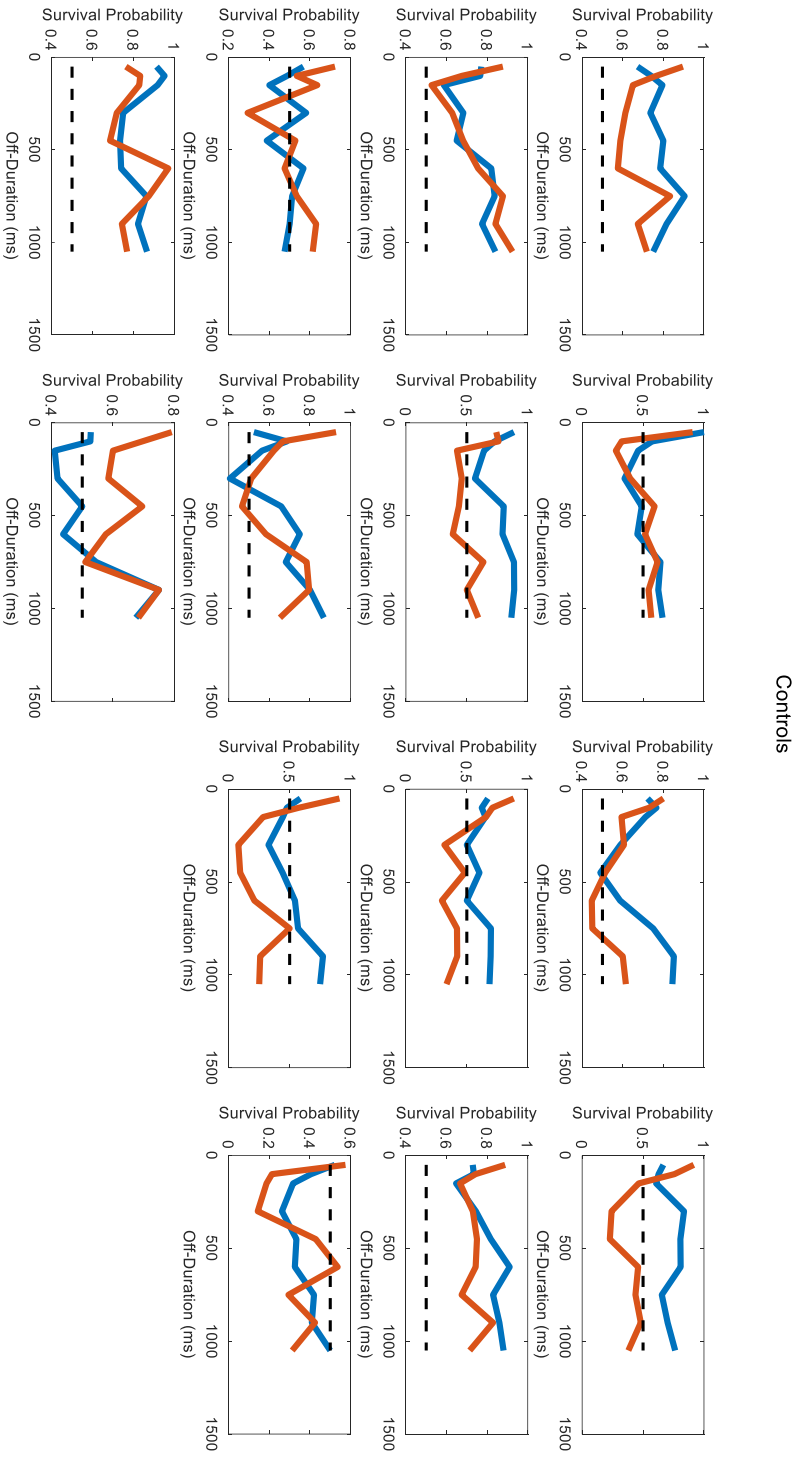


*Figure S3: Stabilization curves for controls (individuals).*

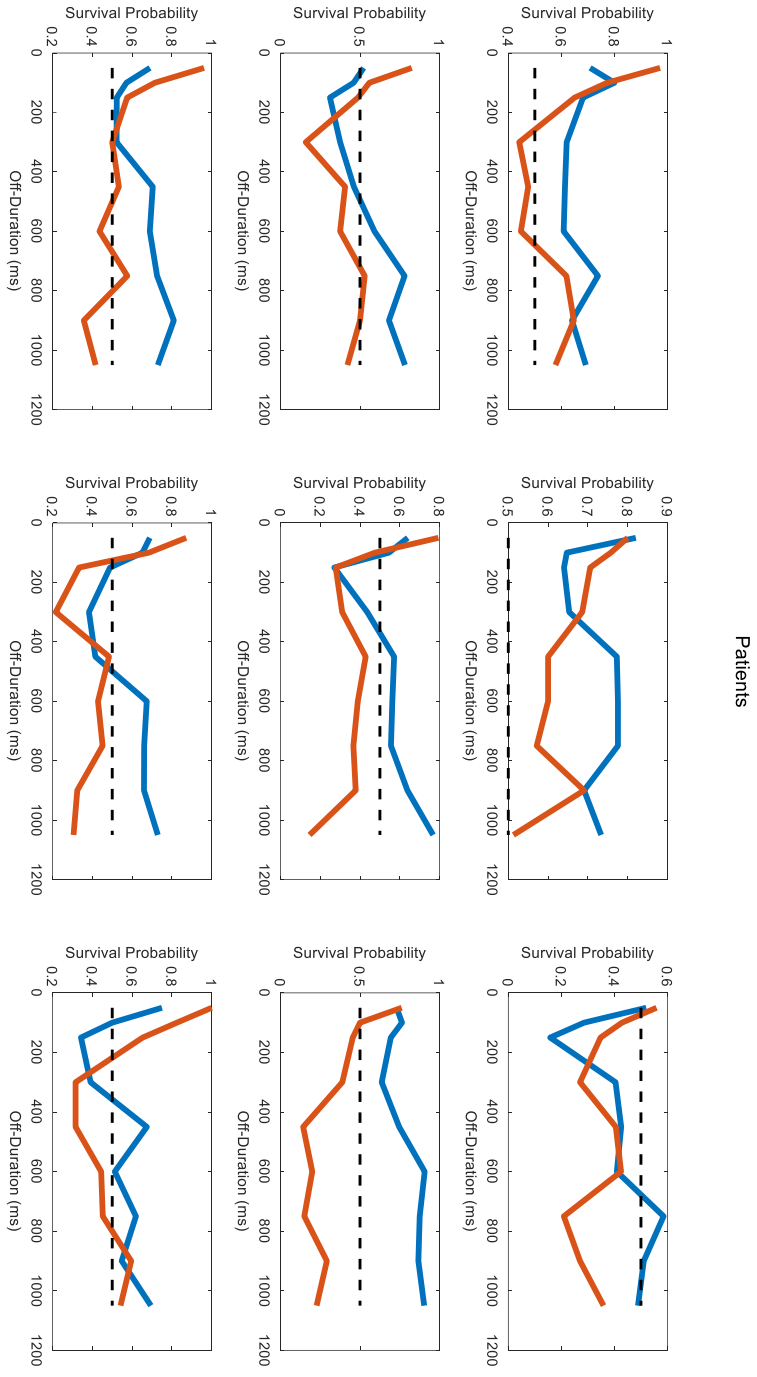


*Figure S4: Stabilization curves for patients (individuals).*

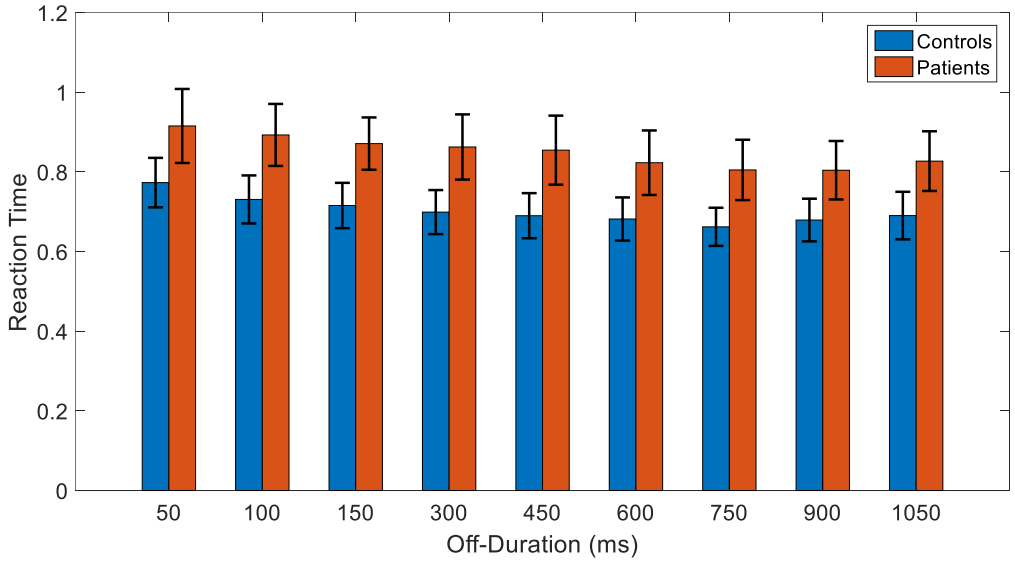


***Figure S5: Reaction times for different OFF-Durations in the DNC experiment.*** *As expected, patients were slower than controls (statistical trend), while there was also a small effect of cue (RT decreases for longer intervals).*

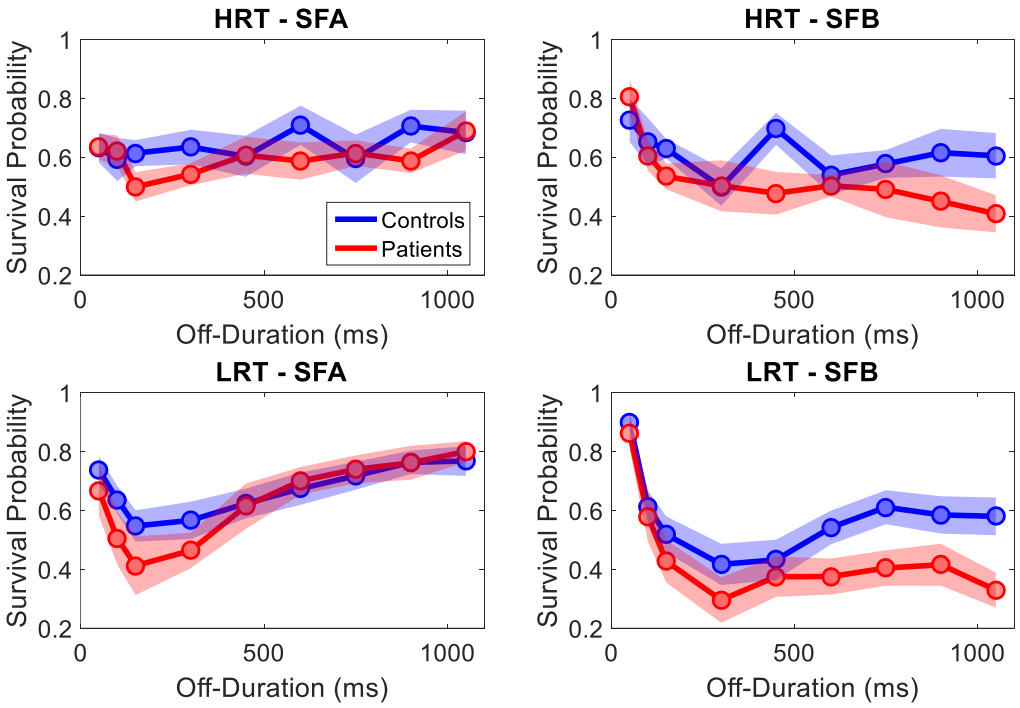


***Figure S6: Results of DNC experiment for LRT and HRT subgroups.*** *As expected, the HRT curves were flatter and closer to chance. On the other hand, LRT curves had the same shape as the overall curves* ***(Figure 5)***

## References


1. McGrath J, Saha S, Chant D, Welham J. Schizophrenia: A concise overview of incidence, prevalence, and mortality. Epidemiol Rev. 2008;30: 67–76. doi:10.1093/epirev/mxn001

2. Crow TJ. Positive and negative schizophrenia symptoms and the role of dopamine. Br J Psychiatry. 1981;139: 251–254. doi:10.1192/bjp.139.3.251

3. Liddle PF. Schizophrenic syndromes, cognitive performance and neurological dysfunction. Psychol Med. 1987;17: 49–57. doi:10.1017/S0033291700012976

4. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci. 2009;10: 48–58. doi:10.1038/nrn2536

5. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. Front psychiatry. 2013;4: 47. doi:10.3389/fpsyt.2013.00047

6. Karvelis P, Seitz AR, Lawrie SM, Seriès P. Autistic traits, but not schizotypy, predict increased weighting of sensory information in Bayesian visual integration. Elife. 2018;7. doi:10.7554/eLife.34115

7. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The Predictive Coding Account of Psychosis. Biol Psychiatry. Elsevier Inc; 2018; 1–10. doi:10.1016/j.biopsych.2018.05.015

8. Jardri R, Denève S. Circular inferences in schizophrenia. Brain. 2013;136: 3227–41. doi:10.1093/brain/awt257

9. Jardri R, Hugdahl K, Hughes M, Brunelin J, Waters F, Alderson-Day B, et al. Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? Schizophr Bull. 2016;42: 1124–1134. doi:10.1093/schbul/sbw075

10. Leptourgos P, Denève S, Jardri R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Curr Opin Neurobiol. 2017;46: 154–161. doi:10.1016/j.conb.2017.08.012

11. Lisman J. Excitation, inhibition, local oscillations, or large-scale loops: What causes the symptoms of schizophrenia? Curr Opin Neurobiol. Elsevier Ltd; 2012;22: 537–544. doi:10.1016/j.conb.2011.10.018

12. Foss-Feig JH, Adkinson BD, Ji JL, Yang G, Srihari VH, McPartland JC, et al. Searching for Cross-Diagnostic Convergence: Neural Mechanisms Governing Excitation and Inhibition Balance in Schizophrenia and Autism Spectrum Disorders. Biol Psychiatry. Elsevier Inc.; 2017;81: 848–861. doi:10.1016/j.biopsych.2017.03.005

13. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference

in schizophrenia. Nat Commun. 2017;8: 14218. doi:10.1038/ncomms14218

14. Mamassian P, Goutcher R. Temporal dynamics in bistable perception. J Vis. 2005;5: 361–75. doi:10:1167/5.4.7

15. Orbach J, Ehrlich D, Heath HA. Reversibility of the Necker Cube: I. An examination of the concept of "satiation of orientation." Percept Mot Skills. 1963;17: 439–458. doi:10.2466/pms.1963.17.2.439

16. Leopold DA, Wilke M, Maier A, Logothetis NK. Stable perception of visually ambiguous patterns. Nat Neurosci. 2002;5: 605–609. doi:10.1038/nn851

17. International statistical classification of diseases an related health problems. 10th ed. World Health Organization; 2010.

18. Sheehan D V., Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The mini-international neuropsychiatric interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry. 1998;59: 22–33.

19. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. Schizophr Bull. 1987;13: 261–276. doi:10.1093/schbul/13.2.261

20. Wallwork RS, Fortgang R, Hashimoto R, Weinberger DR, Dickinson D. Searching for a consensus fi ve-factor model of the Positive and Negative Syndrome Scale for schizophrenia. Schizophr Res. Elsevier B.V.; 2012;137: 246–250. doi:10.1016/j.schres.2012.01.031

21. Peters E, Joseph S, Day S, Garety P. Measuring Delusional Ideation : The 21-Item Peters et aL Delusions Inventory ( PDI ). Schizophr Bull. 2004;30: 1005–1022.

22. Laroi F, Marczewski P, Van der Linden M. Further evidence of the multi-dimensionality of hallucinatory predisposition: factor structure of a modified version of the Launay-Slade Hallucinations Scale in a normal sample. Eur Psychiatry. 2004;19: 15–20.

23. Leucht S, Samara M, Heres S, Patel MX, Furukawa T, Cipriani A, et al. Dose Equivalents for Second-Generation Antipsychotic Drugs : The Classical Mean Dose Method. Schizophr Bull. 2015;41: 1397–1402. doi:10.1093/schbul/sbv037

24. Sostmann HJ, Sostmann H, Crevoisier C, Bircher J. Dose equivalence of midazolam and triazolam. A psychometric study based on flicker sensitivity, reaction time and digit symbol substitution test. Eur J Clin Pharmacol. 1989;36: 181–187.

25. Sundareswara R, Schrater P. Perceptual multistability predicted by search model for Bayesian decisions. J Vis. 2008;8: 1–19. doi:10.1167/8.5.12.Introduction

26. Brascamp JW, Klink PC, Levelt WJM. The 'laws' of binocular rivalry: 50 years of Levelt's propositions. Vision Res. 2015;109: 20–37. doi:10.1016/j.visres.2015.02.019

27. Gigante G, Mattia M, Braun J, Del Giudice P. Bistable perception modeled as competing stochastic integrations at two levels. PLoS Comput Biol. 2009;5: 1–9. doi:10.1371/journal.pcbi.1000430

28. Kim CY, Lee G, Choi H, Goh J. Repeated Measures of Reaction Times among Patients with Schizophrenia. Clin Psychopharmacol Neurosci. 2009;7: 20–22.

29. Kogo N, Hermans L, Stuer D, van Ee R, Wagemans J. Temporal dynamics of different cases of bi-stable figure-ground perception. Vision Res. 2015;106: 7–19. doi:10.1016/j.visres.2014.10.029

30. Woodward TS, Moritz S, Cuttler C, Whitman JC. The Contribution of a Cognitive Bias Against Disconfirmatory Evidence ( BADE ) to Delusions in Schizophrenia. J Clin Exp Neuropsychol. 2006;28: 605–617. doi:10.1080/13803390590949511

31. Klink PC, van Ee R, van Wezel RJ a. General validity of Levelt's propositions reveals common computational mechanisms for visual rivalry. PLoS One. 2008;3: e3473. doi:10.1371/journal.pone.0003473

32. Al-dossari M, Blake R, Brascamp JW, Freeman AW. Chronic and acute biases in perceptual stabilization. J Vis. 2015;15: 1–11. doi:10.1167/15.16.4.doi

33. Deneve S, Jardri R. Circular inference: Mistaken belief, misplaced trust. Curr Opin Behav Sci. 2016;11: 40–48. doi:10.1016/j.cobeha.2016.04.001

34. Keil A, Elbert T, Rockstroh B, Ray WJ. Dynamical aspects of motor and perceptual processes in schizophrenic patients and healthy controls. Schizophr Res. 1998;33: 169–178.

35. McBain R, Norton DJ, Kim J, Chen Y. Reduced cognitive control of a visually bistable image in schizophrenia. J Int Neuropsychol Soc. 2011;17: 551–6. doi:10.1017/S1355617711000245

36. Miller S, Gynther B, Heslop K, Liu G, Mitchell P, Ngo T, et al. Slow binocular rivalry in bipolar disorder. Psychol Med. 2003;33: 683–692.

37. Hoffman RE, Quinlan DM, Mazure CM, Mcglashan TM. Cortical Instability and the Mechanism of Mania : A Neural Network Simulation and Perceptual Test. Biol Psychiatry. 2001;49: 500–509.

38. Schmack K, Gòmez-Carrillo de Castro A, Rothkirch M, Sekutowicz M, Rössler H, Haynes J-D, et al. Delusions and the role of beliefs in perceptual inference. J Neurosci. 2013;33: 13701–12. doi:10.1523/JNEUROSCI.1778-13.2013

39. Schmack K, Schnack A, Priller J, Sterzer P. Perceptual instability in schizophrenia : Probing predictive coding accounts of delusions with ambiguous stimuli. Schizophr Res Cogn. Elsevier B.V.; 2015;2: 72–77. doi:10.1016/j.scog.2015.03.005

40. Schmack K, Rothkirch M, Priller J, Sterzer P. Enhanced predictive signalling in schizophrenia. Hum Brain Mapp. 2017;38: 1767–1779. doi:10.1002/hbm.23480

41. Toppino TC. Reversible-figure perception: mechanisms of intentional control. Percept Psychophys. 2003;65: 1285–1295. doi:10.3758/BF03194852

42. Li H-H, Rankin J, Rinzel J, Carrasco M, Heeger DJ. Attention model of binocular rivalry. Proc Natl Acad Sci. 2017;114: E6192–E6201. doi:10.1073/pnas.1620475114

43. Galaverna F, Morra CA, Bueno AM. Attention in patients with chronic schizophrenia : Deficit in inhibitory control and positive symptoms. Eur J Psychiatry. 2012;26: 185–195.

44. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annu Rev Psychol. 2004;55: 271–304. doi:10.1146/annurev.psych.55.090902.142005

45. Chater N, Tenenbaum JB, Yuille A. Probabilistic models of cognition: conceptual foundations. Trends Cogn Sci. 2006;10: 287–91. doi:10.1016/j.tics.2006.05.007

46. Ma WJ. Organizing probabilistic models of perception. Trends Cogn Sci. Elsevier Ltd; 2012;16: 511–518. doi:10.1016/j.tics.2012.08.010

47. Kording KP. Bayesian statistics : relevant for the brain ? Curr Opin Neurobiol. Elsevier Ltd; 2014;25: 130–133. doi:10.1016/j.conb.2014.01.003

48. Lochmann T, Deneve S. Neural processing as causal inference. Current Opinion in Neurobiology. 2011. pp. 774–781. doi:10.1016/j.conb.2011.05.018

49. Huq SF, Garety PA, Hemsley DR. Probabilistic judgements in deluded and non-deluded subjects. Q J Exp Psychol. 1988;40A: 801–812. doi:10.1080/14640748808402300

50. Moritz S, Woodward TS. Jumping to conclusions in delusional and non-delusional schizophrenic patients. Br J Clin Psychol. 2005;44: 193–207. doi:10.1348/014466505X35678

51. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science (80- ). 2017;357: 596–600. doi:10.1126/science.aan3458

52. Speechley WJ, Ngan ET, Moritz S, Woodward TS. Impaired Evidence Integration and Delusions in Schizophrenia. J Exp Psychopathol. 2012;3: 688–701. doi:10.5127/jep.018411

53. Woodward TS, Buchy L, Moritz S, Liotti M. A Bias Against Disconfirmatory Evidence Is Associated With Delusion Proneness in a Nonclinical Sample. Schizophr Bull. 2007;33: 1023–1028. doi:10.1093/schbul/sbm013

54. Woodward TS, Moritz S, Menon M, Klinge R. Belief inflexibility in schizophrenia. Cogn Neuropsychiatry. 2008;13: 267–277. doi:10.1080/13546800802099033

55. Corlett PR, Taylor JR, Wang XJ, Fletcher PC, Krystal JH. Toward a neurobiology of

delusions. Prog Neurobiol. Elsevier Ltd; 2010;92: 345–369. doi:10.1016/j.pneurobio.2010.06.007

56. Butler PD, Silverstein SM, Dakin SC. Visual Perception and Its Impairment in Schizophrenia. Biol Psychiatry. 2008;64: 40–47. doi:10.1016/j.biopsych.2008.03.023

57. Kwon Y, Nelson S, Toth L, Sur M. Effect of stimulus contrast and size on NMDA receptor activity in cat lateral geniculate nucleus. J Neurophysiol. 1992;68: 182–196.

58. Stephan KE, Friston KJ, Frith CD. Dysconnection in Schizophrenia : From Abnormal Synaptic Plasticity to Failures of Self-monitoring. Schizophr Bull. 2009;35: 509–527. doi:10.1093/schbul/sbn176

59. Chambon V, Pacherie E, Barbalat G, Jacquet P, Franck N, Farrer C. Mentalizing under influence: abnormal dependence on prior expectations in patients with schizophrenia. Brain. 2011;134: 3728–41. doi:10.1093/brain/awr306

60. Teufel C, Subramaniam N, Dobler V, Perez J, Finnemann J, Mehta PR, et al. Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. Proc Natl Acad Sci. 2015;112: 1–6. doi:10.1073/pnas.1503916112

# Chapter 7

## A multiscale approach to psychedelics based on circular inference

In preparation for publication as:

Leptourgos P., Deneve S., Jardri R., (in prep.). A multiscale approach to psychedelics based on circular inference

## Abstract

Psychotomimetic drugs are known to mimic psychosis in non-clinical individuals. Among these agents, serotonergic agonists (i.e., psychedelics) can distort perception and induce hallucinations without confusion. Importantly, hallucinations induced by psychedelics exhibit features that are profoundly different from those observed in common psychiatric disorders. These experiences are primarily multisensory and often associated with synaesthesia, which appears in clear contradiction with the high prevalence of distressing voices heard by schizophrenia patients. In this paper, we introduce a unifying account for these different experiences based on *circular inference,* a form of suboptimal probabilistic inference in which information gets erroneously amplified due to dysregulations of the neural excitatory-to-inhibitory (E/I) balance. Using in-silico simulations, we show that psychedelics could destabilize the E/I balance in the cortical hierarchy in favour of prior information (a phenomenon named "descending loops"). This prior overcounting tends to accumulate in sensory areas, which become over-integrated. Considering that the brain builds representations through message-passing between neuronal populations, this over-integration gives rise to prior-driven aberrant experiences with a strong crossmodal character. A different mechanism might occur in schizophrenia, based on the formation of "climbing loops" in the cortical hierarchy. By amplifying sensory information and enhancing segregation of the sensory representations, this second form of E/I imbalance results in two well-known features of the disorder, i.e., unimodal hallucinations and a reduced vulnerability to illusions. Crucially, this distinction between drug-induced and schizophrenia-related hallucinations also reveals a missing link between erroneous message-passing and neuromodulation. To fill this gap, we propose a canonical micro-circuit implementing circular belief-propagation in which the two inference loops previously introduced could be controlled by different neuromodulators acting on the E/I balance: the "descending loops", under serotonergic control (the primary target of psychedelics), and the "climbing loops" regulated by dopamine (whose role in the pathophysiology of schizophrenia is widely accepted).

## Introduction

Hallucinations can be defined as percepts occurring while the person is awake and without external stimulation of the relevant sensory organ. Even if these experiences can be observed in non-clinical populations [1], hallucinations often constitute the hallmark of psychiatric disorders, such as schizophrenia [2] or borderline personality [3], and are common symptoms in neurodegenerative diseases [4]. Interestingly, hallucinatory experiences can also be induced using psychotomimetic drugs.

A particular class of hallucinogenic drugs, known as "classic psychedelics" [5], has fascinated science for more than a century. Major psychedelics include naturally occurring chemicals such as mescaline (extracted from the peyote cactus), psilocybin ("magic mushrooms") and N,N-Dimethyltryptamine (DMT), as well as synthetic compounds such as lysergic acid diethylamide (LSD) [6]. Long before the first scientists' experimentations with mescaline, various cultures used the psychoactive properties of these drugs either to improve their physical performance in hunting (e.g. Cashinahua people in Brazil and Peru) or to gain spiritual guidance (e.g. Shipibo shamans) [7,8]: shamans typically drink the ayahuaska brew (which contains DMT) sat in a dark place and use songs and perfumes to shape their visions [8]. Interestingly enough, those Amazonian tribes recognized the capacity of psychedelics to enhance the talk between sensory modalities long before the discovery of LSD-induced synaesthesia [9].

All classic psychedelics are serotonergic agonists with a high affinity to $5HT_{2A}$ receptors [7,10]. Those receptors mediate most of the psychoactive effects of psychedelics, as demonstrated by the blocking ability of 2A antagonists, such as Ketanserin [11], even if other receptors, including $5HT_{2C}$, $5HT_{1A}$, $5HT_{5A}$ as well dopaminergic and beta adrenergic receptors, were also proposed to play a role in these effects [12–14]. $5HT_{2A}$ receptors are found in both the cortex and subcortical regions, but they are predominantly expressed in cortical layer V pyramidal cells, suggesting a cardinal involvement of deep layers in the phenomenology of psychedelics [10,15].

From a neurophysiological point of view, serotonergic drugs induce increased activation in a variety of cortical regions (including primary visual cortex and more frontal areas [16,17]) as well as profound changes in the functional connectivity in the Default-Mode Network and within/between Resting-State networks and Task-Positive networks [17]. Finally, psychedelics can decrease the power of alpha-band oscillations (although a reduction in a wider range of

frequencies has also been discovered; [17]), which has been interpreted as an increased excitability in the absence of external stimulation [18].

At the phenomenological level, psychedelics induce profound changes to the people who consume them [6]. They notably induce perceptual, emotional and cognitive alterations [19], while they can also generate mystical experiences and result in a diminished sense of self ("ego-dissolution") and a feeling of unboundedness [20,21]. Perceptual abnormalities comprise elementary and complex hallucinations (mostly visual or crossmodal), visual illusions, intensification of perceptual experience and mental imagery, together with synaesthesia, a (otherwise) rare perceptual phenomenon in which activation of one modality leads to subjective experiences in other modalities as well [22].

Interestingly, the content of these hallucinatory experience (e.g. "the spirits" in the case of the Amazonian Shipibo shamans) can be modulated by the activation of other sensory modalities (a phenomenon called "effect of setting"; e.g. singing of songs or spraying of perfumes) but also by the emotional state of the consumer prior the administration of the drug (named "effect of set") [8,23]. In summary, serotonergic hallucinogens generate rich experiences, including a dominant crossmodal component (complex hallucinations with synaesthesia), and also in some cases a top-down component with increased mental imagery and emotional effects [24–26]).

This description appears to be very different from the psychotic experiences observed in schizophrenia [27,28]. At the molecular level, schizophrenia has been linked to an increased presynaptic storage and release of striatal dopamine [29]. Glutamatergic [30,31], gabaergic [32] and serotonergic [28] abnormalities were also occasionally found associated to these dopaminergic dysregulations. At the phenomenological level, patients with schizophrenia mainly report hearing voices with a dominant negative affective content, although a minority of patients also describe multisensory (usually audio-visual) hallucinations [33–35]. In schizophrenia, these experiences are regularly found coupled with a reduced sensitivity to illusions [36].

These differences immediately raise new questions: What links exist between serotonergic agonism and the aberrant crossmodal experiences previously described? Is drug-induced psychosis functionally and mechanistically linked to schizophrenia-related psychosis? And if so, what mechanism(s) is(are) at the roots of this phenomenological variability? Are the different neurotransmitters (dopamine and serotonin) directly involved in such variability [28]? The recent third wave of psychedelic science together with the burgeoning field of

computational psychiatry [37] recently brought those questions to light and a number of insightful theories started to address them (Corlett *et al.*, 2009; Carhart-Harris, 2018; see Swanson, 2018 for a comprehensive review). Despite those efforts, a unifying, multiscale account of psychosis ranging from psychedelics to schizophrenia is still lacking.

In a first section of this paper, we will integrate available findings in a unique computational framework, able to capture the different facets of these psychotic experiences. We will notably defend the idea that *circular inferences* (CI), a form of suboptimal hierarchical probabilistic inference in which likelihood and prior corrupt and amplify each other [40,41], can offer a holistic and functional explanation for psychosis, beyond schizophrenia. Using in silico simulations, we will show how different suboptimal inferences may be linked to various forms of hallucinations. This will allow us to establish a link between observations made at the meso-scale (e.g., the "erroneous" message-passing between neurons involved in representations' building) and those made at the macro-scale (e.g., the behavioural and phenomenological manifestations of psychosis).

In a second section, we will review empirical evidence supporting a link between meso-scale and micro-scale findings, in other words between the different types of false inferences introduced and the modulation exerted by serotonin and dopamine. Our demonstration will build upon the critical role played by the balance between excitatory (E) and inhibitory (I) inputs in information processing within neural circuits. We state that one of the overarching goal of serotonin and dopamine systems is to regulate the neural E/I balance and to consequently control feedforward and feedback flows of information. More particularly, we will defend the idea that the amplification of feedback information (later called "descending inference loops") is controlled by serotonin (linked to the effects of psychedelics [7]), while the overcounting of feedforward information (later called "climbing loops") is controlled by dopamine (linked to schizophrenia [29]). This reformulation will allow us to implement the CI model as a canonical microcircuit able to integrate different scales of understanding hallucinations and hopefully pave the way for future biophysically detailed models of these phenomena.

## The circular inference framework

Because a detailed description of the *circular inference* (CI) framework has already been made available [40–42], we will only summarise the basic concepts needed to reformulate the problem of drug-induced psychosis.

The brain presents a highly recurrent architecture in which lateral/feedback connections actually dominate feed-forward inputs coming from sensory areas with a ratio of 9:1 [43]. These circuits naturally generate large levels of spontaneous neural activity [44], directly questioning how the system disentangles self-generated signals from true/new sensory events. This problem seems particularly acute for perceptual inference, in which sensory cues are integrated with prior expectations [40,45–48]. Such integration requires both feed-forward and feedback connections, incidentally creating internal information loops [49]. According to the *circular inference* (CI) framework, a finely tuned balance between neural excitation (E) and inhibition (I), a well-known property of brain circuits [50], could keep the information flow under surveillance, removing all redundant messages.

A dysregulation of the E/I balance (due to impaired inhibition, too much excitation or disruptions in the neuromodulation systems responsible for the tuning of the E/I balance [51–53]) results in the uncontrolled recruitment of these loops. In this case, sensory and prior information are reverberated in the neural circuit and eventually get aberrantly corrupted and over-counted [40,41]. A "descending loop" (quantified by parameter $a_P$; see **Supplementary Material**) is defined as the corruption of the feed-forward sensory information by the feedback (top-down) information, leading to an amplification of the priors. Conversely, a "climbing loop" ($a_S$) is generated when the sensory evidence corrupts the prior, leading to the amplification of the likelihood.

We previously showed that such circularity could be an important feature of perceptual inference in humans (Leptourgos et al, submitted) while in extreme cases, it generates psychotic symptoms, including hallucinations and delusions [54,55]. This idea is in line with related theories which postulate that schizophrenia may result from an impairment in brain's predictive mechanisms [56].

## Building a two-parallel-hierarchies' generative model

When we formalise brain function as hierarchical Bayesian inference, we assume that the brain learns the causal structure of the world [47,57]. This generative model is roughly reflected in the cortical hierarchy from primary sensory areas to association areas, with each level representing variables of increasing abstractness [40]. Inference on the other hand corresponds to the inversion of this model, which detects the most probable cause of the sensory evidence (in CI, inference is implemented as belief propagation).

Previous work on CI focused on simple generative models which consisted of one single hierarchy (e.g. the pairwise graph: Forest→Tree→Leaf→Colour green [40]). Those simplified models formalised the inferential processes in one single sensory modality, but failed to account for cross-modal phenomena, like the ones dominating the phenomenology of psychedelics. Here we extend the generative models previously used by considering 2 parallel hierarchies, each of them forming a pairwise graph reflecting a different sensory modality (e.g. audition and vision). Note that the presented results can be generalized to graphs with more than one parent [40]. The two modalities share a common node at the top and through this node stimuli in one modality affect inferences in the other. The nodes within each of the two hierarchies can be interpreted as different sensory areas (e.g. the ventral visual stream from V1 to V4) while the top node could coincide with an higher-order association cortex, where multisensory integration occurs (e.g. the superior temporal sulcus or the occipital-temporal junction [58–60]). For illustration purpose, we will consider the example of the audio-visual stimulus of a bird and a bird song. This complex stimulus results from the integration of a visual (bird) and an auditory (bird song) signals, each of them producing sensory evidence processed in each relevant modality. **Figure 1(a,b)** illustrates the generative model with corresponding anatomical interpretations.

For the sake of simplicity we will consider belief propagation (with or without circularity) in graphical models specific to binary variables, suitable for modelling decision making in "2-alternative forced choice tasks" (Jardri and Denève, 2013b; Leptourgos et al, submitted). Each node represents a variable and corresponds to a binary decision (i.e., "is the variable present or absent?"). All probabilistic quantities (likelihoods, priors and posteriors) are expressed as log-ratios. Sensory evidence is provided as a message clamped at the bottom of the hierarchy, with very positive/negative values corresponding to strong evidence that the bottom variable is present/absent and log-values close to zero corresponding to high uncertainty. Likewise, one can add priors to the system as messages clamped at the top of the hierarchy (e.g.

expectations, memories, emotional cues etc, reaching the association cortex from higher levels, such as the prefrontal cortex, not included explicitly in the model). The conditional probabilities that quantify the strength of the reciprocal causal links between connected nodes can be interpreted as weights (feed-forward ($w_S$) and feedback ($w_P$) weights). In our simulations they take values between 0.9 and 0.95. Finally, in all the presented simulations both modalities consisted of four layers (nodes). Note however that, since amplification depends on the number of connections, hierarchies with different number of nodes might be differentially affected by the loops (**Figure S3**) [61].

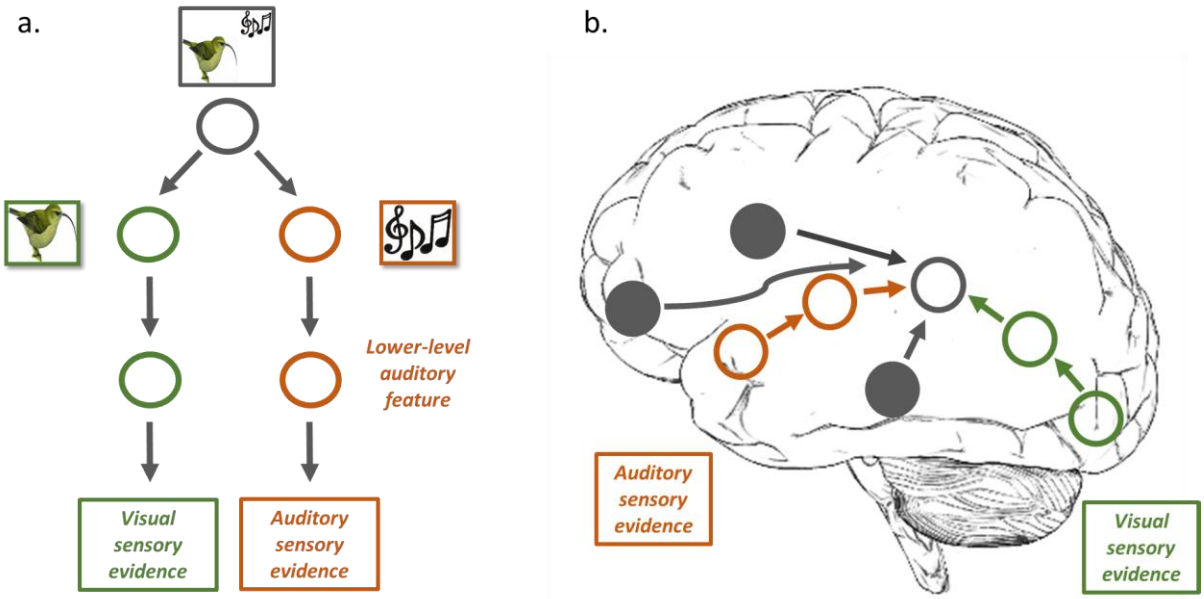


***Figure 1: Generative model and cortical representation: (a.):*** *Contrary to previous accounts of circular inference* [40,41] *(see also* ***Chapters 2,3****), here we consider a model with two hierarchies ,each representing a different sensory modality (e.g. audition (orange) and vision (green)). The two modalities are connected through the top node (grey), which corresponds to the locus of multisensory integration (e.g. association cortex). Our example illustrates how different stimulations of the two sensory modalities might have arisen from the same "multi-modal" stimulus. Inference corresponds to the inversion of this forward model.* ***(b.):*** *A potential implementation of the generative model in (**a.**) by the brain's hierarchical structure. According to Bayesian accounts of perception, the brain learns the causal structure of the world, which is represented in the cortical hierarchy. Filled nodes correspond to higher regions (OFC, ACC and hippocampus), potentially sending different kinds of feedback information to the sensory association cortex.*

## Synaesthesia, hallucinations and visual illusions

In order to establish a link between the meso-scale (i.e., probabilistic computations implemented by a message-passing algorithm) and the macro-scale (i.e., phenomenological varieties of the psychotic experience under psychedelics and in schizophrenia), we ran several *in silico* simulations for different scenarios. Those scenarios correspond to well-known effects of psychedelics and can be linked to aberrant inferences:

- In the "synaesthesia" scenario, we stimulated one modality (e.g., audition) with a strong sensory evidence ($L_A = 3$) while the other modality (e.g., vision) received weaker negative evidence ($L_V = -1.7$).
- In the "sensory-driven hallucination" scenario, both modalities were stimulated only by noise, but in one of the two modalities the value of the sensory evidence was slightly positive ($L_A = 0.4,\ L_V = -0.3$; in this context, noise corresponds to unreliable information). In both cases, to avoid additional confounding effects, we did not consider any prior ($L_P = 0$).
- Finally, in the "visual illusion" scenario, there was a contradiction between the sensory stimulation and the prior ($L_A = -1.4, L_V = -1.4, L_P = 1,$).

## Message passing with and without loops

Belief propagation works by iteratively calculating probabilistic messages and beliefs (log-posterior-ratios) (for technical details about belief propagation with and without loops, please refer to the **Supplementary Material** and to relevant books and papers [40–42,49]). In general (i.e., in the absence of loops), sensory information climbs the cortical hierarchy, moving from sensory to association areas, and conversely prior information descends the hierarchy (in the opposite direction). In the current model, two parallel hierarchies can talk to each other via the top node. In other words, because of the (potential) binding, the presence of a stimulus in one modality increases the probability that there is a stimulus in the other modality too [58]. Once the sensory information reaches the association cortex, it does not stop there but can enter the opposite hierarchy as a prior (**Figure 2a**). In summary and in the absence of loops, each sensory modality receives three types of information: (i) its own sensory evidence, (ii) the sensory evidence from the other modality (computed as a prior), and (iii) prior knowledge that reaches the association cortex from the top.

Adding loops to the model disrupts the conventional message-passing. Because of the loops, information is not only propagated in one direction but instead gets reverberated and counted multiple times. Examples of the message-passing schemas in the presence of descending and climbing loops are provided in **Figure 2(b,c)**.

Descending loops cause reverberation of the feedback information (priors and/or sensory information coming from the opposite modality) which re-climbs the hierarchy (**Figure 2b**; curved arrows). This redundant message re-enters the opposite sensory modality, in which it gets reverberated again due to the descending loops, and this vicious circle continues until beliefs reach their saturation point [41], making them almost indistinguishable.

In the case of climbing loops, sensory information is reverberated and re-descends the original hierarchy as feedback (**Figure 2c**; curved arrows). Contrary to what we just described for the descending loops, when the system is corrupted only by climbing loops, the reverberated message does not re-enter in the opposite modality, but remains trapped in the original one. As a result, we observe a unimodal amplification of sensory evidence, leaving the opposite modality practically unaffected.

As we will explain in the next sections, this difference in the locus of the amplification of information has major effects on the phenomenology of the resulting experiences.

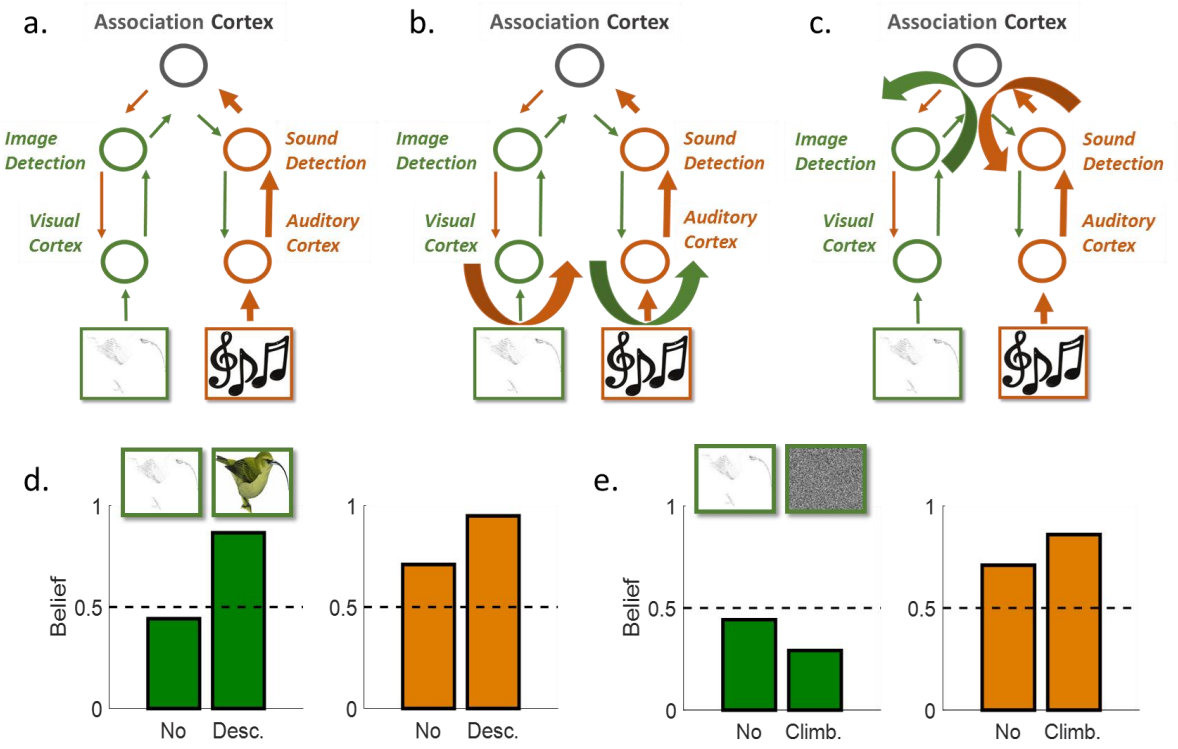

***Figure 2*: Circular inference and synaesthesia: (a.-c.):** *Belief propagation (without* **(a.)** *and with loops* **(b.,c.)**). *Information from the sensory organs climbs the hierarchy and enters in the opposite hierarchy, due to the multisensory integration occurring in the association cortex. In the presence of descending loops* **(b.)**, *information gets amplified in both modalities rendering the two modalities almost indistinguishable. Conversely, climbing loops* **(c.)** *force information to reverberate only inside the original modality, enhancing segregation between modalities. In synaesthesia, we consider what happens when one modality (e.g. audition) is strongly stimulated by unambiguous sensory evidence (e.g. birdsong;* $L_A = 3$*) while the other modality (e.g. vision) receives negative evidence (absence of bird;* $L_V = -1.7$*).* **(d.,e.)**: *Results of simulations for synaesthesia. In the absence of loops, the system hears the birdsong (belief above 0.5) but is more uncertain regarding the presence of a bird (belief close to 0.5). Because of the cross-amplification caused by descending loops (**d.**), the strong auditory information results in both beliefs increasing towards 1, eliciting an inversion in the case of vision. Thus, the system perceives the image of a bird, although only audition is stimulated by the birdsong (synaesthesia). On the contrary, climbing loops cannot generate such an inversion (or synaesthesia), because self-amplification inside the visual hierarchy reduces the visual belief towards 0.*

## Different types of loops for different clinical properties?

We explored the effects of loops in different scenarios, corresponding to different patterns of activation of the CI model. Since we were interested in how the two modalities interact to produce unimodal or crossmodal aberrant experiences, we focused on the beliefs generated by each modality (i.e., at the level of top nodes before higher-order association cortex, corresponding to the most abstract unimodal stimuli; e.g. objects and sounds/voices). We tested the models with different parameters' values (i.e., weights, strength of the loops, likelihoods and prior) which did not affect the qualitative impact of loops on the beliefs. Results were obtained after 50 iterations of the algorithm, with each iteration corresponding to one exchange of messages in both directions between all the connected nodes.

First, we explored the consequences of strongly activating one of the two modalities. This scenario is illustrated in **Figure 2(a-c)**, with relevant beliefs shown in **Figure 2(d,e)**. In our example, the system receives a strong auditory activation from a birdsong (orange hierarchy), but no corresponding visual stimulation (evidence supports the absence of a bird; green hierarchy). When the system does exact inference (belief propagation without loops; **Figure 2a**), those two pieces of sensory evidence climb their respective hierarchy, reach the association cortex (i.e., the grey node) and enter the opposite hierarchy where they are fed-back as priors (**Figure 2a**). Such a system predicts (and experiences) the presence of a birdsong ($P(birdsong|S_A, S_V) \gg 0.5$) and the absence of a bird ($P(bird|S_A, S_V) < 0.5$) (**Figure 2(d,e)**; left bars for each modality).

The addition of descending loops to the system results in amplification of both types of sensory information in each modality (**Figure 2b**). Because of this cross-amplification, the system as a whole is dominated by the strongest sensory input, which in our example is the positive auditory stimulation. Consequently, there is an artificial rise in both beliefs, resulting in an over-confidence regarding the presence of a birdsong and an inversion of the belief about the bird (**Figure 2d**; right bar for each modality). In short, the presence of descending loops enhances the communication between the different sensory modalities, which results in a concomitant experience in the second modality, a phenomenon that corresponds to synaesthesia [9].

On the contrary, climbing loops degrade the communication between modalities, as a result of the loopy amplification of information within the modality of origin. This makes synaesthetic experiences impossible. In our example the positive evidence for the birdsong is amplified within the auditory modality, whereas the negative evidence for the bird is amplified within the visual modality. Therefore, the system is over-confident that there is a birdsong but without the image of a bird (**Figure 2e**, right bar for each modality).

In the second and third scenarios, we wondered whether the added loops could generate strong beliefs even in the absence of strong sensory stimulations (hallucinations). We thus tested the case when both modalities are stimulated by noise, without the presence of priors (**Figure S1**, i.e., both sensory evidences are close to chance (zero)). In the absence of any convincing information, a system that does exact inference remains practically indecisive (beliefs close to 0.5). However, in agreement with previous results [40], loops generate strong beliefs, with unique patterns for each type of impairment. Descending loops generate a strong, crossmodal and sensory-driven experience (a multisensory hallucination combining a bird with a birdsong) whereas climbing loops result in segregated sensory modalities, carrying opposite results (clear presence of birdsong combined with clear absence of bird; sensory driven unimodal hallucination). Here it is important to highlight that climbing loops do not exclude multisensory experiences and could also generate a "crossmodal" hallucination. This particular case would need a concomitant stimulation of both modalities with noisy evidence whose value is (even slightly) above chance. Nevertheless, this does not correspond to a pure crossmodal phenomenon, since the two percepts are generated by unrelated causes.

Finally, for the sake of completeness, the third scenario probed the effect of a prior that contradicts the sensory evidence (**Figure S2**). This could correspond to illusory perception (i.e., a misperception caused by strong priors [62]), but also to the phenomenon of mental imagery

or even to what we could name a prior-driven hallucination (if sensory evidence is absent) [24,25]. We consider structurally identical sensory hierarchies, consequently the beliefs about the two stimuli are identical (but see also **Figure S3**). Importantly, descending loops amplify the prior, resulting in more illusions and stronger mental imagery (or prior-driven hallucinations). On the contrary, climbing loops force the system to resort more to its sensory evidence, which leads to more veridical percepts (decreased susceptibility to illusions) [40].

Previous research has shown that both climbing and descending loops can generate psychotic symptoms such as hallucinations and delusions [40,41]. However, the above considerations speak to a clear distinction between the phenomenological properties of the aberrant experiences generated by these two kinds of loops. Descending loops enhance communication between sensory modalities (potentially between cognitive modules as well), generating strong multisensory, sensory-driven or prior-driven experiences, such as simultaneous crossmodal hallucinations, synaesthesia, mental imagery and more visual illusions. On the other hand, climbing loops intensify segregation between the sensory hierarchies, while they also amplify sensory information, resulting in unimodal aberrant experiences and less vulnerability to illusions. Interestingly, the former appears closer to the phenomenological properties of the drug-induced psychosis (psychedelics) while the latter shares important properties with the phenomenology of schizophrenia (see also [54], for experimental evidence from schizophrenia patients).

Based on these simulations, we would like to suggest that psychedelics (and serotonergic agonism in general) generate transient descending loops in cortical circuits implementing belief propagation, while neurodevelopmental and genetic abnormalities could instigate more permanent climbing loops in patients with schizophrenia. Note however that in schizophrenia, different life trajectories could result in different impairments, potentially underlying the phenomenological variability observed among schizophrenia patients: e.g. prominent descending loops might also generate combined audio-visual hallucinations in a minority of patients [35] and where also found associated with more negative symptoms [54]. In the next sections, we will specifically investigate the links between the meso-scale and the micro-scale (neural circuits), suggesting detailed implementations for the different types of loops.

## From computations to implementations: loops are modulated by different neuromodulators

Since the seminal paper by Schultz et al [63], that linked dopamine with reward prediction errors, the functional role of the different neuromodulation systems has been vividly debated. The goal of this section is to put forward a novel conceptualisation of two neurotransmitters (serotonin and dopamine), as regulators of the E/I balance in different parts of the cortical circuits. An exhaustive review of the relevant literature on neuromodulation can be found elsewhere [64–66] and is beyond the scope of the paper.

In the previous section we suggested a unification of the different psychotic experiences based on the computational principle of *circular inference*. We argued that the hallucinogenic capacity of psychedelic drugs could be linked with their ability to generate descending loops in cortical circuits, while psychosis in schizophrenia (at least, in a majority of patients) could be related to an amplification of sensory evidence (climbing loops). This statement has a direct consequence: serotonin (the main neurotransmitter affected by psychedelics [12]) should be involved in the regulation / prevention of the descending loops whereas dopamine (linked to the pathophysiology of schizophrenia [29]) should be related to climbing loops.

Serotonin is synthesized primarily in the raphe nuclei [67]. Serotonergic receptors are expressed in various cortical and sub-cortical regions, including the claustrum and parts of the frontal, temporal, parietal and occipital lobes [10], with high concentration of the $5HT_{2A}$-type receptors found in layer V pyramidal cells and in middle layer interneurons [15]. Importantly, serotonin has been linked to a variety of processes [66] with a special focus on aversive learning (contrasting with dopamine; [68,69]) and mood regulation [70,71]. Although direct evidence about the involvement of serotonin in perceptual processing is sparser, such a connection is strongly implied by both the abundance of serotonergic receptors in sensory cortices [72] and the effects of the serotonergic drugs [73,74].

Crucially, Moreau et al showed that serotonin has laminar specific (and receptor-specific) effects on the E/I balance in the rat visual cortex [52], in agreement with neurophysiological considerations of the CI framework [41]. In addition to that, various studies have established a link between psychedelics and bistable perception [75,76], a phenomenon that has already been related to circularity (Leptourgos *et al.*, 2017; Leptourgos et al, submitted). In particular, psilocybin was shown to increase persistence in bistable perception tasks, a result consistent with the idea that serotonergic agonists could enhance descending loops. Finally,

optogenetic activation of 5HT neurons in the dorsal raphe nucleus promoted exploitative behaviour in mice, in line with prior-driven behaviours that can be induced by descending loops [77].

Similarly to serotonin, dopamine has also been related to various cognitive processes, mostly non-perceptual given the sparse dopaminergic projections to sensory cortices [78]. From very early, dopamine was associated with the representation of reward prediction errors [63,79,80]. More recently, different experimental studies also demonstrated a direct modulatory effect of dopamine on the visual cortex, comparable to attention [81] while theoretical considerations (mostly Bayesian in nature) started shifting away from reward-learning and towards probabilistic (perceptual) inference [66]. Such accounts include dopaminergic activation as a representation of uncertainty [82–84] or as sensory prediction error [85] and are closely related to the idea that dopamine controls the propagation of sensory information, that we suggest here. Besides, theories of schizophrenia conceptualizing the hyperdopaminergic tone as aberrant salience [86] or as over-precise prediction error [87–89], speak directly to the notion of climbing loops [40,42]. Finally, it's worth highlighting some results by Happel et al, suggesting that dopaminergic modulation via D1/D5 receptors regulates positive, sensory-dependent, thalamo-cortical feedback, offering some direct evidence for a possible neural substrate of long-range climbing loops in the auditory sensory hierarchy [90].

In this section, we briefly reviewed some evidence corroborating our suggestion that serotonin and dopamine could modulate the mechanisms controlling descending and climbing loops respectively. This suggestion has far-ranging consequences, from the functional role of neuromodulation to the potential development of new antipsychotic treatments (see **Discussion**). In the next section we present a reformulation of the *circular inference* algorithm, which anatomically distinguishes between the different types of loops. Based on that, we will put forward a detailed microcircuit, implementing circular inference.

## From computations to implementations: loops are mapped on different types of inhibition

We have highlighted that belief propagation (and also the CI framework) depends entirely on a precise correction of the propagated information, which prevents the formation of loops [40,49]. Previous formulations of the algorithm considered a correction at the level of the messages (eq. S2). Nonetheless, this formulation has two important drawbacks. First, it is not

clear how the system recognises which part of the information is redundant at each time-step. Importantly, the use of auxiliary nodes (a node per message, representing the rectified belief; eq. S2) is improbable, due to increased anatomical complexity. Second, it doesn't clearly differentiate between climbing and descending loops at the anatomical and neurophysiological levels, rendering the quest for neural substrates almost impossible.

Here, we overcome both obstacles by suggesting a novel formulation of the algorithm, based on the idea that the correction occurs at the level of the beliefs and not inside the messages (please refer to **Supplementary Material** for further details). For a pairwise graph, the resulting beliefs can be written as follows:

$$B_n = M_{(n-1)\to n} + M_{(n+1)\to n} - f(B_n, B_{n-1}, a_P) - g(B_n, B_{n+1}, a_S) \quad (1)$$

Messages are simply sigmoid functions of the beliefs of the sending nodes (without correction), while the last two terms correspond to the subtraction of the redundant prior ($f$ term) and sensory ($g$ term) information.

The neural interpretation of eq. 1 is straightforward. Belief at level $n$ (e.g. represented by pyramidal cells in a certain cortical/sensory area) is generated by integrating excitatory inputs from the levels above and below, which are balanced by inhibitory inputs from interneurons at the same level. Inhibition is driven by reciprocal excitation from the same level and inputs from the levels above and below (**Figure 3**). Interestingly, a single interneuron could receive input from multiple areas (each representing different variables), compatible with the observation that excitatory cells are 4 times more numerous than the inhibitory cells.

Crucially, these correction terms are not equivalent. The prior term depends on the belief of the node at the same level and that of the level below. Conversely, the sensory term depends on beliefs at the same level and the level above. That speaks to an important anatomical difference between the two inhibitory mechanisms: interneurons removing descending loops are driven by lateral and feed-forward connections; vice-versa, those responsible for climbing loops' control are driven by lateral and feedback connections (**Figure 3**).

Based on this argument and our knowledge about the intra-laminar connectivity of the sensory cortex, we would like to introduce a specific neural implementation of the CI framework.

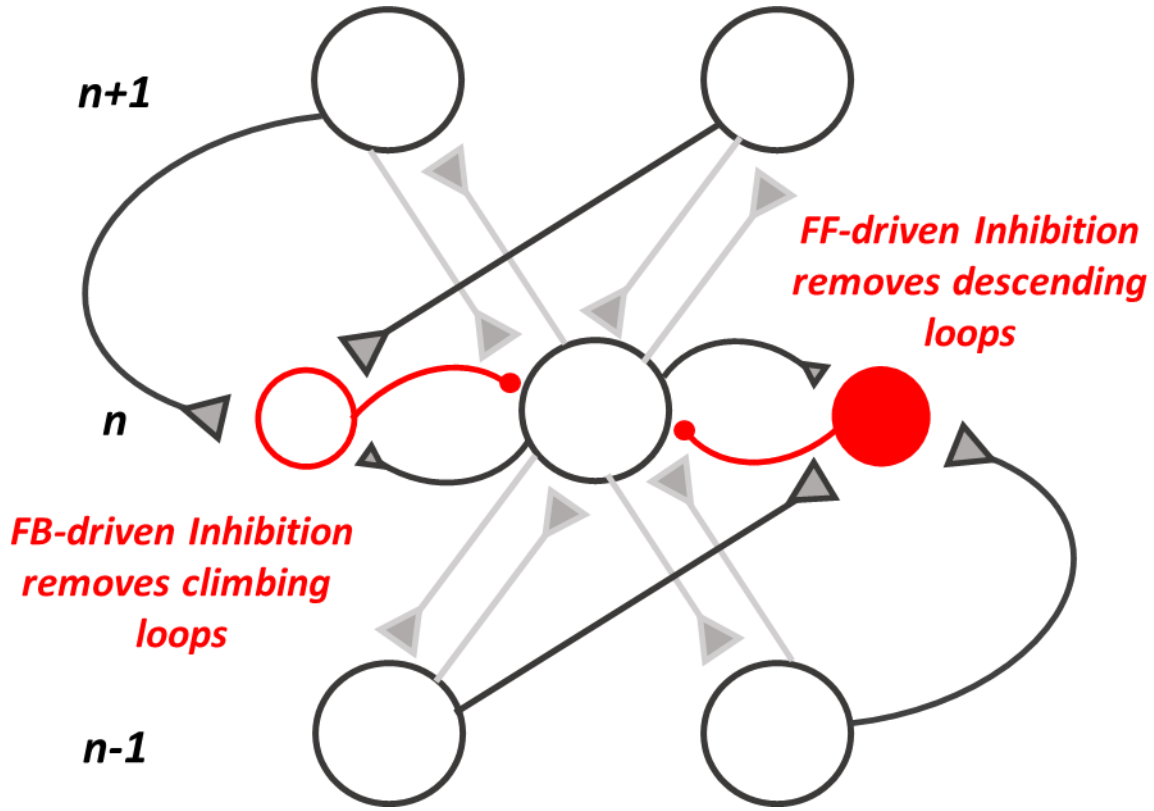


***Figure 3: Different types of inhibition are responsible for descending and climbing loops:*** *The corrections that remove redundant information can take place at the level of the beliefs, resulting in a reformulation of the circular inference algorithm. According to that, feedback-driven inhibition regulates climbing loops, while feed-forward-driven inhibition removes descending loops. This distinction allows to draw links with anatomical structures implementing the different correction mechanisms. [light grey: E-E connections; dark grey: E-I connections].*

## A canonical microcircuit implementing circular inference in the sensory cortex

The cortex is widely viewed as a hierarchical structure [91], whose networks are organized in a laminar specific way, leading to the notion of "canonical" microcircuits [92,93]. Those repeated circuits have long been viewed as the basis of many cortical computations [94,95]. What is the structure of those microcircuits and how are they linked to CI?

**Figure 4** illustrates the (simplified) canonical microcircuit implementing CI. We suggest that pyramidal cells in the superficial layers act as integrators [96], receiving all the available information and generating the beliefs. Both pyramidal cells and interneurons play important roles [97] and exhibit strong laminar specificity [93,98,99]. Feed-forward and feedback connections consist of pyramidal cells' axons [100], mostly targeting other pyramidal cells [101].

Following the dominant view, feed-forward information originates from L2/3 pyramidal cells (and thalamus) and mainly targets L4, projecting both on pyramidal cells and on interneurons in a non-selective way [102]. Those neurons then project on superficial layers (Markov *et al.*, 2014; but see also Pluta *et al.*, 2015) and from there information reaches deep layers and especially L5 [105]. Opposite inter-laminar connectivity within an area (e.g. dashed line from L5/6 to L4) is less frequent. Nevertheless, strong connections exist between L5/6 pyramidal cells and L4 interneurons [106].

In the opposite direction, feedback is less selective: it originates predominantly from the deep layers and targets all layers except L4 [103,107,108] but also non-specific thalamic nuclei [95]. Importantly, a lot of feedback connections terminate on interneurons in L2/3 [101] but also in L1 [109], which then form reciprocal connections with pyramidal cells in superficial layers [98,99].

This description of the cortical microcircuits illustrates in a dramatic way how the recurrent connectivity of the brain leads to the generation of information loops (**Figure 4(b,c)**). Top-down information re-climbs the hierarchy, trapped in descending loops (L2/3(V2) – L5/6(V2) – L2/3(V1) – L4(V2) – L2/3(V2)). Similarly, sensory information forms a positive feedback, involving cortical (L2/3(V2) – L4(V4) – L2/3(V4) – L5/6(V4) – L2/3(V2)) or thalamo-cortical (L2/3(V2) – L5/6(V2) – Thalamus – L4(V2) – L2/3(V2)) climbing loops [90].

More importantly, this illustration gives crucial hints about the implementation of the inhibitory mechanisms controlling the propagation of information. As described before, descending loops are balanced by inhibition driven by feed-forward excitatory inputs. This description fits nicely with L4 (and potentially deep layer) interneurons (**Figure 4b**). Hypo-activation of those interneurons (e.g. because of aberrant modulation of deep layers by serotonin, as observed with psychedelics) would lead to dis-inhibition of this part of the cortical circuits, resulting in an amplification of priors and to crossmodal aberrant experiences.

Likewise, climbing loops are balanced by feedback-driven inhibition. This description points to L1 interneurons [93], with the possible involvement of L2/3 interneurons as well. Impairments of inhibition in superficial layers (e.g. due to dopaminergic abnormalities in schizophrenia) would cause amplification of sensory information and thus more segregation of the sensory modalities. Note that this suggestion is compatible with the influential "dysconnectivity hypothesis" [110] and especially with a variation of this theory implicating thalamo-cortical loops [111].

## General discussion

The goal of this paper was twofold: (i) first, to propose a new multiscale theory of psychedelics, based on the *circular inference* framework, and second (ii) to address different open issues of the framework, related to phenomenology and neural implementation. Overall, we argue for a link between the macro-scale (phenomenological experience), the meso-scale (computational mechanisms) and the micro-scale (cortical microcircuits and neuromodulation), putting forward a unifying account of psychosis under the prism of circular belief propagation.

Previous work has linked climbing loops with psychotic symptoms in schizophrenia, including auditory hallucinations, persecutory delusions, jumping-to-conclusions bias and less vulnerability to illusions [36,40,54]. Additionally, we have suggested that mild (descending) loops might play an important role in normal brain function [54], underlying common perceptual phenomena such as bistable perception (Leptourgos *et al.*, 2017; Leptourgos et al, submitted). Here, we extended those accounts by showing that different loops can generate experiences with different phenomenological properties. In agreement with our previous results on schizophrenia, we demonstrated that climbing loops increase segregation between sensory modalities, generating sensory-driven unimodal hallucinations and less susceptibility to illusions and imagery. Conversely, our simulations suggested that descending loops lead to over-integrated sensory hierarchies which result in prior-driven or sensory-driven crossmodal hallucinations, synaesthesia, visual illusions and increased mental imagery, all common features in psychedelics-induced psychosis (other common properties such as the effect of set / effect of emotions on perception, could also be explained by amplified top-down effects [26]). We concluded that, while climbing loops might be a prominent impairment at the roots of schizophrenia symptoms, descending loops could underlie the rich phenomenology induced by serotonergic agonism.

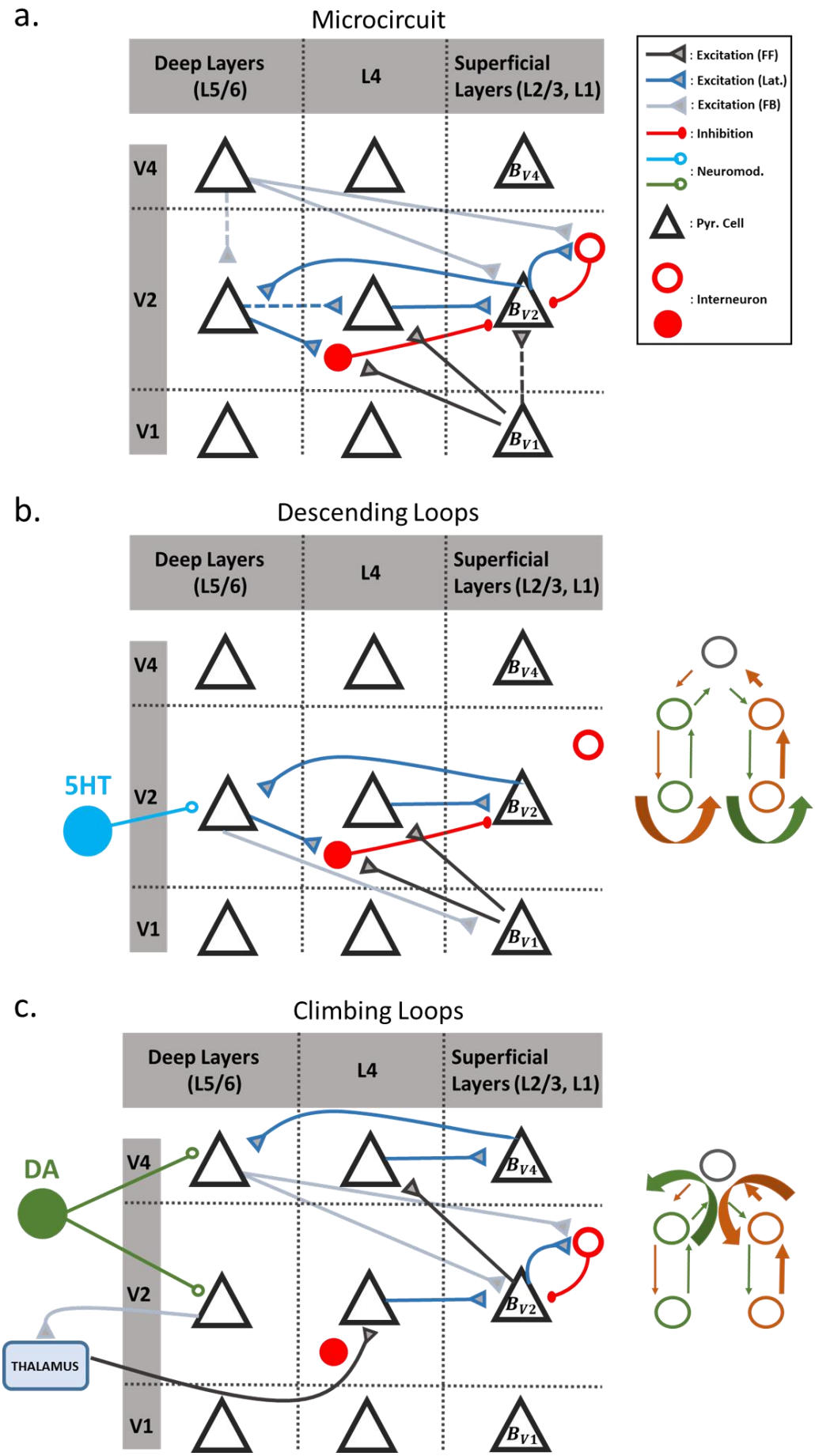
a.
Microcircuit
Deep Layers (L5/6)
L4
Superficial Layers (L2/3, L1)
V4
V2
V1
$B_{V4}$
$B_{V2}$
$B_{V1}$
: Excitation (FF)
: Excitation (Lat.)
: Excitation (FB)
: Inhibition
: Neuromod.
: Pyr. Cell
: Interneuron
b.
Descending Loops
5HT
c.
Climbing Loops
DA
THALAMUS

***Figure 4*: Canonical microcircuit implementing circular inference: (a.):** *Superficial layers act as integrators, receiving all information and generating the beliefs. Feedforward information originates from pyramidal cells in superficial layers (e.g. in V1) and targets pyramidal cells and interneurons in L4 (e.g. in V2), which then project on superficial layers. Less often, superficial layers target directly superficial layers of the level above (especially when the cortical regions are far apart). Feedback information originates from pyramidal cells in deep layers (e.g. in V4) and targets all layers except L4. Most often it targets pyramidal cells and interneurons in L2/3 but also in L1, which form reciprocal connections with each other. Within a cortical level, superficial layers project directly on deep layers, which then drive inhibition in L4.* ***(b.,c.):*** *This reciprocal connectivity between levels generates loops (L2/3(V2) – L4(V4) – L2/3(V4) – L5/6(V4) – [Thalamus – L4(V2)] – L2/3(V2): climbing loop, L2/3(V2) – L5/6(V2) – L2/3(V1) – L4(V2) – L2/3(V2): descending loop), which can be avoided if inhibition successfully removes all redundant information (balances excitation). Inhibition driven by bottom-up information, mediated by interneurons in L4 (potentially also interneurons in L5/6) removes descending loops whereas feedback-driven inhibition, mediated by interneurons in L2/3 (and/or L1) is responsible for climbing loops. We hypothesize that the former is regulated by serotonin while the latter is mainly regulated by dopamine. Serotonergic agonists such as LSD dys-regulate the control of information, mainly through the $5HT_{2A}$ receptors (predominantly expressed in pyramidal cells in L5), leading to the generation of descending loops. In the case of schizophrenia, alterations in the dopaminergic pathways (mesolimbic and/or meso-cortical) could affect information processing in cortical microcircuits (with the possible implication of long range connections with thalamus, striatum or PFC), resulting in climbing loops.*

This theoretical distinction between schizophrenia and psychedelics has important implications for the neural substrates of these inference loops. The involvement of serotonin (and more particularly of the $5HT_{2A}$ receptors) in the psychotic effects of psychedelics has been undisputed for almost 20 years [11]. Similarly, successive dopaminergic theories of schizophrenia have been widely accepted, in view of the efficacy of typical antipsychotics (mostly $D_2R$-antagonsists) and the psychotomimetic effects of amphetamines (DA-agonists) [29]. Given the suggested link between psychedelics/schizophrenia and descending/climbing loops, we further postulated that serotonin modulates the former, while dopamine modulates the latter, both by regulating the E/I balance in different parts of the cortical microcircuits, responsible for the integration of sensory evidence and priors. Remained the question of the structure of those microcircuits.

In the third and final part of the paper, we addressed this issue by delineating a canonical microcircuit which implements CI, further linking computation with neural implementation (see also Bastos *et al.*, 2012, for a related microcircuit implementing predictive coding). Inspired by a novel formulation of the CI equations, in which different loops are regulated by different types of inhibition, we argued that feedback-driven inhibitory interneurons situated in superficial layers (L2/3 and/or L1) mediate climbing loop-control. Conversely, inhibition in the deeper layers (L4 and/or L5/6), driven by feed-forward information, is mainly responsible for

descending loop-regulation. In combination with the previous arguments, we deduced that impairments in the climbing loop control, potentially underlined by dopaminergic hyper-function, mediate the segregated pattern that we observe in schizophrenia. On the other hand, impairments in the descending loop regulation, triggered by serotonergic agonists, result in the integrated pattern that we see in psychedelics.

Interestingly, this unifying framework is related to a number of different theories that addressed the problem of psychosis [6], many of which built on the idea that hallucinatory phenomena result from impairments in predictive mechanisms of the brain [19,38,112,113]. In one study, Muthukumaraswamy et al suggested that enhanced priors, mediated by over-activation of deep layers, generate subjective effects of psychedelics [19], while in another study, Corlett and colleagues associated the same effects with impaired bottom-up processing (combined with intact top-down processing), mediated by enhanced AMPA signalling [38].

The present account is also compatible with another influential contemporary theory of psychedelics, the entropic brain theory (EBT; [39,114]). EBT suggests that psychedelics increase the entropy of brain activity, rendering it more chaotic and susceptible to intrinsic and extrinsic influences, while they also increase connectivity between unrelated brain networks, in agreement with the enhanced integration induced by descending loops [16,17]. Note however that CI, contrary to EBT, is a functional theory directly derived from normative principles.

It's important to highlight that the CI framework is (among other things) a theory of psychedelics, and as such it interprets drug-induced synaesthesia, while it remains unclear whether it can also interpret developmental synaesthesia (e.g. grapheme-color synaesthesia, experienced by a small number of people without any drug consumption). Consequently, the link between our theory and theories of developmental synaesthesia remains debatable at best [22]. That said, it is difficult to disregard the similarity between the descending loops (amplification of priors) and ideas such as the "disinhibited prior" [115].

Importantly, the CI framework introduced in this paper makes a number of new testable predictions. First, it offers a tentative explanation to the phenomenological and neurobiological variability observed in schizophrenia. Indeed, although the majority of patients experience auditory hallucinations, a sub-group of patients (around 30%) also experience both auditory and visual hallucinations [34,116–119]. Additionally, although most of the patients respond well to typical antipsychotic medication (DA-antagonists), one on four exhibits refractory hallucinations [120]. Crucially, most of these "treatment-resistant" hallucinations still respond well to Clozapine, an atypical antipsychotic characterized by a high affinity for serotonin

[28,120]. Although evidence for a link between these two groups (i.e., patients exhibiting complex multisensory hallucinations, which are also drug-refractory) is for now very sparse [4], it's tempting to suggest that most schizophrenia patients exhibit a dopaminergic dysregulation which generates predominant climbing loops in the cortical hierarchy that could be corrected by first-line antipsychotics (type A schizophrenia; [121]). A minority of patients, with primary or associated serotonergic impairments and predominant descending loops, would express more crossmodal and treatment-resistant hallucinations, which better respond to agents with serotonergic antagonism properties, such as Clozapine or even Ketanserin (a good candidate for type B schizophrenia; [121]). A recently published cases-report supports this claim [122].

The aforementioned prediction calls for two important methodological comments. First, phenomenology is crucial when studying schizophrenia, e.g. for building computational assays [123]. Different abnormalities might underlie different groups of patients with differences in symptomatology [54] or phenomenology. If not taken into account, this variability could at minima contaminate the results, leading to contradicting evidence (e.g. prior-driven vs sensory-driven symptoms). Second, despite the importance of multisensory hallucinations as a potential diagnostic tool, few studies have studied them systematically [33,35]. As a result, it's difficult to evaluate objectively their prevalence (as opposed to serial hallucinations), both in schizophrenia and under psychedelics [4].

Another important prediction of the model comes from the fact that descending loops cause amplification of information in both modalities. As presented in **Figure 2**, this results in a general over-confidence, affecting both modalities. Interestingly, this is a unique prediction, since different models (e.g. those based on increased prior weights; not presented here) would only generate over-confidence in the non-stimulated modality and an under-confidence in the stimulated one, a different prediction that can be easily tested behaviourally in future works.

Finally, one more testable prediction concerns the laminar and input specificity of inhibition related to schizophrenia and to drug-induced psychosis. Although standard imaging techniques do not possess the necessary spatial resolution to test so precise assumptions, recent advances in high-field laminar fMRI should allow for arbitration between different implementations or theories [124].

We need to acknowledge some limitations to this work. The present model was designed to account for differences in the number of recruited modalities during hallucinations, still it does not directly address the question why hallucinations are mostly auditory in schizophrenia but visual under psychedelics. We note that such a difference might be less related with the

mechanisms and more related with structural constraints of each modality, including the number of cortical connections (**Figure S3**) [61], modality-dependent distinctions in cortical microcircuits or dissimilar expression of the implicated receptors.

Additionally, we defined in this work crossmodal hallucinations as aberrant experiences occurring simultaneously in more than one modalities and having a common cause. Future work will have to distinguish between crossmodal hallucinations with additional fusion (e.g. seeing a bird and hearing a birdsong, as with anti-muscarinic drugs) and crossmodal hallucinations in which binding is less obvious, as with psychedelics (e.g. hearing a voice and seeing the content of the voice). In this vein, we suggest that the incorporation of aberrant learning in the psychedelics' CI model, caused by descending loops, would strengthen non-existent associations, leading to bizarre crossmodal combinations [40].

Finally, a cautious approach is needed regarding the potential neural substrates of CI. First, our reconceptualization of dopamine's function does not attribute separate roles to D1 and D2 receptors, given the lack of evidence regarding their differential contributions to the dopaminergic effects on sensory cortices [29]). In addition to that, the suggested microcircuit is necessarily simplified, ignoring less frequent connections, interneuron specificities (e.g. differences between fast-spiking interneurons and adaptive interneurons [99]) and within layer details (e.g. detailed connectivity within L2/3), potentially underlying complementary functions such as amplification for sustained activity or filtering.

Overall, we put forward a unifying, trans-nosographic and multiscale account of psychosis, with a special focus on psychedelics. We feel that future work could complete this effort, by adding even more psychotic experiences (e.g. hallucinations occurring in Parkinson's disease or experiences related to anti-muscarinic drugs) and considering alternative underlying mechanisms (e.g. the aberrant weighting of priors or sensory evidence in the context of circular inference or predictive coding).

## Supplementary Material 1: Additional simulations (scenarios 2,3 / Different number of nodes)

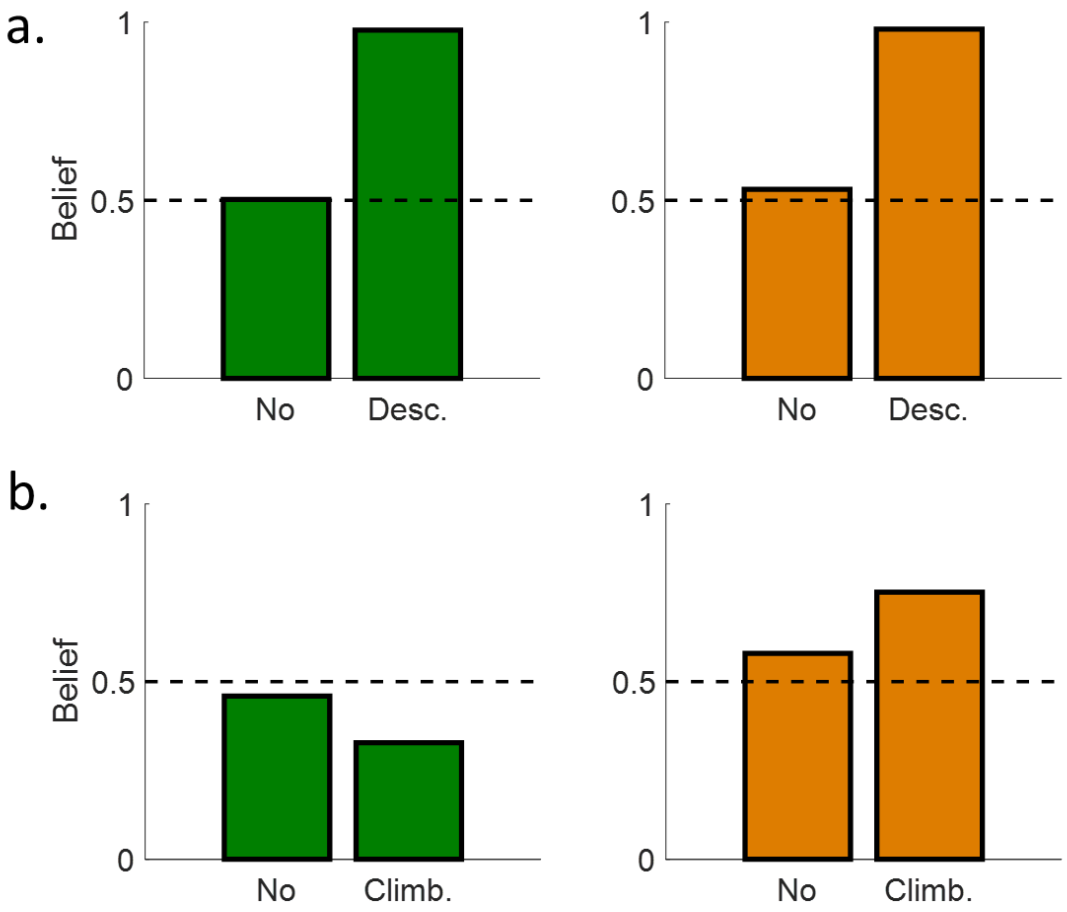


***Figure S1: Circular Inference and hallucinations:*** *The scenario is similar to* ***Figure 2*** *(synaesthesia), except that both modalities are stimulated by very unreliable evidence (noise;* $L_A = 0.4; L_V = -0.3$*). As in synaesthesia, descending loops* ***(a.)*** *cause simultaneous activation of the two modalities (because of the cross-amplification), eliciting a simultaneous cross-modal hallucination. On the other hand, climbing loops* ***(b.)*** *have different effects in the two modalities, resulting in a unimodal, auditory hallucination. We note that in* ***(b.)****, we present the belief at the bottom node, in which the effect of the climbing loops is more easily observable (for the top node, the effect is qualitatively the same but quantitatively weaker).*

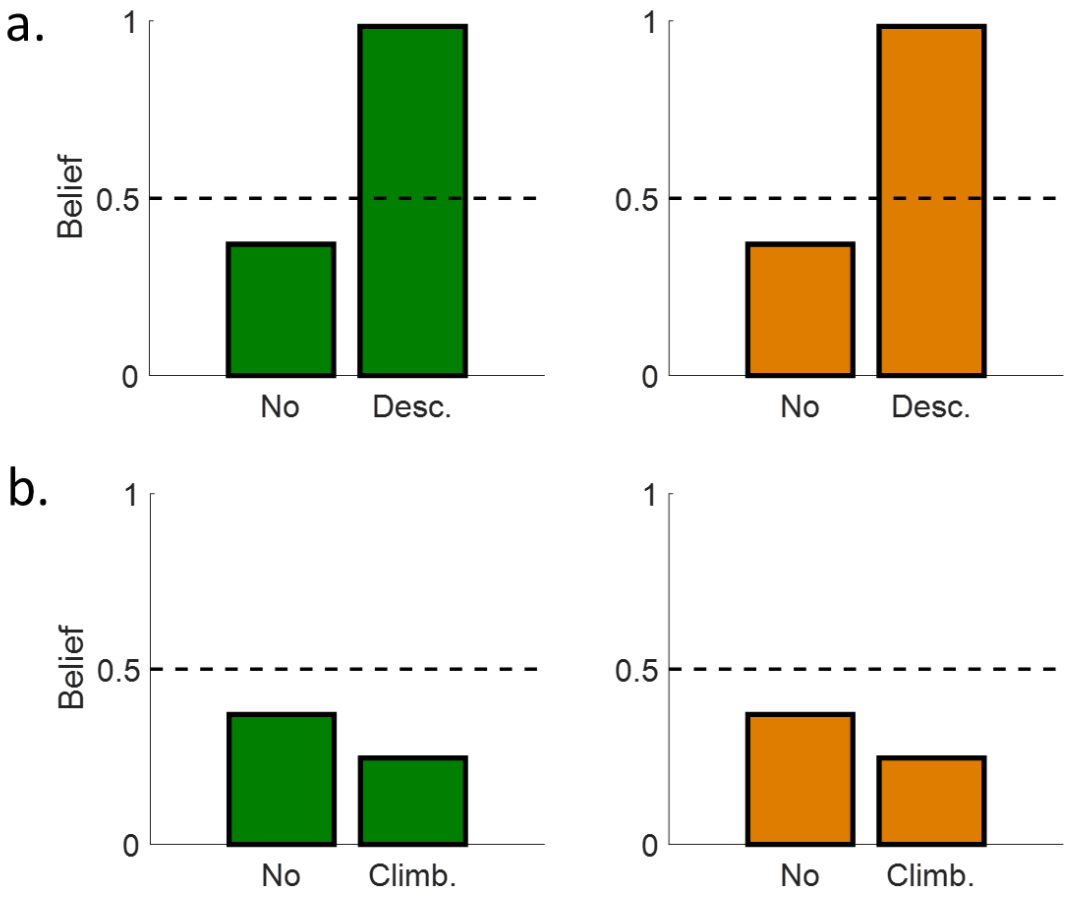


***Figure S2: Circular inference and illusions (mental imagery or prior driven hallucinations):*** *Because the (contradicting) prior is weaker than the sensory information ($L_A = L_V = -1.4$; $L_P = 1$), in the absence of loops both beliefs are below 0.5 (both the bird and the birdsong are absent; note that the beliefs are equal in the two modalities because the two hierarchies are identical and they receive equally strong stimulation).* ***(a.)*** *When the prior is amplified (descending loops), inference is dominated by the feedback, resulting in beliefs close to 1 (that could correspond to an illusion, enhanced mental imagery or a prior-driven hallucination, depending on the context).* ***(b.)*** *Climbing loops (amplification of sensory information) have the opposite effect (less vulnerability to illusions, weaker mental imagery and no prior-driven hallucinations)* [40].

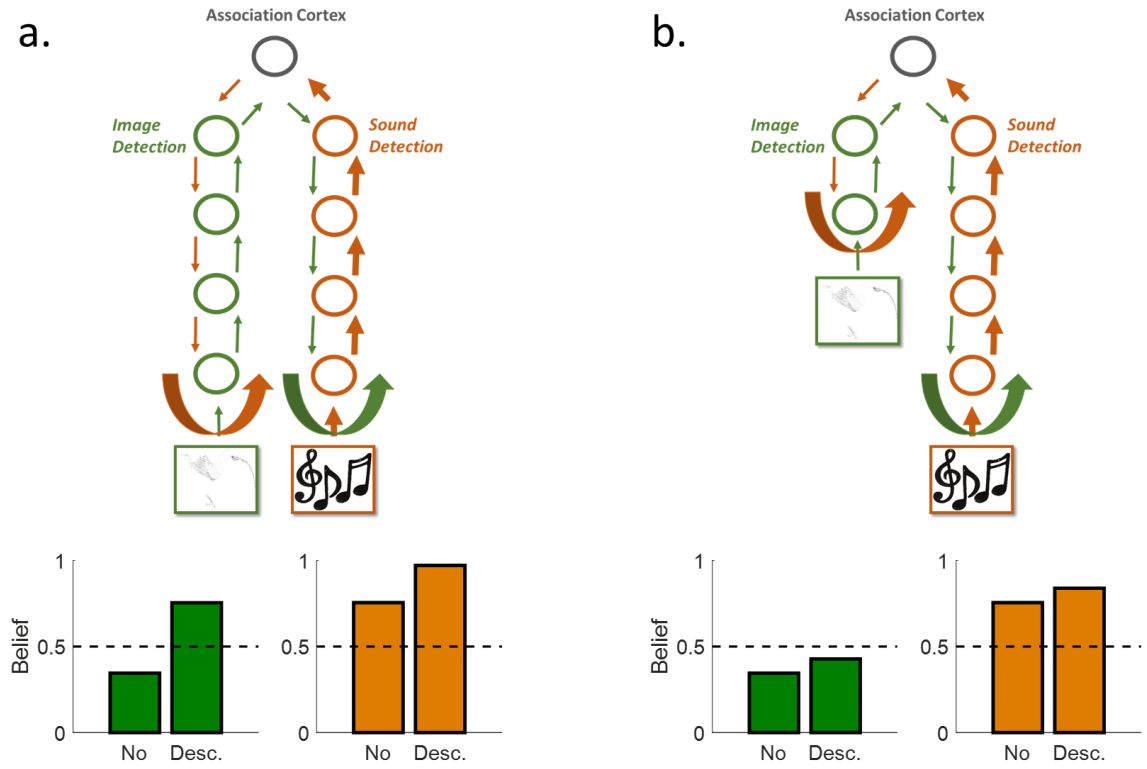


***Figure S3: The number of loops affects the amount of amplification.*** ***(a.):*** *Descending loops in a model with four levels / variables per hierarchy result in synaesthesia.* ***(b.):*** *Exactly the same simulation, but this time the visual hierarchy has only 2 levels / variables, resulting in reduced amplification (in the visual hierarchy) due to descending loops. Consequently, synaesthesia does not occur. Similar results can be obtained for the other simulations presented in the* ***Main Text****. [La = 2.1; Lv = -1.7].*

## Supplementary Material 2: The reformulated CI algorithm

In previous works [40,41], we suggested that the circular inference algorithm can be written in the following form:

$$B_n = \sum_{k=\{n+1,n-1\}} M_{k\to n} \quad (S1)$$

$$M_{k\to n} = F(B_k - M_{n\to k}, w) \quad (S2)$$

where, $F(B,w) = log\left(\frac{we^B+(1-w)}{(1-w)e^B+w}\right)$ is a sigmoid function.

Those two equations calculate iteratively the posterior probabilities for each variable ($S1$) and the probabilistic messages exchanged by nodes ($S2$), in a way that draws an analogy with neural processing in recurrent hierarchical networks. Importantly, eq. ($S2$) tells us that messages are a function of the belief of the node that sends the message, corrected by the message sent in the opposite direction. This correction is crucial, because it controls the propagation of information, making sure that no message gets counted more than once (due to loops).

Despite its efficiency, this formulation has two main drawbacks. First, such a correction is difficult to be implemented in cortical circuits, because it assumes the additional calculation, for each node, of the k ($B_k - M_{n\to k}$) terms. Given the complexity of real-life generative models, such a solution seems extremely inefficient (e.g. in terms of metabolic cost), but also it makes the system vulnerable to small perturbations (failure in any of the nodes would lead to a cascade of miscalculations, resulting in completely aberrant inferences). In addition to that, although it has been postulated that different mechanisms control the different types of potential loops [40], this formulation gives no information about their potential anatomical differences.

To account for those drawbacks, we suggest here a novel formulation of equations ($S1, S2$), where we assume that corrections occur at the level of the beliefs ($S1$). The details of such an algorithm are beyond the scope of this paper and they will be presented in future work. In general, we suggest that equations ($S1, S2$) can be rewritten as follows:

$$B_n = M_{(n-1)\to n} + M_{(n+1)\to n} - f(B_n, B_{n-1}, a_P) - g(B_n, B_{n+1}, a_S) \quad (S3)$$

$$M_{k\to n} = F(B_k, w) \quad (S4)$$

Those equations also describe the iterative calculation of posteriors and messages, but now corrections appear as separate (non-linear) terms ($f$ and $g$) in ($S3$). Consequently,

reciprocal excitation generates loops, but redundant information is removed by inhibitory interneurons targeting directly the neurons that calculate beliefs (pyramidal cells in L2/3, according to our microcircuit). More particularly, inhibition learns to track excitation (E/I balance; [125]), while neuromodulation might be driving this learning. Importantly, inhibition tracking excitation from the two streams (feed-forward and feedback) has different properties: interneurons that remove descending loops are driven by lateral and feed-forward excitation (indeed descending loops are generated between nodes $n$ and $(n-1)$), while interneurons removing climbing loops are driven by lateral and feedback connections (from nodes $n$ and $(n+1)$) (note that a simple subtraction of the opposite message is not possible, given the inequality $F(B-M) \neq F(B) - F(M)$).

## References


1. Larøi F, Sommer IE, Blom JD, Fernyhough C, Ffytche DH, Hugdahl K, et al. The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: State-of-the-art overview and future directions. Schizophr Bull. 2012;38: 724–733. doi:10.1093/schbul/sbs061

2. Waters F, Allen P, Aleman A, Fernyhough C, Woodward TS, Badcock JC, et al. Auditory hallucinations in schizophrenia and nonschizophrenia populations: A review and integrated model of cognitive mechanisms. Schizophr Bull. 2012;38: 683–692. doi:10.1093/schbul/sbs045

3. Niemantsverdriet MBA, Slotema CW, Blom JD, Franken IH, Hoek HW, Sommer IEC, et al. Hallucinations in borderline personality disorder: Prevalence, characteristics and associations with comorbid symptoms and disorders. Sci Rep. Springer US; 2017;7: 1–8. doi:10.1038/s41598-017-13108-6

4. Waters F, Collerton D, Ffytche DH, Jardri R, Pins D, Dudley R, et al. Visual hallucinations in the psychosis spectrum and comparative information from neurodegenerative disorders and eye disease. Schizophr Bull. 2014;40: 233–245. doi:10.1093/schbul/sbu036

5. Osmond H. a Review of the Clinical Effects of Psychotomimetic Agents. Ann N Y Acad Sci. 1957;66: 418–434. doi:10.1111/j.1749-6632.1957.tb40738.x

6. Swanson LR. Unifying theories of psychedelic drug effects. Front Pharmacol. 2018;9. doi:10.3389/fphar.2018.00172

7. Nichols DE. Psychedelics. Pharmacol Rev. 2016;68: 264–355.

8. Fortier M. Sense of reality, metacognition and culture in schizophrenic and drug-induced hallucinations. In: Proust J, Fortier M, editors. Metacognitive diversity: An interdisciplinary approach. Oxford / New York: Oxford University Press; 2018. pp. 343–379.

9. Luke DP, Terhune DB. The induction of synaesthesia with chemical agents: a systematic review. Front Psychol. 2013;4: 1–12. doi:10.3389/fpsyg.2013.00753

10. Nichols DE. Hallucinogens. Pharmacol Ther. 2004;101: 131–181. doi:10.1016/j.pharmthera.2003.11.002

11. Vollenweider FX, Vollenweider-Scherpenhuyzen MFI, Bäbler A, Vogel H, Hell D. Psilocybin induces schizophrenia-like psychosis in humans via a serotonin-2 agonist action. Neuroreport. 1998;9: 3897–3902. doi:10.1097/00001756-199812010-00024

12. Vollenweider FX, Kometer M. The neurobiology of psychedelic drugs: Implications for the treatment of mood disorders. Nat Rev Neurosci. Nature Publishing Group; 2010;11: 642–651. doi:10.1038/nrn2884

13. Kozlenkov A, Gonzalez-Maeso J. Animal models and hallucinogenic drugs. In: Jardri R, Cachia A, Thomas P, Pins D, editors. The Neuroscience of Hallucinations. New York, NY: Springer New York; 2013. pp. 253–277.

14. Lowe NG, Rapagnani MP, Mattei C, Stahl SM. The psychopharmacology of hallucinations: Ironic insights into mechanisms of action. In: Jardri R, Cachia A, Thomas P, Pins D, editors. The Neuroscience of Hallucinations. New York, NY: Springer New York; 2013. pp. 471–492.

15. Weber ET, Andrade R. Htr2a gene and 5-HT2A receptor expression in the cerebral cortex studied using genetically modified mice. Front Neurosci. 2010;4: 1–12. doi:10.3389/fnins.2010.00036

16. De Araujo DB, Ribeiro S, Cecchi GA, Carvalho FM, Sanchez TA, Pinto JP, et al. Seeing with the eyes shut: Neural basis of enhanced imagery following ayahuasca ingestion. Hum Brain Mapp. 2012;33: 2550–2560. doi:10.1002/hbm.21381

17. Carhart-Harris RL, Muthukumaraswamy S, Roseman L, Kaelen M, Droog W, Murphy K, et al. Neural correlates of the LSD experience revealed by multimodal neuroimaging. Proc Natl Acad Sci. 2016;113: 4853–4858. doi:10.1073/pnas.1518377113

18. Kometer M, Schmidt A, Jancke L, Vollenweider FX. Activation of Serotonin 2A Receptors Underlies the Psilocybin-Induced Effects on Oscillations, N170 Visual-Evoked Potentials, and Visual Hallucinations. J Neurosci. 2013;33: 10544–10551. doi:10.1523/JNEUROSCI.3007-12.2013

19. Muthukumaraswamy SD, Carhart-Harris RL, Moran RJ, Brookes MJ, Williams TM, Errtizoe D, et al. Broadband Cortical Desynchronization Underlies the Human Psychedelic State. J Neurosci. 2013;33: 15171–15183. doi:10.1523/JNEUROSCI.2063-13.2013

20. Griffiths RR, Richards WA, McCann U, Jesse R. Psilocybin can occasion mystical-type experiences having substantial and sustained personal meaning and spiritual significance. Psychopharmacology (Berl). 2006;187: 268–283. doi:10.1007/s00213-006-0457-5

21. Halberstadt AL. Recent advances in the neuropsychopharmacology of serotonergic hallucinogens. Behav Brain Res. Elsevier B.V.; 2015;277: 99–120. doi:10.1016/j.bbr.2014.07.016

22. Hubbard EM, Ramachandran VS. Neurocognitive mechanisms of synesthesia. Neuron. 2005;48: 509–520. doi:10.1016/j.neuron.2005.10.012

23. Carhart-Harris RL, Roseman L, Haijen E, Erritzoe D, Watts R, Branchi I, et al. Psychedelics and the essential importance of context. J Psychopharmacol. 2018; doi:10.1177/0269881118754710

24. Albright TD. On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception. Neuron. Elsevier Inc.; 2012;74: 227–245. doi:10.1016/j.neuron.2012.04.001

25. Powers AR, Kelley M, Corlett PR. Hallucinations as Top-Down Effects on Perception. Biol Psychiatry Cogn Neurosci Neuroimaging. Elsevier; 2016;1: 393–400. doi:10.1016/j.bpsc.2016.04.003

26. O'Callaghan C, Kveraga K, Shine JM, Adams RB, Bar M. Predictions penetrate perception: Converging insights from brain, behaviour and disorder. Conscious Cogn. Elsevier Inc.; 2017;47: 63–74. doi:10.1016/j.concog.2016.05.003

27. Andreasen NC, Flaum M. Schizophrenia: The Characteristic Symptoms. Schizophr Bull. 1991;17: 27–49. doi:10.1093/schbul/17.1.27

28. González-Maeso J, Sealfon SC. Psychedelics and schizophrenia. Trends Neurosci. 2009;32: 225–232. doi:10.1016/j.tins.2008.12.005

29. Howes O, McCutcheon R, Stone J. Glutamate and dopamine in schizophrenia: An update for the 21st century. J Psychopharmacol. 2015; doi:10.1177/0269881114563634

30. Corlett PR, Honey GD, Krystal JH, Fletcher PC. Glutamatergic model psychoses: Prediction error, learning, and inference. Neuropsychopharmacology. Nature Publishing Group; 2011;36: 294–315. doi:10.1038/npp.2010.163

31. Anticevic A, Gancsos M, Murray JD, Repovsf G, Driesena NR, Ennisg DJ, et al. NMDA receptor function in large-scale anticorrelated neural systems with implications for cognition and schizophrenia. Proc Natl Acad Sci. 2012;109: 16720–16725. doi:10.1073/pnas.1208494109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1208494109

32. Lewis DA, Pierri JN, Volk DW, Melchitzky DS, Woo TUW. Altered GABA neurotransmission and prefrontal cortical dysfunction in schizophrenia. Biol Psychiatry. 1999;46: 616–626. doi:10.1016/S0006-3223(99)00061-X

33. Lim A, Hoek HW, Deen ML, Dirk J, GROUP I, Bruggeman R, et al. Prevalence and classification of hallucinations in multiple sensory modalities in schizophrenia spectrum disorders. Schizophr Res. The Authors; 2016;176: 493–499. doi:10.1016/j.schres.2016.06.010

34. Llorca PM, Pereira B, Jardri R, Chereau-Boudet I, Brousse G, Misdrahi D, et al. Hallucinations in schizophrenia and Parkinson's disease: An analysis of sensory modalities involved and the repercussion on patients. Sci Rep. Nature Publishing Group; 2016;6: 1–9. doi:10.1038/srep38152

35. Dudley R, Aynsworth C, Cheetham R, McCarthy-Jones S, Collerton D. Prevalence and characteristics of multi-modal hallucinations in people with psychosis who experience visual hallucinations. Psychiatry Res. 2018;269: 25–30.

36. Notredame C-E, Pins D, Denève S, Jardri R. What visual illusions teach us about schizophrenia. Front Integr Neurosci. 2014;8: 1–16. doi:10.3389/fnint.2014.00063

37. Huys QJM, Maia T V., Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016;19: 404–413. doi:10.1038/nn.4238

38. Corlett PR, Frith CD, Fletcher PC. From drugs to deprivation: a Bayesian framework for understanding models of psychosis. Psychopharmacology (Berl). 2009;206: 515–30. doi:10.1007/s00213-009-1561-0

39. Carhart-Harris RL. The entropic brain - revisited. Neuropharmacology. Elsevier Ltd; 2018; doi:10.1016/j.neuropharm.2018.03.010

40. Jardri R, Denève S. Circular inferences in schizophrenia. Brain. 2013;136: 3227–41. doi:10.1093/brain/awt257

41. Leptourgos P, Denève S, Jardri R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Curr Opin Neurobiol. 2017;46: 154–161. doi:10.1016/j.conb.2017.08.012

42. Deneve S, Jardri R. Circular inference: Mistaken belief, misplaced trust. Curr Opin Behav Sci. 2016;11: 40–48. doi:10.1016/j.cobeha.2016.04.001

43. Douglas RJ, Koch C, Mahowald M, Martin KAC, Suarez HH. Recurrent excitation in neocortical circuits. Science (80- ). 1995;269: 981–985. doi:10.1126/science.7638624

44. Hupé J, James A, Payne B, Lomber S, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature. 1998;394: 784–787. Available: https://search.proquest.com/openview/09bbe4a11f10a4097276702l9c4b017b/1?pq-

origsite=gscholar&cbl=40569

45. Lee TS, Mumford D. Hierarchical bayesian inference in the visual cortex. J Opt Soc Am A. 2003;20: 1434–1448.

46. Friston K. Hierarchical Models in the Brain. PLoS Comput Biol. 2008;4. doi:10.1371/journal.pcbi.1000211

47. Lochmann T, Deneve S. Neural processing as causal inference. Current Opinion in Neurobiology. 2011. pp. 774–781. doi:10.1016/j.conb.2011.05.018

48. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, et al. Uncertainty in perception and the Hierarchical Gaussian Filter. Front Hum Neurosci. 2014;8: 1–24. doi:10.3389/fnhum.2014.00825

49. Bishop C. Pattern Recognition and Machine Learning. Springer; 2006.

50. Denève S, Machens CK. Efficient codes and balanced networks. Nat Neurosci. 2016;19: 375–382. doi:10.1038/nn.4243

51. Lucas-Meunier E, Monier C, Amar M, Baux G, Frégnac Y, Fossier P. Involvement of nicotinic and muscarinic receptors in the endogenous cholinergic modulation of the balance between excitation and inhibition in the young rat visual cortex. Cereb Cortex. 2009;19: 2411–2427. doi:10.1093/cercor/bhn258

52. Moreau WA, Amar M, Le Roux N, Morel N, Fossier P. Serotoninergic fine-tuning of the excitation-inhibition balance in rat visual cortical networks. Cereb Cortex. 2010;20: 456–467. doi:10.1093/cercor/bhp114

53. Pfeffer T, Avramiea AE, Nolte G, Engel AK, Linkenkaer-Hansen K, Donner TH. Catecholamines alter the intrinsic variability of cortical population activity and perception. PLoS Biology. 2018. doi:10.1371/journal.pbio.2003453

54. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference in schizophrenia. Nat Commun. 2017;8: 14218. doi:10.1038/ncomms14218

55. Jardri R, Hugdahl K, Hughes M, Brunelin J, Waters F, Alderson-Day B, et al. Are Hallucinations Due to an Imbalance Between Excitatory and Inhibitory Influences on the Brain? Schizophr Bull. 2016;42: 1124–1134. doi:10.1093/schbul/sbw075

56. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The Predictive Coding Account of Psychosis. Biol Psychiatry. Elsevier Inc; 2018; 1–10. doi:10.1016/j.biopsych.2018.05.015

57. Friston KJ, Parr T, de Vries B. The graphical brain: belief propagation and active inference. Netw Neurosci. 2017; 1–78. doi:10.1162/NETN_a_00018

58. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. PLoS One. 2007;2. doi:10.1371/journal.pone.0000943

59. Jardri R, Pins D, Bubrovszky M, Lucas B, Lethuc V, Delmaire C, et al. Neural functional organization of hallucinations in schizophrenia: multisensory dissolution of pathological emergence in consciousness. Conscious Cogn. Elsevier Inc.; 2009;18: 449–57. doi:10.1016/j.concog.2008.12.009

60. Jardri R, Thomas P, Delmaire C, Delion P, Pins D. The neurodynamic organization of modality-dependent hallucinations. Cereb Cortex. 2013;23: 1108–1117. doi:10.1093/cercor/bhs082

61. Jardri R, Denève S. Computational Models of Hallucinations. In: Jardri R, Cachia A, Thomas P, Pins D, editors. The Neuroscience of Hallucinations. New York, NY: Springer New York; 2013. pp. 289–313. doi:10.1007/978-1-4614-4121-2

62. Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nat Neurosci. 2002;5: 598–604. doi:10.1038/nn858

63. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. [Review] [37 refs]. Science (80- ). 1997;275: 1593–1599.

64. Doya K. Modulators of decision making. Nat Neurosci. 2008;11: 410–416. doi:10.1038/nn2077

65. Dayan P. Twenty-Five Lessons from Computational Neuromodulation. Neuron. Elsevier Inc.; 2012;76: 240–256. doi:10.1016/j.neuron.2012.09.027

66. Iglesias S, Tomiello S, Schneebeli M, Stephan KE. Models of neuromodulation for computational psychiatry. Wiley Interdiscip Rev Cogn Sci. 2017;8: 1–22. doi:10.1002/wcs.1420

67. Mohammad-Zadeh LF., Moses L., Gwaltney-Brant SM. Serotonin : a review. 2008; 187–199. doi:10.1111/j.1365-2885.2008.00944.x.REVIEW

68. Cools R, Nakamura K, Daw ND. Serotonin and dopamine: Unifying affective, activational, and decision functions. Neuropsychopharmacology. Nature Publishing Group; 2011;36: 98–113. doi:10.1038/npp.2010.121

69. Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R. Serotonin Selectively Modulates Reward Value in Human Decision-Making. J Neurosci. 2012;32: 5833–5842. doi:10.1523/JNEUROSCI.0053-12.2012

70. Dayan P, Huys QJM. Serotonin, inhibition, and negative mood. PLoS Comput Biol. 2008;4. doi:10.1371/journal.pcbi.0040004

71. Dayan P, Huys QJM. Serotonin in Affective Control. Annu Rev Neurosci. 2009;32: 95–126. doi:10.1146/annurev.neuro.051508.135607

72. Dori I, Dinopoulos A, Blue ME, Parnavelas JG. Regional differences in the ontogeny of the serotonergic projection to the cerebral cortex. Exp Neurol. 1996;138: 1–14. doi:10.1006/exnr.1996.0041

73. Celada P, Puig MV, Artigas F. Serotonin modulation of cortical neurons and networks. Front Integr Neurosci. 2013;7: 1–20. doi:10.3389/fnint.2013.00025

74. Juckel G. Serotonin: From sensory processing to schizophrenia using an electrophysiological method. Behav Brain Res. Elsevier B.V.; 2015;277: 121–124. doi:10.1016/j.bbr.2014.05.042

75. Carter OL, Pettigrew JD, Hasler F, Wallis GM, Liu GB, Hell D, et al. Modulating the rate and rhythmicity of perceptual rivalry alternations with the mixed 5-HT2Aand 5-HT1Aagonist psilocybin. Neuropsychopharmacology. 2005;30: 1154–1162. doi:10.1038/sj.npp.1300621

76. Carter OL, Hasler F, Pettigrew JD, Wallis GM, Liu GB, Vollenweider FX. Psilocybin links binocular rivalry switch rate to attention and subjective arousal levels in humans. Psychopharmacology (Berl). 2007;195: 415–424. doi:10.1007/s00213-007-0930-9

77. Lottem E, Banerjee D, Vertechi P, Sarra D, Lohuis MO, Mainen ZF. Activation of

serotonin neurons promotes active persistence in a probabilistic foraging task. Nat Commun. Springer US; 2018;9: 1–12. doi:10.1038/s41467-018-03438-y

78. Björklund A, Dunnett SB. Dopamine neuron systems in the brain: an update. Trends Neurosci. 2007;30: 194–202. doi:10.1016/j.tins.2007.03.006

79. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996;16: 1936–1947. doi:10.1.1.156.635

80. Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005;47: 129–141. doi:10.1016/j.neuron.2005.05.020

81. Noudoost B, Moore T. Control of visual cortical signals by prefrontal dopamine. Nature. Nature Publishing Group; 2011;474: 372–375. doi:10.1038/nature09995

82. de Lafuente V, Romo R. Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. Proc Natl Acad Sci. 2011;108: 19767–19771. doi:10.1073/pnas.1117636108

83. Friston KJ, Shiner T, FitzGerald T, Galea JM, Adams R, Brown H, et al. Dopamine, affordance and active inference. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002327

84. Schwartenbeck P, FitzGerald THB, Mathys C, Dolan R, Friston K. The dopaminergic midbrain encodes the expected certainty about desired outcomes. Cereb Cortex. 2015;25: 3434–3445. doi:10.1093/cercor/bhu159

85. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, denOuden HEM, et al. Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. Neuron. Elsevier Inc.; 2013;80: 519–530. doi:10.1016/j.neuron.2013.09.009

86. Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. Am J Psychiatry. 2003;160: 13–23. Available: http://ajp.psychiatryonline.org/doi/abs/10.1176/appi.ajp.160.1.13

87. Corlett PR, Murray GK, Honey GD, Aitken MRF, Shanks DR, Robbins TW, et al. Disrupted prediction-error signal in psychosis: Evidence for an associative account of delusions. Brain. 2007;130: 2387–2400. doi:10.1093/brain/awm173

88. Murray GK, Corlett PR, Clark L, Pessiglione M, Blackwell a D, Honey G, et al. Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. Mol Psychiatry. 2008;13: 239, 267–76. doi:10.1038/sj.mp.4002058

89. Corlett PR, Taylor JR, Wang XJ, Fletcher PC, Krystal JH. Toward a neurobiology of delusions. Prog Neurobiol. Elsevier Ltd; 2010;92: 345–369. doi:10.1016/j.pneurobio.2010.06.007

90. Happel MFK, Deliano M, Handschuh J, Ohl FW. Dopamine-Modulated Recurrent Corticoefferent Feedback in Primary Sensory Cortex Promotes Detection of Behaviorally Relevant Stimuli. J Neurosci. 2014;34: 1234–1247. doi:10.1523/JNEUROSCI.1990-13.2014

91. Felleman DJ, Van Essen DC. Distributed hierachical processing in the primate cerebral cortex. Cereb Cortex. 1991;1: 1–47. doi:10.1093/cercor/1.1.1

92. Douglas RJ, Martin KAC. Neuronal Circuits of the Neocortex. Annu Rev Neurosci. 2004;27: 419–451. doi:10.1146/annurev.neuro.27.070203.144152

93. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical

Microcircuits for Predictive Coding. Neuron. Elsevier Inc.; 2012;76: 695–711. doi:10.1016/j.neuron.2012.10.038

94. Raizada RDS, Grossberg S. Towards a Theory of the Laminar Architecture of Cerebral Cortex: Computational Clues from the Visual System. Cereb Cortex. 2003;13: 100–113. doi:10.1093/cercor/13.1.100

95. Haeusler S, Maass W. A statistical analysis of information-processing properties of lamina-specific cortical microcircuit models. Cereb Cortex. 2007;17: 149–162. doi:10.1093/cercor/bhj132

96. Meyer G. Forms and spatial arrangement of neurons in the primary motor cortex of man. J Comp Neurol. 1987;262: 402–428. doi:10.1002/cne.902620306

97. Isaacson JS, Scanziani M. How inhibition shapes cortical activity. Neuron. 2011;72: 231–243. doi:10.1016/j.neuron.2011.09.027.How

98. Dantzker JL, Callaway EM. Laminar sources of synaptic input to cortical inhibitory interneurons and pyramidal neurons. Nat Neurosci. 2000;3: 701–707. doi:10.1038/76656

99. Yoshimura Y, Callaway EM. Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. Nat Neurosci. 2005;8: 1552–1559. doi:10.1038/nn1565

100. Johnson RR, Burkhalter AA. A polysynaptic feedback circuit in rat visual cortex. J Neurosci. 1997;17: 7129–7140. doi:10.1002/(SICI)1096-9861(19960506)368:3<383::AID-CNE5>3.0.CO;2-1

101. Gonchar Y, Burkhalter A. Distinct GABAergic targets of feedforward and feedback connections between lower and higher areas of rat visual cortex. J Neurosci. 2003;23: 10904–10912. doi:23/34/10904 [pii]

102. Johnson RR, Burkhalter A. Microcircuitry of forward and feedback connections within rat visual cortex. J Comp Neurol. 1996;368: 383–398.

103. Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, et al. Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. J Comp Neurol. 2014;522: 225–259. doi:10.1002/cne.23458

104. Pluta S, Naka A, Veit J, Telian G, Yao L, Hakim R, et al. A direct translaminar inhibitory circuit tunes cortical output. Nat Neurosci. Nature Publishing Group; 2015;18: 1631–1640. doi:10.1038/nn.4123

105. Thomson AM, West DC, Wang Y, Bannister AP. Synaptic Connections and Small Circuits Involving Excitatory and Inhibitory Neurons in Layers 2-5 of Adult Rat and Cat Neocortex: Triple Intracellular Recordings and Biocytin Labelling In Vitro. Cereb Cortex. 2002;12: 936–953. doi:10.1093/cercor/12.9.936

106. Mejias JF, Murray JD, Kennedy H, Wang XJ. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. Sci Adv. 2016;2. doi:10.1126/sciadv.1601335

107. Muckli L, Martino F De, Vizioli L, Petro L, Smith FW, Ugurbil K, et al. Contextual Feedback to Superficial Layers of V1 Report. Curr Biol. The Authors; 2015;25: 2690–2695. doi:10.1016/j.cub.2015.08.057

108. Kok P, Bains LJ, van Mourik T, Norris DG, de Lange FP. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. Curr Biol. Elsevier Ltd; 2016;26: 1–6. doi:10.1016/j.cub.2015.12.038

109. Larkum ME. The yin and yang of cortical layer 1. Nat Neurosci. Nature Publishing Group; 2013;16: 114–115. doi:10.1038/nn.3317

110. Stephan KE, Friston KJ, Frith CD. Dysconnection in Schizophrenia : From Abnormal Synaptic Plasticity to Failures of Self-monitoring. Schizophr Bull. 2009;35: 509–527. doi:10.1093/schbul/sbn176

111. Murray JD, Anticevic A. Toward understanding thalamocortical dysfunction in schizophrenia through computational models of neural circuit dynamics. Schizophr Res. Elsevier B.V.; 2016; doi:10.1016/j.schres.2016.10.021

112. Friston KJ. Hallucinations and perceptual inference. Behav Brain Sci. 2005;28: 764–766. doi:10.1017/S0140525X05290131

113. Carhart-Harris RL, Friston KJ. The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas. Brain. 2010;133: 1265–1283. doi:10.1093/brain/awq010

114. Carhart-Harris RL, Leech R, Hellyer PJ, Shanahan M, Feilding A, Tagliazucchi E, et al. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. Front Hum Neurosci. 2014;8: 1–22. doi:10.3389/fnhum.2016.00423

115. Neufeld J, Sinke C, Zedler M, Dillo W, Emrich HM, Bleich S, et al. Disinhibited feedback as a cause of synesthesia : Evidence from a functional connectivity study on auditory-visual synesthetes. Neuropsychologia. Elsevier Ltd; 2012;50: 1471–1477. doi:10.1016/j.neuropsychologia.2012.02.032

116. David C, Greenstein D, Clasen L, Gochman P, Miller R, Tossell J, et al. Childhood Onset Schizophrenia: High rate of visual hallucinations. J Am Acad Child Adolesc Psychiatry. 2011;50: 681–686. doi:10.1016/j.jaac.2011.03.020.Childhood

117. Amad A, Cachia A, Gorwood P, Pins D, Delmaire C, Rolland B, et al. The multimodal connectivity of the hippocampal complex in auditory and visual hallucinations. Mol Psychiatry. Nature Publishing Group; 2014;19: 184–191. doi:10.1038/mp.2012.181

118. Cachia A, Amad A, Brunelin J, Krebs M, Plaze M, Thomas P, et al. Deviations in cortex sulcation associated with visual hallucinations in schizophrenia. Mol Psychiatry. Nature Publishing Group; 2014; 1–7. doi:10.1038/mp.2014.140

119. Rolland B, Amad A, Poulet E, Bordet R, Vignaud A, Bation R, et al. Resting-State Functional Connectivity of the Nucleus Accumbens in Auditory and Visual Hallucinations in Schizophrenia. Schizophr Bull. 2015;41: 291–299. doi:10.1093/schbul/sbu097

120. Sommer IEC, Slotema CW, Daskalakis ZJ, Derks EM, Blom JD, Gaag M Van Der. The Treatment of Hallucinations in Schizophrenia Spectrum Disorders. Schizophr Bull. 2012;38: 704–714. doi:10.1093/schbul/sbs034

121. Howes OD, Kapur S. A neurobiological hypothesis for the classification of schizophrenia : type A ( hyperdopaminergic ) and type B ( normodopaminergic ). Br J Psychiatry. 2014;205: 1–3. doi:10.1192/bjp.bp.113.138578

122. Sommer IEC, Kleijer H, Visser L, Laar T Van. Successful treatment of intractable visual hallucinations with 5-HT2A antagonist ketanserin. BMJ Case Rep. 2018; doi:10.1136/bcr-2018-224340

123. Stephan KE, Iglesias S, Heinzle J, Diaconescu AO. Translational Perspectives for

Computational Neuroimaging. Neuron. Elsevier Inc.; 2015;87: 716–732. doi:10.1016/j.neuron.2015.07.008

124. Stephan KE, Petzschner FH, Kasper L, Bayer J, Wellstein K V., Stefanics G, et al. Laminar fMRI and computational theories of brain function. Neuroimage. Elsevier Ltd; 2017; doi:10.1016/j.neuroimage.2017.11.001

125. Boerlin M, Machens CK, Denève S. Predictive Coding of Dynamical Variables in Balanced Spiking Networks. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1003258

# Chapter 8

# General Discussion

The goal of this thesis was to study the computational mechanisms underlying perception under ambiguity in the general population, but also the reality distortion (i.e., aberrant percepts and beliefs) which characterises the psychosis spectrum. We hypothesized that *circular inference*, a suboptimal predictive process, is present in both healthy individuals and psychotic patients (or individuals under the influence of psychedelic drugs), albeit in different amounts, affecting the way they perceive the world. The present findings largely support this claim.

## Circular inference in bistable perception in the general population

Bistability occurs under conditions of high ambiguity [38]. Interestingly, previous research has shown that the effect of circularity becomes more noticeable when there is high uncertainty [35]. We hypothesized that if loops exist in the cortical hierarchy of healthy individuals, then we should be able to detect their trace in bistable perception tasks.

In a sequence of three experiments (**Chapters 2** and **4**), we found different types of indication that loops are indeed a fundamental mechanism of our perceptual system. On one hand, we observed that the combination of low-level visual cues with high-level priors follows a *circular inference* rule. This result was substantiated by an interaction between the two effects, which cannot occur under purely Bayesian assumptions, but emerges naturally in the presence of loops [39]. On the other hand, we discovered that the dynamics of the behaviour of healthy participants in an intermittent presentation task (the stimulus is displayed discontinuously on the screen) is not compatible with a pure integrator (Bayesian system), instead it displays the features of a bistable attractor, as expected of a system with descending loops. The question that naturally arises then is the following: How can circular inference explain the existence and the rich phenomenology of bistable perception?

Historically, most theoretical approaches to bistable perception consisted of mechanistic models, focusing either on the biophysical mechanisms or the dynamics [40,41]. We suggested a novel functional model (**Chapter 3**) which can answer epistemological questions, including the *alternation* problem ("Why switches occur?") and the *selection* problem ("Why we perceive one interpretation at a time?") [42]. In addition to that, the *dynamical circular inference* model[2] assigns a fundamental role to descending loops: they turn the leaky integrator into a bistable attractor with two highly trusted states, hence they explain the

[2] it can be distinguished from previous circular inference models, thanks to the presence of dynamics

*confidence* problem ("Why do we perceive clear interpretations in the absence of reliable data?"). Furthermore, this model can explain various qualitative features of bistability, including the Levelt's laws [43] and the stabilisation occurring under discontinuous presentation of the stimulus [44] (see also **Chapter 4**).

Plenty of studies in the past found that the brain functions as an ideal, Bayesian observer ([45–48] to mention a few). The present results suggest that, although the brain is equipped with the necessary machinery for Bayesian inference, there are inherent limitations in its use [49–52]. It must be highlighted though that it is not clear yet whether this suboptimality is due to inherent limitations in the neural mechanisms or whether it subserves a specificfunction (see **Chapter 3** for more details).

## Circular inference in the psychosis spectrum

Circular inference was originally proposed as a model for the positive symptoms of schizophrenia. Using in-silico simulations, Jardri and Denève showed that both climbing loops and descending loops are able to generate aberrant percepts (i.e., hallucinations), false and persistent beliefs (i.e., delusions) but also a state of increased confidence [35]. Based on behavioural arguments (decreased vulnerability to visual illusions [37], combined with a "Jumping To Conclusions" bias [8] in schizophrenia patients, especially those with prominent delusions), they suggested that an amplification of sensory evidence, rather than an amplification of priors, explains better those symptoms. A subsequent probabilistic reasoning task confirmed this hypothesis, also revealing a complex association between the different clusters of symptoms and the different types of loops [36].

In **Chapter 6** of the present thesis, we questioned the fact that an increased amount of climbing loops could be necessary and/or sufficient to explain schizophrenia-related psychosis. Beyond circularity, we wondered if additional mechanisms are at play and especially what is the role of the dynamics, which had been largely ignored by previous studies. We compared patients with prominent positive symptoms with matched healthy controls in two bistable perception tasks, similar to the ones used in healthy participants in the first part of the thesis. Our results support the idea that climbing loops are enhanced in patients. Importantly, climbing loops alone cannot explain all the observations. In particular we observed an increased destabilization of schizophrenia patients for short blank durations in an intermittent presentation task, a finding that goes against the predictions of the model. We suggest that a 2-factor theory explains

all the findings, namely an increase in the sensory-amplification (climbing loops) combined with an overestimation of the environmental volatility. Note that those two impairments might have a hidden link, for example the overestimation of the volatility (increased transition rates) might be caused by aberrant learning, a demonstrated effect of the loops [35].

An intriguing characteristic of psychosis is its phenomenological variability. Psychosis is a prominent clinical dimension in many psychiatric and neurological disorders (schizophrenia, affective psychosis, borderline personality disorder, Parkinson's disease, Charles-Bonnet syndrome, Alzheimer's disease etc.; [53–57]) but can occur in non-clinical populations too [58]. Additionally, it can be triggered by various psychotomimetic agents, such as serotonergic agonists / psychedelics [59]. Despite sharing some common features (they all include a detachment from reality), each psychotic experience has unique phenomenological properties. In **Chapter 7** we argue that different phenomenologies might be underpinned by different mechanisms, in particular different types of loops. We suggest that the crossmodal, mainly prior-driven experiences, common under psychedelics, are triggered by descending loops. On the contrary, we show that climbing loops enhance segregation between sensory modalities, resulting in unimodal, sensory-driven hallucinations, as reported in schizophrenia (and in agreement with our previous findings). Furthermore, based on this phenomenological distinction and to fill the gap between erroneous message-passing and neuromodulation, we propose a canonical micro-circuit implementing circular belief-propagation in which the two inference loops could be controlled by different neuromodulators acting on the E/I balance: the descending loops, under serotonergic control (the primary target of psychedelics), and the climbing loops regulated by dopamine (whose role in the pathophysiology of schizophrenia is widely accepted).

Overall, our results suggest that circular inference is a common feature of the human brain. It explains a variety of observations, ranging from normal to pathological brain functioning and spanning the psychosis spectrum.

## Limitations and future directions

This work makes a number of novel predictions, while it also suffers from several limitations. Both predictions and limitations have been highlighted in the different chapters of the thesis. Here we summarise some of them, which we think could offer new opportunities for scientific inquiry.

- In **Chapter 2**, due to the experimental paradigm and more specifically to the manipulation of the prior (between-group design), we only fitted the models to the average data but not to individuals. It would be interesting to repeat this experiment, using a within-group design (e.g., as in [60]), allowing for model-fitting at the level of individuals. Such a study has been designed and will constitute the work of future students in the team;
- In Bayesian accounts, inference and learning make up the two sides of the same coin [6,31,35]. In our functional model (dynamical circular inference; **Chapter 3**), we only considered inference, but we completely ignored learning. Importantly, the addition of learning could generate secondary effects, for example loops could affect the feedforward weights and the transition rates (see [35] for more details);
- In **Chapter 4**, in the tilted cube experiment, we didn't find an acute bias for the longest blank duration (1050 ms), which indicates the absence of descending loops. Nevertheless, persistence exhibits an increasing trend for long OFF-Durations, implying that such a bias could potentially become visible for longer intervals. A follow-up experiment, testing intervals larger than 1050 ms (e.g. between 1 and 2 s), could resolve this doubt.
- In **Chapter 6**, we only presented preliminary results from a small number of participants. Our findings (especially the negative results and the "Bias Against Disconfirmatory Evidence") will have to be tested in a larger sample (30 participants per group, per experiment). Crucially, a participant per participant fitting procedure along with Bayesian model comparison will also have to confirm our qualitative conclusions. These complementary analyses have been scheduled when we will have reached the final sample-size;
- In **Chapter 7** we made a number of predictions regarding the computational mechanisms of drug-induced and schizophrenia-related psychosis and their neural correlates. Those predictions could be tested experimentally, using both behaviour (e.g. the "Fisher task"; [36]) and imaging (e.g. laminar fMRI; [61]).
- More generally, all the current work on *circular inference* has focused on binary variables. Extending the framework by considering Gaussian variables (continuous variables in general) could give the opportunity to compare it directly with different theories, such as predictive coding [13,42,62] and the Hierarchical Gaussian Filter [63].

## References


1. Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: W. H. Freeman and Company; 1982.

2. Barlow HB. Pattern recognition and the repsonses of sensory neurons. Ann N Y Acad Sci. 1969;156: 872–881.

3. Barlow HB. Single units and sensation: a neuron doctrine for perceptual psychology? Perception. 1972;1: 371–394.

4. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. J Physiol. 1959;148: 574–591.

5. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's viusal cortex. J Physiol. 1962;160: 106–154.

6. Friston K. Does predictive coding have a future? Nat Neurosci. Springer US; 2018;21: 1019–1021. doi:10.1038/s41593-018-0200-7

7. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annu Rev Psychol. 2004;55: 271–304. doi:10.1146/annurev.psych.55.090902.142005

8. Huq SF, Garety PA, Hemsley DR. Probabilistic judgements in deluded and non-deluded subjects. Q J Exp Psychol. 1988;40A: 801–812. doi:10.1080/14640748808402300

9. Lee TS, Mumford D. Hierarchical bayesian inference in the visual cortex. J Opt Soc Am A. 2003;20: 1434–1448.

10. Lochmann T, Deneve S. Neural processing as causal inference. Current Opinion in Neurobiology. 2011. pp. 774–781. doi:10.1016/j.conb.2011.05.018

11. Deneve S, Jardri R. Circular inference: Mistaken belief, misplaced trust. Curr Opin Behav Sci. 2016;11: 40–48. doi:10.1016/j.cobeha.2016.04.001

12. Spratling MW. Reconciling predictive coding and biased competition models of cortical function. Front Comput Neurosci. 2008;2: 1–8. doi:10.3389/neuro.10.004.2008

13. Friston K, Kiebel S. Predictive coding under the free-energy principle. Philos Trans R Soc B. 2009;364: 1211–1221. doi:10.1098/rstb.2008.0300

14. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active Inference : A Process Theory. Neural Comput. 2017;29: 1–49. doi:10.1162/NECO

15. Spratling MW. A review of predictive coding algorithms. Brain Cogn. Elsevier Inc.; 2017;112: 92–97. doi:10.1016/j.bandc.2015.11.003

16. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical

Microcircuits for Predictive Coding. Neuron. Elsevier Inc.; 2012;76: 695–711. doi:10.1016/j.neuron.2012.10.038

17. Brown H, Adams RA, Parees I, Edwards M, Friston K. Active inference , sensory attenuation and illusions. Cogn Process. 2013;14: 411–427. doi:10.1007/s10339-013-0571-3

18. Bishop C. Pattern Recognition and Machine Learning. Springer; 2006.

19. Johnson RR, Burkhalter AA. A polysynaptic feedback circuit in rat visual cortex. J Neurosci. 1997;17: 7129–7140. doi:10.1002/(SICI)1096-9861(19960506)368:3<383::AID-CNE5>3.0.CO;2-1

20. Hupé J, James A, Payne B, Lomber S, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature. 1998;394: 784–787. Available: https://search.proquest.com/openview/09bbe4a11f10a4097276702l9c4b017b/1?pq-origsite=gscholar&cbl=40569

21. Han B, Van Rullen R. Shape perception enhances perceived contrast : evidence for excitatory predictive feedback ? Sci Rep. Nature Publishing Group; 2016;6: 1–10. doi:10.1038/srep22944

22. Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006;9: 1432–1438. doi:10.1038/nn1790

23. Berkes P, Orban G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal inernal model of the environment. Science (80- ). 2011;331: 83–87.

24. Stephan KE, Bach DR, Fletcher PC, Flint J, Frank MJ, Friston KJ, et al. Charting the landscape of priority problems in psychiatry , part 1 : classification and diagnosis. The Lancet Psychiatry. Elsevier Ltd; 2016;3: 77–83. doi:10.1016/S2215-0366(15)00361-2

25. Stephan KE, Binder EB, Breakspear M, Dayan P, Johnstone EC, Meyer-lindenberg A, et al. Charting the landscape of priority problems in psychiatry, part 2 : pathogenesis and aetiology. The Lancet Psychiatry. Elsevier Ltd; 2016;3: 84–90. doi:10.1016/S2215-0366(15)00360-0

26. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012;16: 72–80. doi:10.1016/j.tics.2011.11.018

27. Friston KJ, Stephan KE, Montague R, Dolan RJ. Computational psychiatry: the brain as a phantastic organ. The Lancet Psychiatry. Elsevier Ltd; 2014;1: 148–158. doi:10.1016/S2215-0366(14)70275-5

28. Huys QJM, Maia T V., Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016;19: 404–413. doi:10.1038/nn.4238

29. Elmer GI, Leon Brown P, Shepard PD. Engaging Research Domain Criteria (RDoC):

Neurocircuitry in search of meaning. Schizophr Bull. 2016;42: 1090–1095. doi:10.1093/schbul/sbw096

30. Valton V, Romaniuk L, Steele D, Lawrie S, Seriès P. Comprehensive review: Computational modelling of schizophrenia. Neurosci Biobehav Rev. 2017;

31. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci. 2009;10: 48–58. doi:10.1038/nrn2536

32. Corlett PR, Taylor JR, Wang XJ, Fletcher PC, Krystal JH. Toward a neurobiology of delusions. Prog Neurobiol. Elsevier Ltd; 2010;92: 345–369. doi:10.1016/j.pneurobio.2010.06.007

33. Adams RA, Stephan KE, Brown HR, Frith CD, Friston KJ. The computational anatomy of psychosis. Front psychiatry. 2013;4: 47. doi:10.3389/fpsyt.2013.00047

34. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The Predictive Coding Account of Psychosis. Biol Psychiatry. Elsevier Inc; 2018; 1–10. doi:10.1016/j.biopsych.2018.05.015

35. Jardri R, Denève S. Circular inferences in schizophrenia. Brain. 2013;136: 3227–41. doi:10.1093/brain/awt257

36. Jardri R, Duverne S, Litvinova AS, Denève S. Experimental evidence for circular inference in schizophrenia. Nat Commun. 2017;8: 14218. doi:10.1038/ncomms14218

37. Notredame C-E, Pins D, Denève S, Jardri R. What visual illusions teach us about schizophrenia. Front Integr Neurosci. 2014;8: 1–16. doi:10.3389/fnint.2014.00063

38. Blake R, Logothetis NK. Visual Competition. Nat Rev Neurosci. 2002;3: 1–11. doi:10.1038/nrn701

39. Leptourgos P, Denève S, Jardri R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? Curr Opin Neurobiol. 2017;46: 154–161. doi:10.1016/j.conb.2017.08.012

40. Wilson HR. Computational evidence for a rivalry hierarchy in vision. Proc Natl Acad Sci U S A. 2003;100: 14499–503. doi:10.1073/pnas.2333622100

41. Moreno-Bote R, Rinzel J, Rubin N. Noise-Induced Alternations in an Attractor Network Model of Perceptual Bistability. J Neurophysiol. 2007;98: 1125–1139. doi:10.1152/jn.00116.2007

42. Hohwy J, Roepstorff A, Friston K. Predictive coding explains binocular rivalry: An epistemological review. Cognition. 2008;108: 687–701. doi:10.1016/j.cognition.2008.05.010

43. Brascamp JW, Klink PC, Levelt WJM. The 'laws' of binocular rivalry: 50 years of Levelt's

propositions. Vision Res. 2015;109: 20–37. doi:10.1016/j.visres.2015.02.019

44. Leopold DA, Wilke M, Maier A, Logothetis NK. Stable perception of visually ambiguous patterns. Nat Neurosci. 2002;5: 605–609. doi:10.1038/nn851

45. Ernst M, Banks M. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;415: 429–433. Available: http://www.nature.com/nature/journal/v415/n6870/abs/415429a.html

46. Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nat Neurosci. 2002;5: 598–604. doi:10.1038/nn858

47. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. PLoS One. 2007;2. doi:10.1371/journal.pone.0000943

48. Moreno-Bote R, Knill DC, Pouget A. Bayesian sampling in visual perception. Proc Natl Acad Sci U S A. 2011;108: 12491–12496. doi:10.1073/pnas.1101430108

49. Hudson TE, Maloney LT, Landy MS. Movement Planning With Probabilistic Target Information. J Neurophysiol. 2007;98: 3034–3046. doi:10.1152/jn.00858.2007

50. Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. Neuron. Elsevier Inc.; 2012;74: 30–39. doi:10.1016/j.neuron.2012.03.016

51. Acerbi L, Vijayakumar S, Wolpert DM. On the Origins of Suboptimality in Human Probabilistic Inference. PLoS Comput Biol. 2014;10. doi:10.1371/journal.pcbi.1003661

52. Drugowitsch J, Wyart V, Devauchelle A-D, Koechlin E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. Neuron. Elsevier Inc.; 2016;92: 1–14. doi:10.1016/j.neuron.2016.11.005

53. Larøi F, Sommer IE, Blom JD, Fernyhough C, Ffytche DH, Hugdahl K, et al. The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: State-of-the-art overview and future directions. Schizophr Bull. 2012;38: 724–733. doi:10.1093/schbul/sbs061

54. Waters F, Allen P, Aleman A, Fernyhough C, Woodward TS, Badcock JC, et al. Auditory hallucinations in schizophrenia and nonschizophrenia populations: A review and integrated model of cognitive mechanisms. Schizophr Bull. 2012;38: 683–692. doi:10.1093/schbul/sbs045

55. Reichert DP, Seriès P, Storkey AJ. Charles Bonnet Syndrome: Evidence for a Generative Model in the Cortex? PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1003134

56. Waters F, Collerton D, Ffytche DH, Jardri R, Pins D, Dudley R, et al. Visual hallucinations in the psychosis spectrum and comparative information from neurodegenerative disorders and eye disease. Schizophr Bull. 2014;40: 233–245. doi:10.1093/schbul/sbu036

57. Niemantsverdriet MBA, Slotema CW, Blom JD, Franken IH, Hoek HW, Sommer IEC, et al. Hallucinations in borderline personality disorder: Prevalence, characteristics and associations with comorbid symptoms and disorders. Sci Rep. Springer US; 2017;7: 1–8. doi:10.1038/s41598-017-13108-6

58. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. Science (80- ). 2017;357: 596–600. doi:10.1126/science.aan3458

59. Swanson LR. Unifying theories of psychedelic drug effects. Front Pharmacol. 2018;9. doi:10.3389/fphar.2018.00172

60. Harrison SJ, Backus BT. Uninformative visual experience establishes long term perceptual bias. Vision Res. Elsevier Ltd; 2010;50: 1905–1911. doi:10.1016/j.visres.2010.06.013

61. Stephan KE, Petzschner FH, Kasper L, Bayer J, Wellstein K V., Stefanics G, et al. Laminar fMRI and computational theories of brain function. Neuroimage. Elsevier Ltd; 2017; doi:10.1016/j.neuroimage.2017.11.001

62. Weilnhammer V, Stuke H, Hesselmann G, Sterzer P, Schmack K. A predictive coding account of bistable perception - a model-based fMRI study. PLoS Comput Biol. 2017;13: 1–21. doi:10.1371/journal.pcbi.1005536

63. Mathys C. A Bayesian foundation for individual learning under uncertainty. Front Hum Neurosci. 2011;5: 1–20. doi:10.3389/fnhum.2011.00039

# Résumé

Nous évoluons dans un monde incertain. De ce fait, notre survie dépend de notre capacité à prendre rapidement des décisions, et ce de manière fiable et adaptative. Il est possible de mieux comprendre cette capacité en considérant la perception comme un processus d'inférence probabiliste au cours duquel les informations sensorielles sont combinées à nos attentes pour produire une interprétation plausible de notre environnement. Les théories récentes de psychiatrie computationnelle suggèrent par ailleurs que la grande variabilité des troubles psychiatriques, au rang desquelles figure la schizophrénie, pourrait résulter d'une altération de ces mêmes processifs prédictifs. L'*Inférence Circulaire* est l'une de ces théories. Ce cadre de pensée stipule qu'une propagation incontrôlée d'information dans la hiérarchie corticale pourrait générer des percepts ou des croyances aberrantes. Afin d'explorer le rôle joué par l'*Inférence Circulaire* en condition normale ou pathologique, ce travail de thèse s'est appuyé sur des tâches de prise de décision en conditions perceptives ambigües.

Dans une première partie, nous nous sommes intéressés au rôle joué par la circularité dans la perception bistable. Le phénomène de bistabilité survient lorsque deux interprétations se succèdent à intervalle régulier pour un même percept. Nous présentons les résultats d'une tâche conduite en population saine où nous avons manipulé les informations sensorielles et à priori utilisées par les participants lors de la visualisation d'un cube de Necker (**article 1**). Nous avons pu montrer un effet propre à chaque manipulation, mais également une interaction entre ces deux sources d'information, incompatible avec une intégration Bayésienne optimale. Résultat confirmé par la comparaison de divers modèles computationnels ajustés aux données, qui a pu mettre en évidence la supériorité de l'*Inférence Circulaire* sur les modèles Bayésiens classiques. Nous avons ensuite voulu tester un modèle fonctionnel de la bistabilité (**article 2**). Nous avons donc dérivé la dynamique du modèle et montré que la présence de boucles descendantes dans la hiérarchie corticale, transformait ce qui était jusque là un intégrateur imparfait du bruit sensoriel en *modèle à attracteur bistable*. Ce modèle ne reproduit pas seulement le phénomène de bistabilité, mais également l'ensemble de ces caractéristiques phénoménologiques. Dans un **3ème article**, nous avons testé une prédiction, notamment en cas de présentation discontinue d'un stimulus bistable. Deux expériences complémentaires utilisant un paradigme de présentation intermittente du cube de Necker ont donc été conduites en population générale. Nos resultats etaient compatible avec les prédictions faites par le modèle de *l'Inférence Circulaire Dynamique*, suggérant que la circularité puisse être un mécanisme générique à l'origine de notre façon de voir le monde.

Dans la seconde partie de ce travail, nous avons étudié *l'Inférence Circulaire* en condition pathologique, notamment lors d'expériences psychotiques (schizophrénie, psychédéliques). Nous avons utilisé la perception bistable pour explorer les mécanismes computationnels à l'œuvre dans la schizophrénie (**article 4,5**). Nous avons comparé les performances de patients présentant des symptômes psychotiques à des témoins sains appariés lors d'une tâche de perception bistable. Nous avons pu montrer chez les patients une amplification des informations sensorielles combinée à une surestimation de la volatilité environnementale. Enfin nous terminons ce travail en proposant une approche transversale de l'effet des psychédéliques (**article 6**), sur la base des résultats précédents et de la spécificité clinique de ces expériences sensorielles cross-modales, afin de relier l'échelle macroscopique (i.e., comportement et phénoménologie), mésoscopique (i.e., les boucles inférentielles) et microscopique (i.e., les différents neurotransmetteurs impliqués aboutissant à un microcircuit canonique).



# Mots Clés

Inférence Bayésienne, Inférence circulaire, schizophrénie, psychose, perception bistable, Cube de Necker, psychédéliques, microcircuit canonique, hierarchique, fonctionnel, dynamique



# Abstract

We live in an uncertain world, yet our survival depends on how quickly and accurately we can make decisions and act upon them. To address this problem, modern neuroscience reconceptualised perception as an inference process, in which the brain combines sensory inputs and prior expectations to reconstruct a plausible image of the world. In addition to that, influential theories in the emerging field of computational psychiatry suggest that various psychiatric disorders, including schizophrenia, could be the outcome of impaired predictive processing. Among those theories, the circular inference framework suggests that an unconstrained propagation of information in the cortex, underlain by an excitatory to inhibitory imbalance, can generate false percepts and beliefs, similar to those exhibited by schizophrenia patients. In the present thesis, we probed the role of circular inference from normal to pathological brain functioning, gaining insights from perceptual decision making in the presence of high ambiguity.

In the first part of the thesis, we focused on the role of circularity in bistable perception in the general population. Bistability occurs when two mutually exclusive interpretations compete and switch as dominant percepts every few seconds. In a **1st article**, we manipulated sensory evidence and priors in a Necker cube task, asking how the brain combines low-level and high-level information to form perceptual interpretations. We found a significant effect of each manipulation but also an interaction between the two, a finding incompatible with Bayes-optimal integration. Bayesian model comparison further supported this observation, showing that a circular inference model outperformed purely Bayesian models. Having established a link between circular inference and bistable perception, we then put forward a functional theory of bistability, based on circularity (**2nd article**). In particular, we derived the dynamics of a dynamical circular inference model, showing that descending loops (i.e. a form of circularity resulting in aberrant amplification of the priors) transform what is normally a leaky integration of noisy evidence into a bistable attractor with two highly trusted stable states. Importantly, this model can explain both the existence and the phenomenological properties of bistable perception, making a number of testable predictions. Finally, in a **3rd article**, we tested one of the model's predictions, namely the perceptual behaviour when the stimulus is presented discontinuously. We ran two Necker cube experiments using a novel intermittent-presentation methodology, and we calculated the stabilisation curves (i.e. persistence as a function of blank durations). We found that participants' behaviour was compatible with the model's prediction for a system with descending loops, suggesting that circularity constitutes a general mechanism that shapes the way healthy individuals perceive the world.

In the second part, we studied circular inference in pathological conditions related to psychosis. We notably focused on two varieties of the psychotic experience, namely schizophrenia-related psychosis and drug-induced psychosis. After discussing the links between behaviour, aberrant message-passing and the corresponding neural networks (**4th article**), we used bistable perception to probe the computational mechanisms underlying schizophrenia in a **5th article**. We compared patients with prominent positive symptoms with matched healthy controls in two bistable perception tasks. Our results suggest an enhanced amplification of sensory inputs in patients, combined with an overestimation of the environmental volatility. In the last article (**6th**), we delineated a multiscale account of psychedelics, ultimately linking the macroscale (i.e. phenomenological considerations such as the crossmodal character of the psychedelics experience), the mesoscale (i.e. loops) and the microscale (i.e. neuromodulators and canonical microcircuits).



# Keywords

Bayesian inference, Circular inference, schizophrenia, psychosis, bistable perception, Necker cube, psychedelics, canonical microcircuit, hierarchical, functional, message-passing algorithms, dynamical