Comprendre les idéations complotistes de l'individu au groupe social : dimensions cognitives et computationnelles

Salomé Leclercq

THÈSE DE DOCTORAT

Spécialité : Neurosciences

En vue de l'obtention du grade de DOCTEUR

Préparée à l'École Doctorale Biologie Santé Lille

Unravelling the fabric of conspiracy theories from the individual to the community: Cognitive and computational dimensions

Soutenue par Salomé LECLERCQ

Le 11 Décembre 2023

Dirigée par le Professeur Renaud JARDRI

DEVANT LE JURY COMPOSÉ DE

Dr. Anne GIERSCH, MD, PhD, HDR, DR INSERM, Strasbourg	(Rapporteur)
Dr. Sylvain DELOUVÉE, PhD, HDR, MCF Université Rennes 2	(Rapporteur)
Prof. Frank LARØI, PhD, Professeur, Université d'Oslo	(Assesseur)
Dr. Mélissa ALLÉ, PhD, MCF Université de Lille	(Assesseur)
Prof. Renaud JARDRI, MD, PhD, PUPH, Université de Lille	(Directeur de thèse)

Université de Lille

“A belief is not merely an idea the mind possesses;
it is an idea that possesses the mind.”
Robert Oxton Bolton

“The popularity of conspiracy theories is explained by people’s desire to believe
that there is some group of folks who know what they’re doing.”
Damon Knight

Remerciements

Ces quatre dernières années ont été riches en émotions et il convient en premier lieu d'adresser mes remerciements aux très nombreuses personnes qui ont contribué à ma réussite par leur aide, leur soutien et leurs encouragements.

Mes remerciements reviennent tout d'abord au Professeur Renaud Jardri pour son accompagnement au cours de ces cinq dernières années. Vos grandes qualités humaines et académiques ont été les piliers du succès de ce projet. Merci d'avoir cru en moi depuis le master et de m'avoir offert l'opportunité de poursuivre en doctorat. D'être resté toujours serein et positif pendant les périodes d'incertitude et de m'avoir montré qu'il était possible de s'adapter aux situations les plus inattendues, voire même d'en faire un sujet de thèse ! D'abord fascinée par le sujet des hallucinations, je vous remercie de m'avoir insufflé une véritable passion pour le sujet des croyances et de m'avoir fait découvrir avec patience et pédagogie le monde de l'Inférence Bayésienne. D'avoir toujours trouvé le temps de répondre à mes interrogations malgré un emploi du temps très chargé. De m'avoir appris à allier rigueur scientifique dans l'élaboration d'un projet et élégance dans la présentation des résultats. De m'avoir montré le pouvoir que peuvent avoir une communication impactante et des compétences de vulgarisation scientifique. Plus globalement, je vous remercie de m'avoir fourni un cadre bienveillant dans lequel j'ai pu m'épanouir, m'exprimer et explorer mes idées de projets scientifiques. Veuillez trouver ici le témoignage de ma gratitude et de mon profond respect.

J'aimerais également remercier les membres de mon Jury et de mon Comité de Suivi, le Dr Mélissa Allé, le Dr Sylvain Delouvée, le Dr Anne Giersch, le Dr Anthony Lantian, le Pr Frank Larøi et le Pr Laurent Madelain pour avoir accepté de suivre, de lire et d'évaluer mon travail. De m'avoir challengée avec bienveillance et de m'avoir amenée à me questionner à chaque étape de cette thèse.

La plateforme CURE a été le théâtre de nombreuses joies comme de nombreux doutes depuis ma seconde année de master. J'aimerais particulièrement remercier toutes les personnes que j'ai eu la chance d'y rencontrer. Louise, pour avoir vécu cette aventure qu'est le doctorat avec moi, mais surtout en même temps que moi. Avoir pu échanger sur les nombreux challenges auxquels nous étions toutes les deux confrontées m'a beaucoup aidée à y faire face. Merci de m'avoir apporté un soutien quotidien, en particulier pendant la rédaction de ce manuscrit et merci pour les memes (« Je peux pas j'ai des deadlines »). Merci à Sébastien pour avoir été ma « personne-ressource » à chaque fois que je rencontrais un problème technique et pour les nombreuses heures passées à souffrir avec moi face aux caprices de PsychoPy. Tes bons conseils et ta bonne humeur ont rendu tout ce processus bien plus facile. Merci à Pantelis pour ta patience et ta pédagogie. Alors que j'ignorais tout de ce qu'était un modèle computationnel tu as su répondre à toutes mes questions avec passion et en te mettant toujours à mon niveau. Merci à David pour ton sourire et ton humour, et merci à Charlotte pour ta bienveillance et ton écoute. Merci à Lucie pour ta bonne humeur communicative, ton rire m'a aidée à dédramatiser en toute situation. Merci à Régine pour ton soutien moral et pour ton aide sur l'organisation. Merci à Pierre et Delphine pour le regard critique mais bienveillant que vous avez apporté sur mon projet.

Plus largement, j'adresse mes remerciements à toutes les personnes avec qui j'ai eu l'occasion d'échanger au cours de ces dernières années : internes, stagiaires, personnels et invités de la plateforme CURE, à tous les membres de l'équipe Plasticité et Subjectivité et du laboratoire Lille Neurosciences et Cognition ainsi qu'à mes co-auteurs. Merci à Céline pour m'avoir aidée à naviguer les méandres administratifs avec plus d'aisance et à Sophie pour m'avoir permis de partir en conférences dans de bonne conditions. Merci à Sophie et Vincent pour nos échanges scientifiques passionnés. Merci à Marielle et à Arnaud Cachia pour vos conseils précieux en statistiques. Merci à Maxime, Benjamin, Jacques, Marion, Léa, Antoine, Sylvain, Briac, Arnaud, Jane, Molly, Thibault, Tristan, Edouard, Matéa, Julie et bien d'autres qui avez tous contribué à votre manière à rendre l'atmosphère du laboratoire très agréable à vivre.

Je remercie également tous mes anciens professeurs ainsi que toutes les personnes qui m'ont donné ma chance, m'ont encadrée et m'ont encouragée au cours de mes études et sans qui je ne serai pas arrivée aux portes du Doctorat en premier lieu. Je remercie en particulier Maxime Hédouin pour ses qualités humaines, sa bienveillance et sa bonne humeur communicative. Tu m'as beaucoup appris.

J'adresse mes remerciements tout particuliers à mes parents Isabel et Jean-Jacques sans qui rien n'aurait été possible. Tout d'abord pour avoir toujours cru en moi et m'avoir toujours encouragée à suivre mes passions. Pour avoir trouvé les mots qui ont su me rassurer dans les plus grands moments de doute. Et enfin pour m'avoir apporté un soutien moral et financier conséquent tout au long de mes études en me répétant « c'est normal » à chaque fois que je vous remerciais. Je ne vous le dirai jamais assez, mais je peux déjà commencer ici, merci pour tout.

J'aimerais également adresser mes remerciements à ma sœur Ludivine, pour m'avoir donné l'exemple dans sa propre voie. Merci d'avoir pris le temps de partager ton vécu avec moi et de me rassurer dans les moments difficiles. Je remercie mon frère Damien pour avoir cru en moi et plus largement toute ma famille ainsi que mes amis. J'adresse mes remerciements particuliers à Océane pour son soutien sans faille. Grâce à toi je ne me sens jamais seule. J'aimerais également remercier la personne qui partage ma vie, Maxime, de m'avoir accompagnée avec le sourire, de m'avoir soutenue et de m'avoir aidée à m'aérer la tête au cours de cette période cruciale. Merci d'être là.

Enfin, j'aimerais remercier mon chat, Castiel, qui a constitué mon soutien moral le plus important de ces sept dernières années et dont les ronrons ont rendu cette thèse plus douce, en particulier la première année marquée par les confinements.

UNRAVELLING THE FABRIC OF CONSPIRACY THEORIES FROM THE INDIVIDUAL TO THE COMMUNITY
Cognitive and Computational Dimensions

Abstract

The pace of socio-political crisis has dramatically accelerated over the past decade and was accompanied by a surge in conspiratorial and pseudo-scientific beliefs. These inflexible ideas, widespread to varying degrees in the non-clinical population, can drastically influence a wide range of attitudes from health-related behaviors to political engagement. Furthermore, while it can be observed at an individual level, this global phenomenon of belief rigidification is mirrored at the community level by the major polarization and radicalization of online opinions. Adherence to conspiracy theories (CTs) has been proposed by some authors as a coping strategy aimed at restoring predictability in highly uncertain situations. This compensatory mechanism would be rooted in cognitive and perceptual inference biases that can be captured by Bayesian belief models.

This PhD thesis aimed at deciphering the mechanisms underpinning the emergence and maintenance of unshakeable conspiracy beliefs through three intertwined levels of comprehension: the cognitive, perceptual and computational approaches.

In the first axis, I explored the cognitive mechanisms associated with conspiracy ideations. I notably showed that hypersalience, a cognitive bias that consists in attributing great significance to irrelevant stimuli, was associated with adherence to CTs and vaccine hesitancy in the context of the COVID-19 pandemic. We also demonstrated that the debated link between the perceived lack of control over one's life could be experimentally captured with a behavioral task. I further showed that this association was stress sensitive and could be uncovered by real-world uncertainty.

Drawing on that idea, the second axis aimed at deciphering the dynamics of CTs around distressing and uncertain political events, combining an online bistable perception task with computational model fitting. Using the *Circular Inference* framework, we notably showed that when uncertainty peaks, CTs were associated with an overweighting of sensory information. In an attempt to cope with uncertainty, some participants particularly sensitive to stress adopted an exploration strategy that consisted in searching for simple and intuitive answers to complex issues. Progressively, this exploration strategy could shift to an

exploitation strategy in which increased adherence to CTs is associated with the amplification of prior information leading to a self-reinforcement of the belief system.

Finally, in the third axis, I addressed the question of belief rigidification at the community scale by modeling belief propagation in large social networks, as a form of probabilistic inference. We notably approached the phenomenon of polarization and radicalization observed in online communities as aberrant overconfidence rooted in some form of circularity in messages-passing, considered inherent to the network's structure. Going further, we demonstrated the validity of a novel algorithm, *Circular Belief Propagation*, in countering this aberrant overconfidence using data from Facebook© and Twitter©.

Public Summary

The last decade has seen a notable increase in socio-political crises and in the rise of conspiracy and pseudo-scientific beliefs. It has been suggested that some individuals may turn to conspiracy theories as a coping mechanism when confronted with the uncertainty and unpredictability brought about by such events. These rigid beliefs can have a substantial influence on various aspects of society, ranging from people's health-related behaviors to their political engagement. This PhD thesis aimed at better understanding the mechanisms underlying the emergence and maintenance of these unshakeable beliefs, especially in times of socio-political uncertainty such as the COVID-19 pandemic, or during high stakes events like presidential elections. In this perspective, I used online behavioral tasks and mathematical models to explore the links between specific cognitive and perceptual information processing and conspiracy ideations, across diverse populations experiencing real-world uncertainty.

Keywords

Conspiracy; belief; bistability; circular inference; Bayesian inference

COMPRENDRE LES IDEATIONS COMPLOTISTES DE L'INDIVIDU AU GROUPE SOCIAL
Dimensions cognitives et computationnelles

Résumé

Le rythme des crises sociopolitiques s'est considérablement accéléré au cours des dernières années, et s'est accompagné d'une montée en flèche des croyances conspirationnistes et pseudo-scientifiques. Ces idées inflexibles, largement répandues au sein de la population générale, ont eu un impact non négligeable sur les comportements individuels, qu'il s'agisse de leur choix de santé ou de leur choix d'engagement politique. Ce phénomène global de rigidification des croyances, s'il peut être observé au niveau individuel, trouve également son reflet à l'échelle des groupes sociaux, via la polarisation et la radicalisation des opinions en ligne. Certains auteurs ont formulé l'hypothèse du « complotisme » en tant que stratégie d'adaptation, qui viserait à rétablir la prévisibilité du monde face à des évènements très incertains. Ce mécanisme compensatoire s'appuierait sur des biais d'inférence, cognitifs et perceptifs, que les modèles Bayésiens sont censés pourvoir capturer.

Cette thèse de doctorat visait donc à décrypter les mécanismes qui sous-tendent l'émergence et le maintien des croyances complotistes à travers trois niveaux de compréhension : les approches cognitives, perceptives et computationnelles.

Dans le premier axe, j'ai exploré les mécanismes cognitifs associés aux idéations complotistes (IC). J'y montre notamment que l'attribution aberrante de saillance, un biais cognitif qui consiste à attribuer une trop grande importance à des stimuli non-pertinents, est associée aux IC et à l'hésitation vaccinale lors de la pandémie de COVID-19. Nous démontrons également que le lien, encore débattu, entre IC et perte du sentiment de contrôle sur le monde, peut être capturé expérimentalement par une tâche comportementale. Je montre enfin que cette association est sensible au stress et peut être révélée par l'incertitude du monde réel.

En m'appuyant sur ces résultats, je me suis appliqué dans le deuxième axe du travail à déchiffrer la dynamique des IC autour d'événements politiques incertains, en combinant l'utilisation d'une tâche de perception bistable en ligne et d'un modèle computationnel. En utilisant le modèle de l'*Inférence Circulaire*, nous montrons notamment que lorsque l'incertitude atteint son paroxysme, les IC sont associées à une prise en compte plus

importante des informations sensorielles. Pour faire face à l'incertitude, certains participants, particulièrement sensibles au stress, adopteraient une stratégie d'exploration consistant à rechercher des réponses simples et intuitives à des questions complexes. Progressivement, cette stratégie d'exploration évoluerait vers une stratégie d'exploitation dans laquelle l'adhésion accrue aux théories du complot est associée à l'amplification des connaissances a priori conduisant à un auto-renforcement du système de croyance.

Enfin, dans le troisième axe de la thèse, j'aborde la question de la rigidification des croyances à l'échelle du groupe, en modélisant la propagation des croyances dans les réseaux sociaux comme une forme d'inférence probabiliste. Nous avons notamment abordé les phénomènes de polarisation et de radicalisation communément observés dans les communautés en ligne comme un excès de confiance qui trouverait ses origines dans une forme de circularité inhérente à la structure du réseau. En outre, nous avons pu démontrer la validité d'un nouvel algorithme, le *Circular Belief Propagation* (CBP), capable de contrer cet excès de confiance en utilisant des données issues de réseaux réels, tels que Facebook© et Twitter©.

Résumé grand public

La dernière décennie a été marquée par une augmentation notable des crises sociopolitiques et par la montée en puissance des croyances conspirationnistes et pseudo-scientifiques. Certaines personnes se tourneraient vers les théories du complot pour faire face à l'incertitude et à l'imprévisibilité engendrées par de tels événements. Ces croyances rigides peuvent avoir une influence substantielle sur diverses attitudes, telles que les comportements liés à la santé ou l'engagement politique. Cette thèse de doctorat visait à mieux comprendre les mécanismes qui sous-tendent l'émergence et le maintien de ces croyances rigides, en particulier lors de périodes marquées par une forte incertitude sanitaire ou sociopolitique, telles que la pandémie de COVID-19 ou les semaines encadrant des élections politiques clivantes. Afin de répondre à ces questions, je me suis appuyée sur des tâches comportementales en ligne et des modèles mathématiques permettant d'explorer comment des différences de traitement cognitif et perceptif de l'information pouvaient être associées aux idéations conspirationnistes au sein de diverses populations exposées à une forte incertitude.

Mots-clef

Complotisme; croyance; bistabilité; inférence circulaire; inférence Bayésienne.

Contents

Chapter I

General Introduction

Foreword: Make my PhD great again

I started this thesis project in October 2019. After a master's project aiming at deciphering the inferential impairments underlying rigid beliefs and aberrant percepts in schizophrenia, it was finally time for me to start a PhD. Thrilled to delve deeper in that subject using neuroimaging tools, I happily started researching everything I could about electro-encephalogram (EEG) to elaborate my experimental design. The initial aim was to find associations between (i) delusion and hallucination proneness, (ii) inference mechanisms using the computational model of Circular Inference and (iii) electro-physiological patterns extracted from EEG recordings.

First cases of coronavirus were reported in France no more than three months later, in January. "*No need to worry about a virus that has the name of a Mexican beer*" I foolishly thought. Little did I know I was witnessing the beginning of a world-wide social crisis that would rhythm billions of lives for the next years. On March 2020, the 17th, French president Emmanuel Macron announced the first lockdown. Receiving participants at the lab (in a Hospital nevertheless) to administer a research protocol was deemed a lesser priority except if you were involved in the noble quest for the Holy Grail: a Covid-19 vaccine. How long would it last? Could we go back to the lab soon? Even if we did, would there be other lockdowns? And most of all, what about my thesis? A bunch of questions and no way to answer them using the scientific method… I was baffled by uncertainty.

Luckily, some people started coming up with answers. Internet seemed like the proper way to spread them fast and soon, social networks were full of them. Simple and intuitive answers. That's how we learned that "*the virus escaped from a Chinese lab*". Not long after, we realized that it did not literally escape, but that "*populations were voluntarily infected to target the more vulnerable among us and control the rest using lockdowns and other curfews*". Finally, we started noticing that "*the virus didn't even exist*". It was only a tale used to cover-up "*the mass implementation of trackable devices in our bodies through so-called vaccines*". Conspiracy theories were flourishing everywhere, and we started to wonder: what about these types of rigid beliefs resurging in non-clinical populations? Why are they particularly blooming during uncertain times? Are the underlying mechanisms comparable to those previously

found in delusions with which I was more familiar? Can we use similar tools to investigate them? Lucky for me, this time I could call upon critical thinking, data gathering, and statistical analysis as well as the refinement of online assessment methods to produce satisfactory answers to these questions. Could I make my PhD great again?

The pace of socio-political crisis has exponentially accelerated these last years. Gaining a better understanding of conspiratorial and pseudo-scientific beliefs that can drastically influence a wide range of attitudes ranging from health-related behaviors to political engagement constitutes one of the most crucial issues of our times. Future directions towards interventional and algorithmic procedures targeting the emergence of unshakeable beliefs resistant to scientific counter-evidences can only be elaborated based on a comprehensive overview of the mechanisms underpinning such extreme views. Drawing on this idea, the main objective of this work was to decipher the mechanisms underlying the instalment, maintaining and strengthening of conspiracy ideations and rigid beliefs in non-clinical populations and how they may relate to uncertainty. We decided to combine tools from cognitive & social psychology, psychophysics and computational modelling to address the multifaceted aspects of this complex issue.

In the first axis of my thesis, we explored the cognitive features of conspiracy theories adherence (CTs). I first focused on hypersalience, a cognitive bias that consists in attributing aberrant attention to irrelevant stimuli. Interestingly, this bias has previously been found involved in delusions (Kapur, 2003). We demonstrated that hypersalience was associated with CTs and vaccine hesitancy during the Covid-19 pandemic. We also explored how to address the debated link between the feeling of lack of control and CTs. Using an online bistable task, I showed that this association could be uncovered by conditions of heightened real-world uncertainty. More precisely, I showed that this effect was mediated by stress vulnerability and was domain-specific.

In the second axis of my thesis, we leaned on the use of computational modelling to investigate the dynamics of inferential processes associated with the progressive instalment and amplification of CTs in times of political uncertainty. Using the Circular Inference (CI) model, we showed that when uncertainty reached a climax, an overweighting of sensory information was associated with conspiracy ideations. In order to cope with uncertainty, some participants adopted an exploration strategy that

consisted in searching for simple and intuitive answers to complex issues. Progressively, this exploration strategy gave way to an exploitation strategy in which increased adherence to conspiracy theories was associated with the amplification of prior information leading to a self-reinforcement of the belief system.

Finally, in the third axis we addressed the question of belief rigidification at the community scale by modeling belief propagation in large-scale social networks as a form of probabilistic inference. We notably approached the phenomenon of polarization and radicalization commonly observed in online communities as aberrant overconfidence caused by circularity inherent to the network's structure. Going further, we demonstrated the validity of a novel algorithm, Circular Belief Propagation (CBP), in countering aberrant overconfidence in real networks structures using data from Facebook© and Twitter©.

I – A Bayesian approach of beliefs and perception

The Bayesian approach refers to a computational (or mathematical) framework that aims to understand and model how humans process cognitive and perceptual information to form coherent representations of the world. In this approach, the mind is described as a probabilistic machine that follows precise rules (or algorithms) in order to produce educated guesses, called inferences, regarding the current state of the world. These inferences are iteratively computed and refined through the optimal integration of newly acquired information and prior knowledge. This approach took its roots in probability theory, which consists in evaluating the chances of occurrence of random events. Starting with a way to compute the chances of getting a six in a simple throw of dice, this theory quickly emerged as a formal account of reasoning processes under uncertainty (Gigerenzer, 1989; Hacking, 1975). These models gained increasing attention in the scientific community and can now account for a broad range of inference processes including, but not limited to, symbolic reasoning (Oaksford and Chater, 2001), inductive learning and generalization (Tenenbaum et al., 2006) social cognition (Baker, 2007) and visual perception (Yuille and Kersten, 2006). Although these topics may seem to rely on very distinct mechanisms, the Bayesian formalism

approaches each of these concepts by trying to understand how humans build models of the external world so quickly and efficiently, drawing relevant information from a constant flow of noisy and partial inputs (Chater et al., 2010). In this manuscript, we will specifically explore how biased inference processing at the cognitive and perceptual level can account for the emergence of rigid beliefs, taking a special focus on conspiracy theories. But first, we need to detail how Bayesian models of information processing can account for beliefs formation and perception.

First, let us define beliefs and develop how Bayesian models explain their emergence. Beliefs are subjective phenomena involving the attribution of a certain degree of plausibility to observable events. At the cognitive level, belief generation is thought to be embedded in a Bayesian inference system, i.e., the optimal combination of sensory and a priori information (Breen, 1999). By gathering available information and prior knowledge about the world, we are able to assign a certain level of confidence to concepts such as the existence of a God, the possibility of life appearing on another planet or the validity of the scientific method. The ability to generate and update these beliefs is crucial, as it is by building a unique belief system that we construct our own image of the world and make informed decisions in all aspects of our life.

Let's illustrate this notion with an example. As many people, I am not a big fan of biking under the rain. I also dislike taking the bus when it's sunny. We can thus imagine that whether I take my bike or the bus to the lab today will depend on my belief about the weather. To make this decision, I can rely on the information currently available, such as looking out of my window to check the current weather conditions. However, for my weather predictions to be optimal, I need to combine this information with my prior knowledge of the weather, which includes factors like the season and the meteorological forecasts I watched on TV this morning. Finally, for this integration to be complete, I also need to assign a certain level of plausibility to each piece of information. For example, the extent to which I can trust the weather forecasts from a particular channel based on my past experiences. While this bike-or-bus example may seem trivial (being wrong simply resulting in getting wet), understanding how agents integrate and update information to make decisions and adapt to ever-changing and highly uncertain environments represents a crucial issue that can be applied to a broad range of situations with much higher stakes, such as choosing between candidates in

an election, choosing to get vaccinated or not, or even choosing one's attitude and behavior towards climate change stakes.

Bayesian approach provides a nice mathematical framework specifying how agents solve this integration of information in order to attribute a certain degree of plausibility to any conjecture about the world (Chater et al., 2010). These degrees of plausibility are represented by numerical values ranging from 0 to 1, where 1 corresponds to absolute certainty that an occurrence in the world is true ("*I'm certain it will rain today*") and 0 that it is false ("*I'm certain it won't rain today*"). In simple terms, these values can be derived from the Bayes' Theorem, expressed as follows:

$$(1) \quad P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

This equation represents how we can integrate new information (like observing clouds in the sky) with prior knowledge (pre-existing beliefs about the weather) to update beliefs in an uncertain environment. As such, P(A|B) is the probability of event A occurring, given that event B has occurred. In our example, this is the probability of a rainy day, given that the weather is cloudy. P(B|A) represents the probability of the weather being cloudy (B), given that the day is rainy (A). This is the likelihood of cloudy weather when it's rainy. P(A) corresponds to the prior belief (or prior probability) that a day will be rainy, based on factors like the season, historical data or meteorological forecasts. Finally, in this example, P(B) would be the total probability of cloudy weather, which can be calculated as the sum of the probabilities of cloudy weather in all possible scenarii. If, as a result of these computations, P(A|B) is close to 1, I am more likely to take the bus as I grant a high level of confidence in the assumption that it will rain. If, on the contrary, P(A|B) ends up close to zero, I might prefer taking my bike being confident that the weather will be good!

While Bayesian formalism provides well-suited framework to account for beliefs generation, the same principles can be applied to perceptual processing under uncertainty (Friston, 2005; Kersten et al., 2004; Mamassian et al., 2002). While this is true for any type of perception, we will take a specific focus on visual processing. In Bayesian terms perception can be described as an inferential process that combines

external sensory signals with prior knowledge and beliefs about the world (Helmholtz, 1866). This process would enable the observer to build a stable and unified perceptual experience out of noisy and ambiguous sensory information. A famous example of perceptual inference is the visual perception of depths that consists in the three-dimensional (3D) reconstruction of objects projected from a two-dimensional (2D) retinal image (Helmholtz, 1866).

Perceptual processing of bistable figures such as the *Necker Cube* constitutes a specific case of visual inference under heightened conditions of uncertainty (Mamassian et al., 2002; Sterzer et al., 2009). A bistable figure is an ambiguous or noisy stimulus that does not variate across time but offers two mutually exclusive interpretations (**Figure 1**). Confronted to such an image, the conscious perception of an observer will naturally alternate between the two possible configurations, proving that a basic visual input is combined to a subjective interpretation and contextualization before becoming a conscious percept. This implication of subjective re-interpretation is illustrated by the fact that voluntary control can influence bistable perception (Brouwer and van Ee, 2006; Intaitė et al., 2010). While some authors argue that this perceptual control mechanism that has recently been interpreted as visual flexibility (Rodríguez-Martínez, 2023). depends of executive processes, as selective attention and inhibitory control (Bialystok and Shapero, 2005).

There are several types of bistable images that have been explored and used in experimental paradigms (Bialystok and Shapero, 2005; Long and Toppino, 1981; Rodríguez-Martínez and Castillo-Parra, 2018). Figure-ground reversals correspond to figures where the background of the first salient interpretation can be recognized as an object in the second representation (for example the vase-face illusion **Figure 1-A**). *Meaning-content reversals* are figures that alternates between two configurations equally salient which differ in terms of meaning (for example "*My girlfriend or my mother-in-law*" **Figure 1-C**). Finally, *perspective reversals* corresponds to figures for which the two interpretations correspond to different orientations of the same object (see the Necker Cube or Shröder's staircase, **Figure 1-B,D**). This last category provides particularly interesting stimuli for research, given the fact that the two interpretations are equal in their salience and significance, making perception less likely to be biased towards one or the other.

Figure 1 : Types of bistable images (Rodríguez-Martínez and Castillo-Parra, 2018). From left to right: **(A)** Rubin´s vase-face illusion (figure-ground reversals); **(B)** The Necker cube (perspective reversals); **(C)** "*My girlfriend or my mother-in-law*" (meaning-content reversals); **(D)** The Schröder reversible staircase (perspective reversals).

According to David Marr, concepts such as cognitive reasoning or visual perception could be approached at three complementary levels of comprehension (Marr, 1982, **Figure 2**). First, the *computational level* which consists in formalizing the problem the agent is trying to solve. This level specifies the goal of the operation and the structure of the environment it takes place in. Some examples of the problems we are considering here are the reconstruction of a three-dimensional representation from a two-dimensional visual scene, or determining if a conjecture is true or not. Second, the *algorithmic level* specifies the operations required to achieve this goal from a probabilistic perspective and by breaking down the process in simple steps. For example, in the case of visual perception, it would describe how the hierarchical system extracts basic features from visual receptors, then uses this information to infer the structure of two-dimensional surfaces and their orientations to finally generate an internal representation of a three-dimensional visual scene. This second level of complexity also describes the mathematical algorithms used by the brain to go from one step to another until the goal is achieved. Finally, the third level corresponds to the *implementation or hardware level*. It accounts for the way brain circuitry conveys information according to the algorithms described at the second level at the physiological scale.

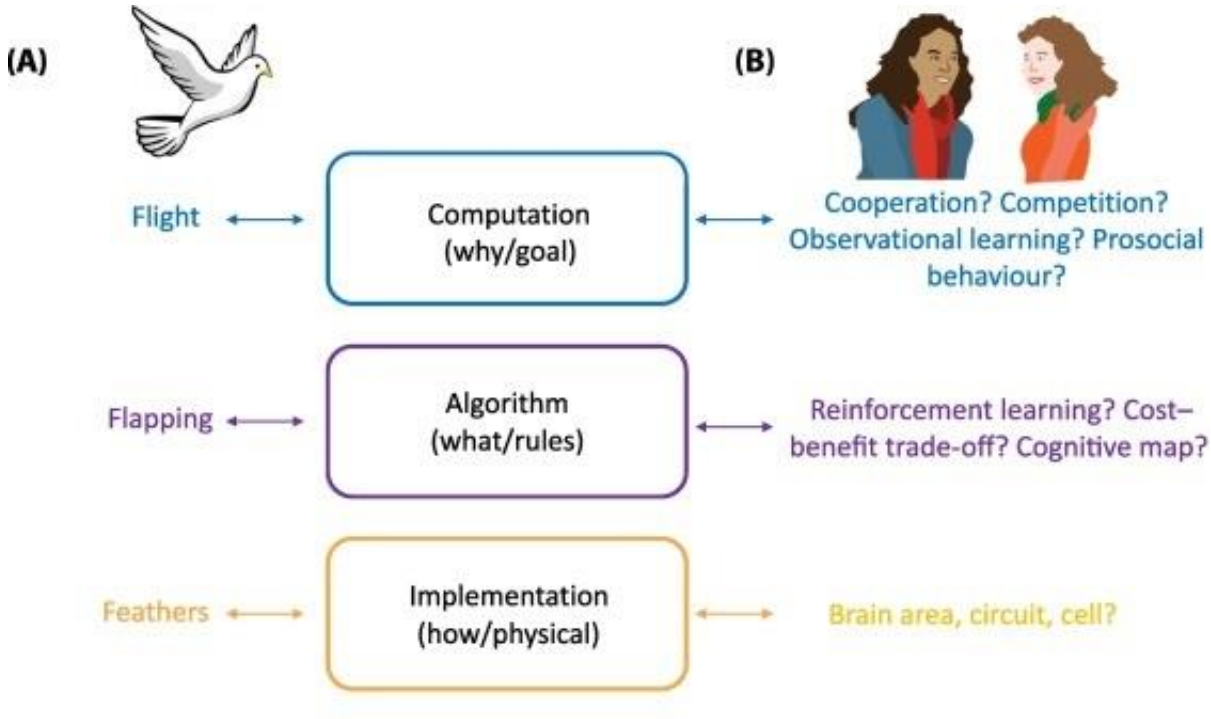


Figure 2 : Marr's Three Levels of Analysis for Non-Social and Social Behaviors (Lockwood et al., 2020). (A) To understand how a bird flies, we can build a model that follows Marr's three levels of analysis. First, we need to understand what the agent is trying to accomplish, in this case flying (computational level). Second, we need to formalize the operations required to fly, for example flapping its wings up and down (algorithmic level). Finally, we can think of the physical structure that conveys such action, like the aerodynamic features of the feathers (implementation level). **(B)** This method can be applied to understand more complex issues such as human social behavior. Computational models can be used to capture the processes involved at the algorithmic level while brain imaging studies are often designed to investigate the implementation level.

In the present manuscript, we will mainly focus on the computational (I) and algorithmic (II) levels of description to decipher the mechanisms involved in the formation of non-clinical rigid beliefs. I will notably try to understand the impact of distinct factors (e.g., uncertainty, stress and lack of control) that could lead to reasoning and perceptual biases. The investigation of the "hardware" level (III) will be later conducted by the team in the years to come, notably by using high-density electroencephalographic recordings concomitant to probabilistic tasks, some examples of which I will introduce in the following sections of the dissertation.

Key concepts

- Bayesian models are mathematical frameworks based on the Bayes' theorem that stipulate the mathematical operations through which the brain computes probabilistic inference.
- Inferences are conceptualized as the optimal combination of prior knowledge and current noisy sensory inputs to generate a mental representation.
- These models of information processing provide a unified framework that can account for the emergence of both beliefs and perception under uncertainty.

While Bayesian models of cognition perfectly describe rational agents and how they would perform exact inference, we know that information processing can frequently be biased (Acerbi et al., 2014; Beck et al., 2012). Perceptual illusions constitute good examples of such inaccurate integration (Notredame et al., 2014). Different levels of probabilistic reasoning can be involved in this inaccuracy, ranging from reduced precision of the external source to the internal inability to update current beliefs. These Bayesian processes can be assessed using experimental paradigms. In the following section, I will briefly describe three types of probabilistic tasks, each aiming to investigate a different aspect of inference processing.

II – A brief focus on experimental paradigms based on Bayesian reasoning

Over time, various experimental paradigms have been proposed to investigate how individuals make probabilistic judgements and update their beliefs in situations of uncertainty. Here, I will focus on three experimental frameworks in particular, that were fitted with computational models in order to characterize bayesian processing differences in clinical and non-clinical samples.

Probabilistic decision tasks constitute a first group of experiments, in which individuals are required to make choices or decisions related to uncertain or risky

outcomes. These type of tasks are designed to study how individuals integrate and combine information to form these probabilistic judgements. The conventional "*Beads task*" (Brett-Jones et al., 1987; Phillips and Edwards, 1966) constitutes an iconic example of this class of experimental paradigms that has been extensively used and described.

In the "draws to decisions" variant of this task, participants are shown two jars containing colored beads (Garety et al., 1991; Huq et al., 1988). Beads have two possible colors (in our example red and blue, see **Figure 3**) and the two jars contain a reversed beads-ratio of each colour. Participants are told that beads are going to be extracted from one of the two randomly-chosen jars, and that they have to guess wich one it is. Each trial starts with the presentation of the jar and the extraction of one bead. Participant can then chose to indicate their decision about the jars if they have reached certainty, or they can aslo wait to see another bead (up to twenty beads).

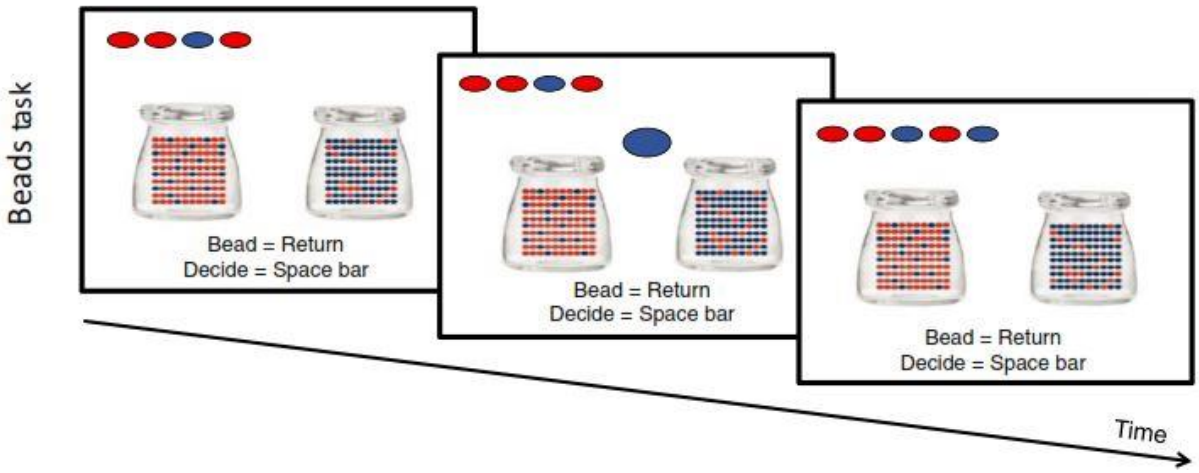


Figure 3: The beads task (Voon et al., 2016). In this experiment, participants are presented with two jars, each filled with red and blue beads. Participants are informed that a random selection will be made between the two jars, and their task is to predict which jar will be chosen. Each trial can be decomposed into three steps. The jars of varying ratio are displayed (1) followed by the extraction of a single bead that is presented (2). At this point, participants have the option to make their prediction regarding the jars if they feel confident, or they can choose to wait for the extraction of another bead (3).

Making a decision based on a very small number of beads was defined as a *jumping-to-conclusions* (JTC) bias (Garety et al., 2005). JTC corresponds to a bias where individuals make hasty probabilistic decisions based on limited sensory

evidence, but with increased levels of confidence in their decisions. Different versions and variants of this task have been used, changing for example the beads-ratio (called contingencies) to represent different levels of contextual uncertainty.

Another key concept in the Bayesian approach to beliefs is their ever-updating nature. Indeed, Breen (Breen, 1999) proposed that considering belief formation through the lens of probabilistic learning could allow us to go beyond fixed conceptions, such as considering beliefs as the simple product the internalization of societal norms and to understand how, on the contrary, beliefs evolve in light of experience. Let's illustrate that using again my weather example: if I look through my window and see a bright blue sky when I wake up, I might rationally believe that it won't rain today. Now, if the sky has turned grey and become laden with heavy clouds while I've been having my morning coffee, my belief about today's weather will certainly change based on this new information and my prior belief that the environment is susceptible to change or is volatile.

Tasks assessing flexibility to changing contingencies are another important paradigm that emerged from this need to study more complex and dynamic decision-making scenarii. Derived from classical learning and decision-making paradigms such as the *Wisconsin Card Sorting Test* (Grant and Berg, 1948) and *Iowa Gambling Task* (Bechara et al., 1994), the "*Reversal-Learning task*" (**Figure 4**) focuses on how participants integrate feedback from past probabilistic decisions and their ability to adapt to changing contingencies. During the task, participants are usually presented with two options, such as a choice between symbols or decks of cards from which they can draw. Each option is associated with a given probabilistic outcome. For example, choosing the first option might lead to a reward (positive outcome) 80% of the time and a punishment (negative outcome) 20% of the time. Conversely, choosing the second option is associated with a different set of probabilities, this ratio being referred to as the "option's contingency". During the initial learning phase, participants must infer these contingencies along trials and adapt their behavior accordingly. During the following reversal phase, the outcome probabilities are reversed. Participants must detect this change and adapt their choices to these new contingencies.

Several parameters representing participants performance can be extracted from this type of paradigm, such as the learning rate (how quickly participants learn which option is the most rewarding), reversal detection (the ability to recognize a

change in contingencies), adaptation (how efficiently participants adapt their choices after detecting the reversal) and persistency (how likely participants are to continue to choose the previously rewarded option out of habit after a reversal), a proxy of cognitive flexibility. Some variants of this task including a social component have also been proposed, asking for example participants to choose a game partner instead of a symbol.

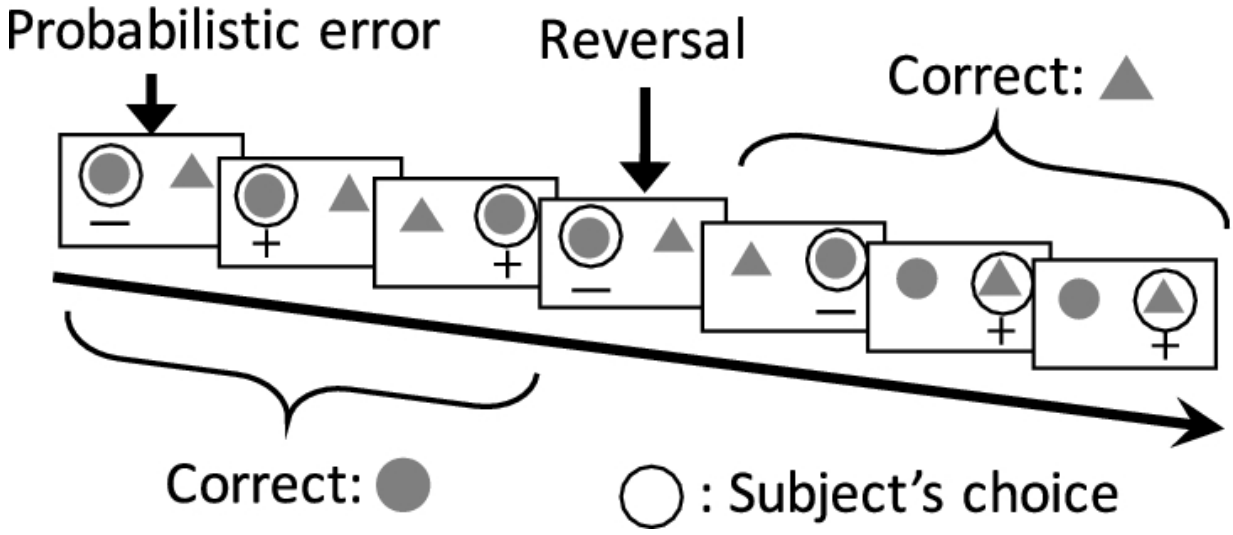


Figure 4 : Example of a reversal-learning task (Masumi and Sato, 2021). In this example, participants are asked to find out the correct option between two stimuli under changing contingencies. Each trial can be decomposed in three steps. The participants are presented with a circle and a triangle (1) and use their keyboard to choose between these two options (2) before getting a feedback indicating which was considered as the "correct" answer. This illustration represents a changing in contingencies across trials. In first three trials, the circle was defined as the "correct" stimulus. In fourth trial, contingency between stimuli and reward is reversed, and the triangle is defined as "correct" stimulus.

Finally, **bistable perception tasks** can be used to assess the perceptual side of inference processing. While the use of the different bistable images we presented in the previous section can present various advantages, we will focus here on the *Necker Cube* (NC) paradigm.

During the task, participants are usually presented with a projection of the NC displayed on the screen, either continuously or intermittently. At each trial,

participants are asked to report their subjective interpretation of the cube, either *seen from above* (SFA) or *seen from below* (SFB - **Figure 5**). Prior expectations and visual features of the NC, like contrast levels or tilt degree, can be manipulated to reduce uncertainty or bias perception towards an interpretation or the other (Dobbins and Grossmann, 2010; Leptourgos, Notredame, et al., 2020). Assessing the impact of these manipulations over perceptual behavior can provide substantial information about Bayesian processing of visual information under conditions of maximized or reduced uncertainty.

Different metrics can be extracted from the participants' answers to characterize their perceptual behavior. Among those criteria we can cite the *relative predominance* on one hand that correspond to the probability of experiencing each percept and constitutes a direct reflection of perceptual preferences. On the other hand, the mean phase duration, mean switch rate and survival probability are three metrics providing additional information about the temporal dynamics of inference processing. *Mean phase duration* corresponds to the time during which the perception of an interpretation persists before a perceptual switch (i.e., a change in the NC interpretation). The *switch rate* is a close parameter that can be defined as the mean number of switches experienced during a given period. Finally, the *survival probability*, also referred to as ***perceptual stability***, corresponds to the probability that a percept persists from one trial to the next. The interest for this later metric is driven by the idea that according to the Markov model, the current percept (i.e., one of the two cube interpretations, SFA or SFB) depends on the previous percept, but also its updating by sensory observation. This assumes a circularity in information processing, where the percept at time *t* becomes an a priori information at time *t+1*.

However, while the Bayesian inference framework provides a nice account for the mechanisms underpinning bistability dynamics, elementary features of the visual system such as eye-movements have also been shown implicated in perceptual switches. Based on this idea, ocular temporal windows, that can be derived from the dynamics of ocular fixations, were recently proven an interesting alternative to simple manual responses to account for perceptual dynamics (Polgári et al., 2020).

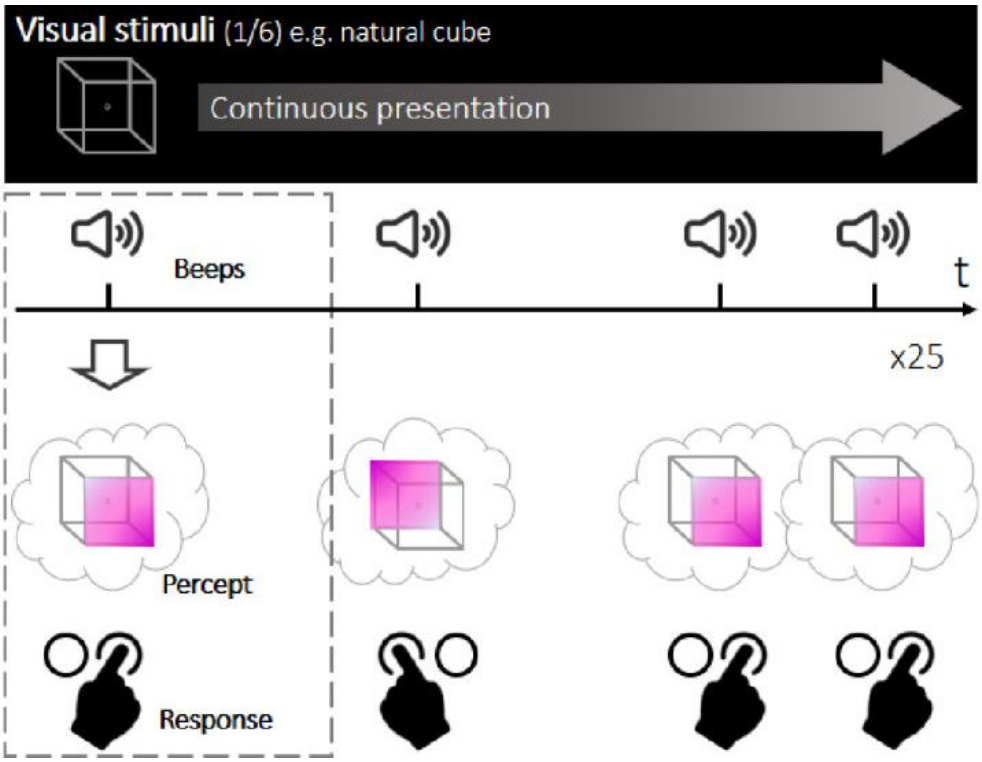


Figure 5 : Example of a bistable perception task (Leptourgos, Notredame, et al., 2020). Here, a Necker Cube (NC) is continuously presented to participants. Each trial can be decomposed into three steps. After a pseudorandomized duration (1), an auditory cue is presented (2) indicating that participants are asked to report their current interpretation of the stimulus: 'seen from above' (SFA) or 'seen from below' (SFB), using the right or left arrow of their keyboard, respectively (3).

Key concepts

- Real-world information processing can be biased and inaccurate.
- Various experimental set-ups can be used to study how people make probabilistic judgments and update their beliefs in uncertain situations.
- Probabilistic decision tasks evaluate how participants makes choices related to uncertain outcomes.
- Flexibility to changing contingencies paradigms assess how participants update their belief to changing probabilities and contingencies.
- Bistable perception tasks focus on the dynamics of perceptual inference processing.

In this PhD thesis, I will rely on a bistable perception task to assess deviations from optimal inference. Across each chapter, I will put efforts to better understand how distorted inferential processes could lead to rigid beliefs, choosing to put a special emphasis on conspiracy ideations and how they crystallize around highly uncertain contexts.

This experimental work found its inspiration in previous work conveyed on inflexible beliefs in the pathological context of psychosis. In the next sections of the thesis, we will subsequently (1) try to elucidate the differences characterizing rigid (albeit) non-clinical beliefs, such as conspiracy ideations, from pathological beliefs such as delusions, (2) develop the main findings advocating for the implication of impaired inference in the psychosis spectrum, mainly focusing on paranoid ideations, and (3) summarize the information processing biases that have been found associated with non-clinical beliefs, especially conspiracy beliefs.

III – "Nothing is worse than a rigid belief": change my mind.

Inflexible non-clinical ideas, such as paranormal ideations, are considered by some people as plausible and by others as irrationals. Should we nevertheless define such beliefs as pathological? In a broader perspective, what distinguishes a rational, socially adapted belief from an aberrant, pathological one? This question is of both philosophical and medical interest and represents a major societal stake. Indeed, the normalization or pathologizing of a belief can drastically influence how the community views that belief. According to Bortolotti (Bortolotti, 2023), the pathologization of beliefs such as conspiratorial ideas is not without consequences. This author explains that if society recognizes an idea as the pathological expression of a dysfunction, the latter risks losing its place in public debate, no longer being considered as an opinion that can be discussed or countered, but as an abnormality of the mind to be treated. In this way, he points to the danger for people expressing these ideas of simply being strongly discredited. In the following section we quickly review some of the main approaches of normality that were proposed and their limits before highlighting the interest of a dimensional approach based on conviction degree and belief flexibility.

Attempting to address this question, different criteria have been proposed (notably by psychiatrists) to distinguish between a rational, socially adapted belief and an aberrant, pathological one. According to Devereux (Devereux, 1970), normative behaviors are prescribed by the society in which an individual evolves, meaning that deviance is culturally coded. Drawing on this idea, normative approaches posit that a belief could be considered as "normal" (as far as this notion exists) if it has a collective value (i.e., if it is widely shared among the group members). This idea was quickly set aside considering that a shared consensus is context-dependent and thus can drastically vary across times and cultures. As an example, Galileo's claim that Earth was not the centre of the universe was considered so misplaced regarding the belief system in place at his time that he was judged and condemned for heresy (Finocchiaro, 2005). Nevertheless, the same belief would be considered perfectly adapted nowadays and it is thus difficult to qualify such a belief as aberrant. Moreover, culture provides individuals with models of conduct and misconduct codifying what is usually considered as an expression of illness. Each culture can have a different representation of illness, associated with a dedicated repertoire of pathologies and cares. A good example of cultural influence on the way of interpreting and treating disorders is *Wendigo*. The Wendigo is a form of psychosis, specific to the native-American tradition, in which the individual believes to be possessed by the Wendigo spirit and exhibits cannibalism (Volkan et al., 2021). Acknowledging that our modern western system provides a model of normality and pathology inherently tied to our culture thus immediately raises the question of a lack of objectivity in such a normative approach.

A second approach posits that a good criterion to define pathology resides in the coherent nature of a belief system. The problem with this claim is that a system can be coherent without being necessary true. This is notably the case for different famous fantasy universes like *Lord of the Rings* (Tolkien, 1954) or *Harry Potter* (Rowling, 1997), where every magical or supernatural element of the story meticulously follows an arbitrary set of interconnected rules defined by the authors. While these literary masterpieces are impressively consistent (and I patiently waited for my letter to Hogwarts as a kid), I learned soon enough that magic wasn't likely to exist. This point also stands for more tangible and subversive structures like cults. These groups of people typically base their beliefs on a set of religious or metaphysical

ideas separated from those of the larger group to which they belong. Even if the system displays a coherent nature in which "*everything apparently fits together*", these social structures often isolate individuals in order to strengthen the cognitive, emotional or financial grip the organization holds on them which can of course have dramatically nefarious effects on their life (Stein, 2021).

Despite their interest, those two first attempts to define pathological beliefs are still unsatisfactory. A more clinical approach proposed to consider the nefarious impact of beliefs over one's well-being as a good criterion of pathology. In this perspective, distress, excessive preoccupations and behavioral interference were proposed as central dimensions of delusional ideations (Brett-Jones et al., 1987; Chadwick and Lowe, 1994; Peters et al., 2016). Pathological beliefs are a major source of distress for the individuals who holds them. For instance, persecutions ideas make the subject feels she/he is the target of other's harmful intents, resulting in constant fear. Those individuals may then choose to withdraw from social interactions, out of fear and distrust. This isolation in turn promotes feelings of loneliness and anxiety and has a negative impact on the person's relationships. Furthermore, the delusions themselves can put a strain on social relationships. The sometimes-odd nature of the beliefs and the inability of others to convince the patient that her/his ideas are irrational might lead to misunderstanding, stigmatization, poor social integration and conflicts, again triggering feelings of anxiety in a totally vicious circle. Another important dimension of beliefs is *preoccupation*, which refers to the intensity and persistence of recurring thoughts about them. In other words, "how much" or "how often" one's thought content relates to the belief. There is also an intrusive nature to preoccupations where thoughts can substitute themselves to other contents and disrupt attentional processes (Kalivas and Paulus, 2021). Drawing on that idea, the amount of influence held on volitional processes and behavioral manifestations has been proposed as a complementary dimension of pathological beliefs referred to as *behavioral interference*. Such interferences can be dramatically time-consuming as exemplified by the engagement in rituals in obsessive-compulsive disorders (Hollander, 1997). Patients with delusional beliefs might also engage in dangerous or self-harming behaviors (Pahuja et al., 2020). In the most severe cases, interferences can manifest through a loss of functionality and autonomy where delusions interfere with patient's abilities to attend their job or maintain daily activities or to take care of themselves.

While characterizing the deleterious nature of beliefs is of higher importance in a clinical context, this approach provides only a partial answer to the question we want to address. This is why, I would like to focus on two complementary dimensions that will enable us to characterize the degree of adaptation or aberration of a given belief, and which can be conceptualized using a Bayesian approach: conviction and belief flexibility. Approaching belief formation and maintenance through these dimensions provides three substantial advantages. First, it provides a satisfactory framework in characterizing aberrant and pathological beliefs independently of societal contents or subjective judgements about the plausibility of the content. Second, it allows to go beyond a simple "clinical / non-clinical" dichotomy and considers beliefs as a complex product of cognitive mechanisms situated along a continuum ranging from normal to pathology. Finally, the Bayesian framework can nicely account for differences in conviction and flexibility. Indeed, probabilistic reasoning biases can affect these dimensions of belief leading to excessive conviction, referred to as overconfidence and altered flexibility conceptualized as failures in updating our beliefs according to new inputs.

Conviction is defined as the degree of certainty that a belief is true (American Psychiatric Association, 2013; Appelbaum et al., 2004; Kendler et al., 1983) or resistant to modification (Leeser and O'Donohue, 1999). A high degree of conviction constitutes one of the main features of pathological beliefs, such as delusions, and can also be observed in obsessive-compulsive disorders (Brakoulias and Starcevic, 2011) or dysmorphophobia (Rossell et al., 2020). Regarding non-clinical populations, Robert Abelson (Abelson, 1988) approached conviction considering socio-political issues and proposing three components to characterize beliefs: (i) emotional commitment, (ii) ego preoccupation and (iii) cognitive elaboration. Emotional commitment corresponds to the strength of the individual's feeling or emotional response towards an issue, while ego preoccupation involves the subjective levels of importance and concerns attributed to and associated with the issue. Finally, cognitive elaboration can be defined as the subjective feeling of being knowledgeable about the issue. Belief flexibility on the other hand can be defined as the individual's ability to question and criticize a belief and to express doubts about it. On the contrary, rigidity would correspond to an aberrant resistance of the belief to reasonable doubt and counter-evidence. Belief flexibility can

also be approached in a mechanistic way through the assessment of belief updating dynamics (Fromm et al., 2023).

The concept of clinical threshold appears relevant in medical contexts since it helps to determine when to intervene and allows for the proposal of adapted treatments and care. Nevertheless, considering beliefs through a dimensional approach (as opposed to the classic categorical medical model), that takes into account the intensity of these different dimensions, allows us to consider their underlying mechanisms beyond an "all or nothing" perspective. It also implies that some disorders are more severe than others, but more importantly that we can situate beliefs such as magical and conspiratorial ideations on a common continuum, ranging from normal to pathology, independently of their exact content.

Key concepts

- Various criteria have been proposed to distinguish rational from pathological beliefs.
- Conviction, or the degree of certainty in a belief, is a key aspect of beliefs' characterization.
- Belief flexibility involves the ability to question and criticize a belief, while rigidity resists doubt and counter-evidence.
- Some biases in probabilistic reasoning can affect these dimensions causing overconfidence and sub-optimal belief updating.
- Characterizing beliefs through these dimensions allows us to position non-clinical beliefs along a continuum, ranging from normal to pathology.

In the next sections we will progressively highlight some inferential mechanisms underlying rigid beliefs along this continuum. We will start by describing biases associated with rigid beliefs in the psychosis spectrum that inspired the present work, notably putting a special emphasis on paranoid concerns. Then, we will present how these biases have also been found associated with non-clinical rigid beliefs, focusing preferentially on conspiracy ideations.

IV – Down the rabbit-hole: aberrant inference in the schizotypal spectrum

Schizophrenia (and more broadly, psychotic disorders) can be defined as a severe mental illness that manifests through an impoverishment of the associations between thoughts, emotions and behaviors, along with alterations in the sense of reality. The idea that personality traits, cognitive patterns and perceptual experiences related to these psychotic states are not "all-or-nothing" phenomena, but rather form parts of a continuum that is widely present in the non-clinical population, is gaining momentum (Guloksuz and van Os, 2018). The term "*schizotypy*" that has been successively conceptualized as a milder form of schizophrenia (Meehl, 1962; Rado, 1953), a personality dimension (Eysenck, 1960), and both a healthy variation and a predisposition to psychosis (Claridge, 1997) refers to this continuum, ranging from non-clinical to pathological experiences. On the other hand, the "*psychosis spectrum*" term is mostly used to account for the multidimensional and polymorphic nature of psychosis. Biased inference can be captured at different degrees along this continuum and might account for specific symptoms of psychosis and schizotypy related to beliefs and perceptions.

Among these symptoms, delusional thinking constitutes the ideal example of rigid and unshakeable beliefs. The occurrence of these inflexible ideas has been extensively investigated and found repeatedly related to differences in inference processing that we will detail in the following paragraphs. Delusions are defined in the 5th edition of the DSM (American Psychiatric Association, 2013) as "*fixed beliefs that are not amenable to change in light of conflicting evidence*". These inflexible beliefs can revolve around various themes such as grandiosity or somatic concerns. Among this repertoire, persecutory ideas occur in nearly 90% of first-episode psychosis patients (Freeman, 2007). Also referred to as paranoia, persecutory ideations represent the most common type of delusions associated with psychosis and correspond to the belief of being harassed or even harmed by an individual, a group, or an organization. These paranoid concerns occur at varying degrees throughout the general population and may serve an adaptive function related to the ability to detect potential social threats (Darwin et al., 2011; Freeman et al., 2005; Gilbert et al., 2005). Raihani and Bell (Raihani and Bell, 2019) notably found that in competitive situations,

the presence of coalitions and group coordination might elicit cognitive mechanisms aimed at detecting, anticipating and avoiding social threats.

Delusional and paranoid concerns have also been found associated with different categories of cognitive biases. We will preferentially focus on two of these altered forms of inference: motivated reasoning and jumping-to-conclusions.

Motivated reasoning is a cognitive bias that consists in evaluating information and evidence congruently with pre-existing beliefs or desired outcomes. Delusions have been consistently found associated with *confirmation biases*, defined as the tendency to preferentially consider evidence confirming a pre-existing belief (McLean et al., 2017) but also *disconfirmatory biases*, i.e., the tendency to disregard evidence that goes against the established belief (Garety et al., 1991). Similarly, delusions have also been found associated with *bias against confirmatory evidence* that corresponds to the tendency to discard new evidence that reinforces the true probability (McLean et al., 2017). This tendency for motivated reasoning could be conceptualized as a default in belief updating processes.

As previously described, *Jumping to Conclusions* (JTC) is a cognitive bias in which individuals tend to draw quick conclusions with great confidence based on limited evidence. It has been found associated with delusional and paranoid concerns (Dudley et al., 2016; Garety et al., 1991; Huq et al., 1988; McLean et al., 2017) meaning that paranoid individuals might require less evidence to believe that other people are hostile or that a situation is dangerous, contributing to a cycle of suspicious thinking.

Along these cognitive biases, altered forms of perceptual inference have repeatedly been found implicated in delusional beliefs. As we demonstrated earlier in this dissertation (see section "*A Bayesian approach of beliefs and perception*") new sensory information is combined with prior expectations to generate predictions (Helmholtz, 1866). The predictive coding framework posits that whenever these predictions are violated by sensory input, a prediction-error signal is triggered, leading to the updating of the internal model's predictions. Delusions have been shown associated with differences in the processing of these prediction-error signals (Corlett et al., 2007; Fletcher and Frith, 2009; Kapur, 2003).These imprecise predictions would render some sensory events overly surprising and salient, a phenomenon that can be referred to as aberrant salience attribution (Kapur, 2003) or hypersalience. Salience,

in this context, refers to the distinctiveness of a stimulus based on its physical properties or associated emotional and motivational factors. Aberrant salience attribution, also known as hypersalience, involves attributing inappropriate significance to neutral events, a phenomenon that varies in degrees across populations. The cognitive effort to make sense of such aberrant salience would results in the formation of delusional beliefs.

To fully understand this theory, let's come back to our initial bike-or-bus example (presented in the section "*A Bayesian approach of beliefs and perception*"). If you remember correctly, I was trying to predict whether it was going to rain or not based on meteorological forecasts broadcasted on TV and current states of the sky I could see through my window. This prediction was updated by newly available sensory information like clouds that appeared in the sky while I was sipping my coffee. Let's now imagine a single little cloud appearing in the sky. The new prediction I would compute in the light of this new piece of data wouldn't land too far from the first prediction I made. Comparing these predictions would then only trigger a small prediction-error. This would result in me considering the cloud as irrelevant to take in account in my decision-making process. Now, if the prediction-error triggered by this cloud is aberrantly amplified, its presence might pass as a salient and important information. The system would try to make sense of this cloud and might be biased in predicting a rainy weather. This system's tendency to detect and respond to salient stimuli, such as a sudden loud noise or a potential predator long conditioned survival. Drawing on this idea, hypersalience might have an adaptative value potentially enhancing threat detection. Aberrant salience attribution however has been found to play a crucial part in delusions (Holt et al., 2006; Kapur, 2003) and to also be associated with schizotypal traits (Chun et al., 2019; Haselgrove et al., 2016).

Key concepts

- Beliefs can be situated along a continuum, ranging from flexible and adapted, to rigid and pathological ones.
- Rigid beliefs can be found at various degrees across the schizotypal spectrum, such as paranoid concerns characterized by overconfidence and belief-updating failures.
- Delusions constitute a pathological form of inflexible beliefs, rooted in cognitive and perceptual inference biases.
- Among those biases, the attribution of great significance to random stimuli has been called hypersalience.

While the inferential biases associated with rigid beliefs in the schizotypal spectrum have been thoroughly investigated, less is known about the underlying mechanisms of other types of non-clinical beliefs. This thesis proposes to fill this gap focusing primarily on conspiracy-related beliefs, which were particularly thriving these past years. While these beliefs are usually not considered severe enough to require clinical intervention, they share common features with pathological constructs such as *paranoia* (Darwin et al., 2011) and have been proposed to be underlaid by similar cognitive biases constituting a good model for non-clinical rigid beliefs.

IV – Curiouser and curiouser: Debunking the adaptative value of conspiracy beliefs

It is now time to give a proper definition of what we consider as *Conspiracy Theories*. Adherence to conspiracy theories (CTs) rely on the general idea that a group of malevolent individuals are acting together in secret with hopes of achieving evil goals. Conspiracy ideations can be defined as a generalized tendency to believe in such conspiracy theories (Brotherton and Eser, 2015; Swami et al., 2011). I am not saying that such situations never happened (we can find several examples of political

conspiracies all along History), but we will see that CTs differ from conventional conspiracies in their lack of evidence and their resistance to discrediting (Prooijen and Lange, 2014). Furthermore, CTs often constitute a fertile ground for the rejection of authority, or official and scientific explanations. Adherence to CTs can simply be approached in terms of adherence to disparate conspiracy beliefs intersecting themes like aliens, viruses or new technologies, but in this thesis, I chose to rather approach CTs as a system of thoughts. Indeed, these disparate beliefs can be considered as first-order beliefs regarding the nature of the world, constituting what has been called by some authors, a conspiratorial mentality or mindset (see below).

Congruent with this idea, a first line of research evidenced that when an agent adheres to one conspiracy theory, she/he will be more prone to deem other conspiratorial hypothesis true, whether these conspiracy theories are shared by a wide group of individuals or invented by the experimenters (Drinkwater et al., 2012; Goertzel, 1994; Lewandowsky et al., 2013; Swami et al., 2011). To approach this phenomenon, Swami (Swami et al., 2011) used the notion of *conspiratorial mindset* which corresponds to a self-perpetuating system where one theory feeds another. In line with this conceptualization, it has also been proposed that an agent can firmly believe in several CTs contradicting one another. While the irrational nature of this paradox has been pointed out, we propose to nuance this allegation reminding that sometimes high levels of confidence can be granted to several contradictory theories at the same time by the same individual (Wood et al., 2012, e.g., *Princess Diana faked her death*, and *she was assassinated*).

Overall, this idea suggests that CTs constitute more a style of thinking questioning the "truth" provided by certain institutions, than a system of congruent and interconnected beliefs. Approaching that same idea through Bayesian formalism, we would then have several hypotheses to which a degree of plausibility is granted, together with a bias against the hypotheses validating the common worldview. Such a reasoning profile can be referred to as *conspiratorial mindset*, *conspiracy mentality*, predisposition to conspiracy ideations or other denominations generally measured as one's general degree of belief in CTs.

The scientific literature fosters a discussed, yet consistent, idea that CTs share common features with paranoia and schizotypy. While CTs are not considered pathological per se, these inflexible ideas can nevertheless crystallize around similar

background themes (e.g., mind-controlling new-technologies). Congruent with that assumption, CTs have been found associated with paranoid ideas and other paranoid characteristics such as distrust and fear of outside agents (Darwin et al., 2011). For Meller (Meller, 2002), paranoia and CTs can be both explained by a common mechanism of misjudgement of intentions and causalities. Evidence also suggests that CTs are associated with schizotypy and a wide range of related perceptual-cognitive factors (Barron et al., 2014; Bruder et al., 2013; Swami et al., 2013). CTs have notably been found correlated with delusional ideations (Dagnall et al., 2015) and inflexible beliefs present in non-clinical populations, such as magical ideations and paranormal beliefs (Darwin et al., 2011).

However, the often-controversial nature of CTs makes it difficult to dismiss them as mere "aberrant, false beliefs" for two main reasons. They still differ from paranoia on several points (Byford, 2011; Imhoff and Lamberty, 2017). First, they differ in their content. While paranoid concerns are self-referenced and focused on individual threats coming from everyone; conspiratorial thoughts revolve around the idea that the entire community is being targeted by those in power. Paranoid thoughts also appear to be personal and quickly implausible, while CTs revolve around more social themes and may appear credible. Finally, these two types of beliefs differ in their social values. While delusional individuals are most of the time hermetic to other odd beliefs, CTs are typically shared among a community and conspiracy believers can cluster together, even when they contradict each-other.

Another crucial difference is the adaptative value of CTs, which has been extensively discussed in the literature. In his recent paper, Bortolotti (Bortolotti, 2023) argues that, in a naturalistic approach, CTs cannot be considered as "*malfunctioning beliefs*" but rather as the output of biased cognitive processes. The subtle difference he makes would lie in the fact that a process can be biased and still optimal if the bias itself bypasses the tenets of probabilistic reasoning to fulfil an adaptative function. This author thus considers CTs as "*the imperfect response to psychological and epistemic needs that people experience when facing a significant event that calls for an explanation*". In other words, CTs would address the need to make sense out of threatening and distressing events along the need to gain a sense of agency.

According to Douglas (Douglas et al., 2017), conspiracy theories often speak to epistemic motives to reduce uncertainty by obtaining knowledge, existential motives

to feel safe and in control or social motives to hold one's self and one's groups in high regard. A similar conceptualization has been proposed by Wheeler (Wheeler, 2021) who stated that: "*Many people, because of a strong psychological need for complete explanations when unusual events happen, coupled with intense feelings of needing to belong to a social group, are prone to believing in conspiracies*". This approach highlights another substantial difference between CTs and paranoia. While paranoid beliefs and other delusions are associated with greater levels of distress and social isolation, CTs would not directly induce distress and could even reinforce the bonds between members of the social group sharing the same theories (Bortolotti et al., 2021; Douglas et al., 2017). CTs could also fulfil a larger social adaptative function by detecting potentially dangerous coalitions relying on a wide range of cognitive mechanisms including pattern, agency and alliance detection along threat management (van Prooijen and van Vugt, 2018).

Drawing on this idea, CTs have been proposed to serve as coping mechanisms for stress and loss-of-control when uncertainty increases sharply. In support of this theory, it has been proven that CTs are particularly sensitive to stress. In this sense, some studies highlighted an increase in CTs when anxiety levels rose (Grzesiak-Feldman, 2013; Swami et al., 2016). The COVID-19 pandemic was notably shown associated with a recrudescence in pre-psychotic/ psychotic experiences (Mengin et al., 2020) and conspiracy adherence (Georgiou et al., 2020). A longitudinal study also demonstrated that anxiety, uncertainty aversion and existential threat were associated with stronger conspiracy endorsement. However, an increase in CTs did not in return trigger a decrease in these variables. On the contrary, CTs strengthening could even predict an increase in these negative feelings and other CTs, suggesting a self-reinforcing mechanism (Liekefett et al., 2021).

Furthermore, facing major or recurring life stressors with uncertain outcomes can negatively impact the sense of personal control over one's life, or the unfolding of events. Drawing on this idea, *Compensatory Control Theory* (Kay et al., 2008) posits that CTs can serve as a coping mechanism when confronted with a subjective loss of control over the course of events (LoC). This substantial approach of CTs inspired the second experimental work presented in this thesis. Congruent with this theory, loss of perceived control in times of uncertainty was found associated with changes in beliefs and attitudes (Bukowski et al., 2017; Peluso and Pichierri, 2021; Thompson et al.,

1993; Zhu et al., 2020). More specifically, it has been suggested that, in the event of a global social crisis, blaming an external group could constitute a coping strategy for re-establishing a sense of personal control (Bukowski et al., 2017; Sullivan et al., 2010).

However, the investigation of the CTs-LoC association has yielded mixed results. While it has been validating using various experimental paradigms (Whitson and Galinsky, 2008), notably when control in the political arena is threatened (Kofta et al., 2020; Pantazi et al., 2022; Stojanov et al., 2022; van Prooijen and Acker, 2015), attempts at replicating these findings have not always been successful (Hart and Graether, 2018; Nyhan and Zeitzoff, 2018; Stojanov et al., 2020; van Elk and Lodder, 2018). These literature discrepancies might be explained by hidden variables that can influence both LoC and CTs (Stojanov et al., 2020), such as stress (Swami et al., 2016) and uncertainty (van Prooijen and Jostmann, 2013). We argue that the methodologies used to induce LoC in these research papers did not always induce a threat level high enough to trigger CTs.

Inspired by this idea, considerable efforts have been made to naturally assess the impact of real-world events fraught with uncertainty — such as political elections, natural disasters, or the COVID-19 pandemic — on LoC and simultaneous CT adherence (Šrol et al., 2021; Stojanov et al., 2022). However, experimental designs aimed at elucidating the role of uncertainty in this LoC-CTs association stay very rare (Dow et al., 2022). Furthermore, the widespread use of laboratory procedures provides only limited information on the actual real-world validity of the relationship between control threat and conspiracy ideations (van Prooijen and Acker, 2015), encouraging future studies to adopt an ecological approach to address the central claims of the compensatory control model (Stojanov et al., 2022).

In the light of those considerations, my thesis work aimed at better understanding the associations between CTs and the inferential mechanisms involved in uncertainty processing and how they specifically relate to the distress and LoC induced by real-world uncertainty. Furthermore, despite their adaptative values, I would like to remind that CTs stay some alarming phenomena because of their rigid component. As mentioned earlier, beliefs widely shape attitudes and subsequent behaviors. CTs can thus drastically influence individual and collective decisions related to important issues, such as health care (Bertin et al., 2020; Bird and Bogart, 2005; Jolley and Douglas, 2014; Marinthe et al., 2020; van Mulukom et al., 2022) or climate

change (Uscinski and Olivella, 2017). CTs can also have an impact on the way people perceive contemporary and historical world events (Swami et al., 2010), with again some important consequences in terms of socio-political engagement (Butler et al., 1995; Imhoff et al., 2021).

It has been suggested that CTs could produce a generalized political attitude associated with intentions to challenge status quo and a pejorative view of the elites, perceived as less likeable and more threatening than low-power groups (Imhoff and Bruder, 2014). While CTs can strengthen the bonds among the members of a well-defined group sharing the same views, they can also lead to social isolation through stigmatization. Withdrawal of social engagement can be a direct consequence of CTs, for instance in the case of rejection of COVID screening tests or vaccination. Congruent with this idea, CTs have been shown associated with avoidant coping styles, a maladaptive stress-management type that can lead to temporary disengagement and abandonment of goal-related behaviors (Marchlewska et al., 2021), or difficulties in regulating one's emotions (Molenda et al., 2023). In the context of the COVID-19 pandemic, some authors also demonstrated that an immature defense style, defined as the tendency to use fantasy as a substitute for human relationships or problem solving, was a good predictors of CTs (Gioia et al., 2023).

Key concepts

- CTs are rigid beliefs that share some features with paranoia.
- CTs have been conceptualized as compensatory coping mechanisms designed to face uncertainty and a perceived lack of control.
- Despite their adaptive values, CTs are overly rigid and could have a negative impact on individual and collective decisions.

Given their potential maladaptive values, a substantial amount of papers tried to elucidate CTs determinants. In the next section we will briefly review some of these findings preferentially focusing on the probabilistic biases found associated with CTs.

V – Through the looking-glass: perceptive-cognitive features of rigid beliefs

First, different thinking styles were found associated with CTs. Irwin and Young (Irwin and Young, 2002) differentiate: (a) *analytical/rational processing*, referred to as reality-testing, which is slow, conscious, considered, and nuanced, from (b) *intuitive/experiential processing*, which is fast, pre-conscious, holistic, and spontaneous. CTs appear more associated with intuitive processing (Drinkwater et al., 2012; Georgiou et al., 2021; Pytlik et al., 2020) and reduced critical thinking abilities (Lantian et al., 2021). This could be explained by reality-testing deficits that have been evidenced in conspiracy believers (Lewandowsky et al., 2013).

Interestingly, the different types of cognitive biases we highlighted in delusions (see "*Down the rabbit-hole: aberrant inference in the schizotypal spectrum*" section) have also been found associated with CTs (for a full review, see Gagliardi, 2022). Conspiracy believers are thought to bias the weight they attribute to certain stimuli to avoid uncertainty or to conform to pre-existing world-views, a tendency we previously referred to as *motivated reasoning* (Wycha, 2015). This tendency is illustrated by the association between CTs and confirmation biases (Kuhn et al., 2022; McHoskey, 1995) and disconfirmatory biases (Georgiou et al., 2021; Woodward et al., 2007). Congruent with this idea that conspiracy believers might seek to reduce uncertainty, CTs have finally been found associated with hastier or statistically inaccurate decision making based on slim probabilities, a phenomenon we previously referred to as *jumping-to-conclusion* (JTC) also known as *inferential confusion* (Kabengele et al., 2023; Pytlik et al., 2020). This tendency to formulated biased probabilistic judgement is further illustrated by the association between CTs and conjunction fallacy, defined as the assumption that a combination of two or more specific conclusions is more probable than any one of those conclusions (Brotherton and French, 2014; Dagnall et al., 2017). This bias is also known as the "*Linda problem*" and was assessed by providing

participants a description of a lady named Linda. The conjunction fallacy occurs when participants mistakenly believe that a specific combination of events (such as Linda being a bank teller and a feminist activist) was more likely than one of the individual events (Linda being just a bank teller). This first assumption violates the laws of probability because the more specific statement can never be more probable than the broader one. The authors explain that this bias could illustrate a tendency of CTs believers to infer underlying causal relationships between ostensibly unrelated events.

This proneness to detect causal relationships between unrelated stimuli can also be illustrated by a perceptual bias, *illusory pattern detection* (IPD). Drawing on that idea, the research papers that investigated the potential CTs-IPD association have yielded accumulating, albeit conflicting results. Pattern detection can be defined as the ability to make sense of our environment by identifying significant relationships among stimuli. This capacity to discriminate and internalize associations is, from an evolutionary perspective, central to our survival in terms of detection of potential sources of danger (Beck and Forstmeier, 2007; Mattson, 2014; Shermer, 2011). However, this process can be biased, leading to the perception and internalization of associations between random events coming from the environment (Beck and Forstmeier, 2007; Foster and Kokko, 2009; Shermer, 2011). In fact, there is a natural tendency in the general population to detect patterns in random stimuli (Rieth et al., 2011).

Interestingly, IPD propensity was proven associated with various rigid beliefs (Blackmore and Moore, 1994; van Harreveld et al., 2014; Wiseman and Watt, 2006), including paranoid ideas (Brennan and Hemsley, 1984), as well as conspiracy theories and supernatural beliefs (van Prooijen et al., 2018). Some authors however found contradicting results refuting the link between IPD and adherence to CTs (Dieguez et al., 2015).

Van Prooijen and colleagues (van Prooijen and van Vugt, 2018) posit that these results could be explained by "*sampling differences*", a disharmony in the socio-demographic features of the samples recruited in these different protocols. The correlation between IPD and paranormal beliefs appears more robust, with individuals more inclined to detect faces in noisy images (Krummenacher et al., 2010; Riekki et al., 2013) or to perceive patterns in ambiguous stimuli (Blackmore and Moore, 1994; Brugger et al., 1993; Gianotti et al., 2001). A recent line of research grounded on strong

methodological validity provided new evidence to feed the debate linking paranormal beliefs and COVID-19 conspiracy beliefs with IPD, through *signal detection theory* (Hartmann and Müller, 2023; Müller and Hartmann, 2023).

According to Beck and Forstmeier (Beck and Forstmeier, 2007), this can be explained by the less significant cost of a false alarm in terms of survival (e.g., associating rustling grass with the approach of a predator rather than the wind causing it) than the opposite error (e.g., failing to flee when a predator is indeed approaching). However, there are significant inter-individual variations in this propensity for *illusory pattern detection* (Gosselin and Schyns, 2003). I would like to highlight the similarity between the tendency to detect illusory patterns and hypersalience that inspired the first study I will be presenting in this thesis. Both mechanisms rely on aberrant perceptual processing, play an important role in threat detection, and have been associated with aberrant beliefs. Surprisingly however, to our knowledge no study investigated the potential association between hypersalience and CTs, a gap I will address in the first chapter of this dissertation.

While this evolutionary account provides a nice explanation for the emergence of IPD, it does not however explain the amount of variations in individual susceptibility to this phenomenon. IPD would actually have a second function that inspired the second study presented in this thesis. As we explained earlier in this section, facing major life stressors could render difficult the maintenance of a sense of control over events, at the origin of significant distress. Establishing connections between different elements from the environment, whether illusory or not, could actually help restoring a sense of predictability and therefore a sense of control over the external world (Hogg et al., 2010; Whitson and Galinsky, 2008). Supporting this view, some studies highlighted a link between perceived LoC and IPD (van Harreveld et al., 2014; Whitson and Galinsky, 2008). However, these results were also subject to debates in the literature, as some authors (van Elk and Lodder, 2018) refuted the possible link between perceived lack of control and IPD. Again, we posit that the memory-recall method used to induce LoC might not threaten the overall sense of control sufficiently to trigger compensatory mechanisms such as IPD.

Similarly to the hypersalience mechanism we described in delusions (see section "*Down the rabbit-hole: aberrant inference in the schizotypal spectrum*"), adherence to irrational beliefs would constitute a compensatory strategy aimed at

validating IPD to alleviate the aversive sensation of LoC (Walker et al., 2019). In support of this assertion, the literature established a link between LoC and adherence to supernatural beliefs (Kay et al., 2010; Laurin et al., 2008) but also CTs (Sullivan et al., 2010; Whitson and Galinsky, 2008), notably when control in the political arena is threatened (Kofta et al., 2020; Pantazi et al., 2022; Stojanov et al., 2022; van Prooijen and Acker, 2015). Attempts at replicating these findings however have yielded negative result (Hart and Graether, 2018; Nyhan and Zeitzoff, 2018; Stojanov et al., 2020; van Elk and Lodder, 2018). The LoC-CTs association is likely driven by several common co-variables (Stojanov et al., 2020), including stress (Swami et al., 2016) and uncertainty (van Prooijen and Jostmann, 2013), that where not properly accounted for in these studies, and which might partially explain these discrepancies. The second study presented in this thesis aims to remedy these inconsistencies by assessing the impact of individual stress levels on the LoC-CTs association in a context of real-world uncertainty.

Key concepts

- CTs are associated with biased forms of probabilistic inference such as jumping-to-conclusion and illusory pattern detection (IPD).
- IPD and hypersalience are both related to aberrant perceptual processing and play a role in threat detection. However, no study explored the association between hypersalience and CTs.
- IPD may help restoring a sense of predictability and control in a chaotic world.
- CTs could be a compensatory strategy to validate IPD and thus alleviate the aversive sensation of lack of control (LoC).
- The LoC-CTs association remains however subject to debate and might be influenced by hidden covariates, such as stress and uncertainty.

Different types of probabilistic tasks can be used to investigate the inference biases involved in the emergence of rigid beliefs situated across the normal-to-

pathology continuum. Delving on a deeper level of comprehension, coupling mathematical models to behavioral data collected through these dedicated paradigms could help refine and expand our understanding of the mechanisms underpinning inference biases involved in CTs. To conclude this introduction, I will (1) highlight how computational models have been used to fit data from probabilistic tasks to approach paranoid and conspiratorial ideations, and (2) defend why the *Circular Inference* model might constitute a tool of choice to investigate non-clinical rigid beliefs such as CTs in this context.

VII – When *computational* models strike back

According to the *Computational Theory of Mind*, the probabilistic processing of information underlying cognitive functions can be approached with mathematical models. These models correspond to Marr's second level of analysis mentioned in the first section of this introduction (i.e., the algorithmic level), and aim at simplifying complex neurophysiological processes by reducing them to a sequence of mental operations, similar to logical-mathematical operations, called *computations* (Horst, 2005). The role of such computational models is to provide insight into the sometimes-unclear behavioral data provided by probabilistic tasks. For example, the use of a beads task could highlight JTC phenomena in two different populations but leave aside the potential differences in the underlying mechanisms leading to this bias in each of these groups (e.g., a change in decision threshold versus an increased weight attributed to sensory inputs). Computational models aim at elucidating and quantifying such subtle mechanisms. To do so, a model is designed to compute mathematical values, or parameters for each participant, based on their behavior. These parameters vary depending on the structure of the model. Each of them however reflects a different dimension of Bayesian reasoning, such as the confidence granted in the sensory evidence or the prior belief that the world is bound to change. The goal of the model is to compute for each participant the parameters that best fit behavioral data, in order to characterize how they process information.

Interestingly, computational modeling has already been used to understand phenomenon sharing common features with CTs such as schizotypy, persecutory beliefs and paranoia. Fromm and colleagues (Fromm et al., 2023) took an interest in individuals with psychotic-like experiences, defined as unusual subjective experiences resembling subtle psychotic symptoms without necessarily causing distress. These experiences, that can be conceptually assimilated to schizotypal traits, were associated with altered dynamics of belief updating, specifically slower learning after large prediction-errors. This tendency to disregard environmental changes, would limit the capacity to establish new beliefs in the face of contradictory evidence leading to the emergence of rigid beliefs.

Following on from this literature, an emerging line of research specifically focused on the associations between paranoia, belief-updating and uncertainty processing through a computational lens. Reed and collaborators (Reed et al., 2020) notably addressed this question using reversal-learning tasks in large non-clinical samples, patients with schizophrenia, but also in animal models, using rats exposed to methamphetamine, a psychotomimetic drug. They evidenced a common uncertainty-driven belief-updating mechanism in each of these populations, showing that paranoia was associated with strong priors for high environment volatility. The same team replicated this effect in a real-world setting and found that the successive lock-downs and re-opening during the COVID-19 pandemic were associated with an increase in paranoia and had an impact on belief-updating (Suthaharan et al., 2021). An association between belief-updating and paranoia was finally replicated in a non-clinical sample as well as in patients with schizophrenia (Sheffield et al., 2022).

In a series of papers, Barnby and collaborators (Barnby et al., 2020, 2022) adopted a complementary approach by using social and non-social versions of a probabilistic reversal-learning task in non-clinical participants. They showed that, paranoia was associated with greater uncertainty regarding others' actions and reduced belief-updating regarding harmful intent attributions. These authors posit that when the probability of being targeted by harmful intents increases, it might elicit a cognitive mechanism designed to cope with the uncertainty induced by the situation. In other words, paranoia may arise from selective pressures to infer and avoid social threats, particularly in ambiguous or changing circumstances. In an attempt

to efficiently detect and avoid potential threats, participants exhibiting the higher levels of paranoia, excessively relied on prior beliefs that other's intentions were malevolent.

Even if evidenced for paranoia, the substantial association between CTs and inferential biases immediately raises the question of the Bayesian inference processes underlying these specific beliefs. While attempts to couple computational approaches with behavioral data to explain the underpinning mechanisms of CTs are convincing, they remain very limited. Suthaharan and colleagues (Suthaharan et al., 2021) notably implemented an online protocol during the COVID-19 crisis. Coupling a reversal-learning task with computational modeling, they showed that during this sensitive period of time, CTs were associated with heightened prior expectations for volatility. Another recent research paper (Zhang et al., 2022) proposed that, CTs would be associated with IPD, conceived as a biased form of adaptative learning resulting in causal relationships between random stimuli such as persons and events. These authors suggest that the performance in the reversal-learning task would directly relate to the ability of accurately perceiving the statistical structure of the environment and to adjust decisions accordingly. Coupling an online version of a reversal-learning task (**Figure 6**) with a computational model, they showed that CTs are associated with deficits in the recognition of contingencies underlying the outcomes (inaccurate pattern detection), together with a reduced ability to adapt to changes in these contingencies (associative learning in conditions of heightened volatility).

Figure 6 : A web-based reversal-learning task (Zhang et al., 2022). In this probability-reversal learning task, participants start with 10,000 "gold coins" and their goal is to lose as few gold coins as possible. During the experiment, participants are repeatedly asked to choose between a blue and a red leprechaun. Each of them has a different probability of stealing gold coins from them and holds a bag with a number indicating the potential coin loss. Participants are informed that the likelihood of encountering a thieving blue or red leprechaun depends on recent outcomes. In the stable block, the blue leprechaun has a 75% chance of stealing, while the other leprechaun has a 25% chance. In the volatile block, the stealing probabilities of the leprechauns switch every 20 trials.

Using a different approach, Rigoli (Rigoli, 2022) attempted to challenge the adaptative value of CTs introducing the notion of "*expected consequences*". In their theoretical model, motivated reasoning played a key role in the probabilistic processes underlying belief formation. An accurate hypothesis underpinning a series of events might be rejected in favor of another one if it is too likely to threaten the established belief system. Bolstering this idea, some authors argued that CTs could actually be the result of inference processing that, embedded in a broader conspiracist beliefs system, would not just be "sub-optimal", but rather remarkably efficient at protecting the core belief system itself against disconfirmatory evidence (Poth and Dolega, 2023). In a recent paper, Sato (Sato, 2023) supported this assumption. Referring to the *Free Energy Principle Theory*, he posits that, in order to efficiently adapt to an ever-changing environment, the cognitive processes underpinning beliefs need to reduce surprise (called here free-energy). To accomplish this predictive optimality, CTs would be prioritized depending on the psycho-social environment. This author also argues that, for individuals historically exposed to systemic mistreatment by the authorities, likely to raise negative expectations about the establishment, the risk of adopting an inaccurate conspiracy belief about authorities would not outweigh the induced benefits. This appears congruent with the common idea that discriminated populations subscribe more to CTs (Gkinopoulos and Mari, 2023; Nera et al., 2022).

Beyond the individual, a relatively recent axis of research also examined how information-sharing algorithms on web social-media may account for extreme beliefs propagation across large-scale networks and sub-networks. This part of the literature was particularly interested in phenomena like polarization and radicalization, which dramatically resurfaced in online communities. Polarization refers to a situation in which the distribution of opinions is characterized by two well-separated peaks around the neutral consensus (Baumann et al., 2020). In other words, it corresponds to the emergence of two opposite beliefs regarding divisive issues with no clear answer. Radicalization on the other hand can be conceptualized as an overconfidence in one or the other of these opposing views.

Interestingly, Bayesian models have been proposed to account for polarization around climate change related beliefs (Cook and Lewandowsky, 2016). Evidence suggests that CTs could be the product of rational Bayesian reasoning processes embedded in some sub-optimal constraints such as partial access to the total

information shared across the network (Madsen et al., 2017). The implication of this limited or biased access to information is particularly striking in the context of echo-chambers. *Echo-chambers* can be defined as enclosed spaces where people are only confronted to a given type of belief or opinion, congruent with their own world-views. These beliefs are reverberated within the same, limited circle, inducing a self-reinforcement mechanism. The term mediatic echo-chamber refers to a strategy where the same information is repeated on a different form or across different sources to falsely increase its credibility in the public. While biased exposure to confronting opinions is inherent to the constraints of our social environment, echo chambers have drastically multiplied with the development and popularity of numerical social networks. The substantial influence of such biased social interactions on information processing has recently been investigated by Baumann and colleagues (Baumann et al., 2020). Using a dynamical model of belief propagation, the authors highlighted a mechanism by which agents sharing similar opinions can mutually reinforce each other triggering radicalization of opinions.

Some authors tried to explain how echo-chambers could emerge and maintain themselves in online spaces despite the availability of contradictory information and opinions. Some controversial figures of the public scene, such as Eli Pariser, blames the extreme personalization operated by the algorithms that results in what he calls "*filter bubbles*" where the individuals would only be oriented towards content that reinforce their own ideological and political convictions (Pariser, 2011). Everyone is aware today that GAFAM (Google, Apple, Facebook, Amazon & Microsoft) collect a large amount of personal data from the internet (e.g., from our navigation history) to infer our needs and envies, but also our opinions, and propose us a selective content supposed in line with our interest. One direct consequence of this marketing process could be "*intellectual isolation*". Scientific research indicates that while these algorithms (as well as cognitive and social biases) could play a part in belief strengthening, the size of the network could structurally explain the emergence of echo-chambers, even when the agents operate optimal Bayesian reasoning (Madsen et al., 2018). It has also been suggested that the dynamic component of beliefs propagation across these networks could be a significant factor to consider when accounting for echo-chambers and polarization (Pilditch et al., 2022).

Key concepts

- Computational models offer an ideal framework to refine data-interpretation in behavioral and cognitive science.
- Bayesian models provide evidence for the implication of biased inferential mechanisms in the emergence of rigid beliefs across the schizotypal spectrum.
- Beyond the individual, the Bayesian framework can also account for polarization and radicalization phenomena observed in online communities.
- The use of computational models to understand CTs constitute an emerging research axis.

All of these models provided valid accounts for the cognitive-perceptual inference mechanisms underpinning rigid beliefs. This thesis proposes to build on an alternative approach, i.e., the *Circular Inference* model, that has also proven effective in accounting for rigid beliefs and altered perceptions across the psychosis spectrum. I will now end this introduction by a theoretical presentation of this model, followed by a brief review of the experimental findings it yielded until then. For convenience, the main features of CTs we presented across the previous sections are summarized in **Table 1**, p.59.

VIII – The *Circular Inference* Model

Among the different Bayesian approaches that were proposed to account for rigid beliefs, the *Circular Inference Model* (CI) is of particular interest for us since it was shown able to capture positive symptoms in schizophrenia, i.e., delusions and hallucinations (Jardri and Denève, 2013). At the physiological level, this model posits that subtle imbalances in the neuronal excitatory / inhibitory ratio are responsible for an altered form of causal inference, named "*circular inference*". This theory roots in the idea that the control of the balance between excitatory and inhibitory synaptic inputs generated in response to external stimuli is essential to generate stable mental representations. Indeed, brain activity is subjected to the constant analysis of information flow coming from the senses (but also from internal sources), that can be

more or less relevant to the individual. All this information must be sorted according to its importance in order to be processed optimally by the brain. In this way, some non-relevant information in a given context will be treated as " noise" that can be inhibited, thereby freeing up synapses to propagate this information in the cortical hierarchy. This type of filtering is directly related to the concept of salience we mentioned earlier (see "*Down the rabbit-hole: aberrant inference in the schizotypal spectrum*" section). For example, specific neurons in the visual cortex might be highly selective for bright colours (Shapley and Hawken, 2002) or contrasts (Reynolds and Desimone, 2003), helping the brain to prioritize and respond to these salient features in the environment. Similarly, neurons of the posterior parietal cortex would encode the location of salient stimuli (Constantinidis and Steinmetz, 2001, 2005). This process is furthermore mirrored at the cognitive level. The anterior cingulate cortex has notably been shown implicated in the processing of unexpected threat-related stimuli (Bishop et al., 2004).

To carry out this filtering, called here *neural selectivity*, a subtle interplay of balancing between excitation and inhibition of information occurs. The inhibitory action of gamma-aminobutyric acid (GABA) is crucial in this neuronal mechanism since it prevents synapses from prolonged excitation, counteracting the excitatory effects of Glutamate. Crucially, some imbalances in the GABA/Glutamate equilibrium were demonstrated in several neurological disorders, including epilepsy and cerebral ischemia. Similarly, the possibility of a global dysfunction in neurotransmission mechanisms (dopaminergic, glutamatergic, and GABAergic) to account for the symptoms of schizophrenia has been seriously considered in recent years, and a significant amount of data supports the idea of an excitatory / inhibitory imbalance in this disorder (Brennan and Hemsley, 1984; Coyle, 2006; Hugdahl et al., 2015; Pilowsky et al., 2006; Volk and Lewis, 2002). It is common to observe a decrease in inhibition or an aberrant increase in synaptic excitation within this population at the molecular, genetic and physiological levels, as well as in post-mortem tissues. This global malfunction would alter the excitatory / inhibitory balance, a system that contributes, among other things, to the maintenance of stable perceptual representations (Carcea and Froemke, 2013).

Jardri and Denève proposed a theory that explains how these perturbations in the bE/I at the neuronal level could produce behavioral manifestations similar to those described in schizophrenia (Denève and Jardri, 2016; Jardri and Denève, 2013).

Specifically, they suggest that these perturbations can give rise to a pathological form of inference known as *Circular Inference* (CI). In CI, ascending and descending sensory information, inadequately inhibited, becomes amplified and processed repeatedly by neurons, as though this information appears entirely new rather than simply reverberated within the neural circuitry. In silico simulations have shown that an imbalance in bE/I results in abnormal redundancy in the processing of information, leading to the formation of inferential loops. According to the circular inference model, psychotic symptoms would constitute cognitive and perceptual manifestations of a poorly regulated propagation of the probabilistic messages implemented at the neuronal level.

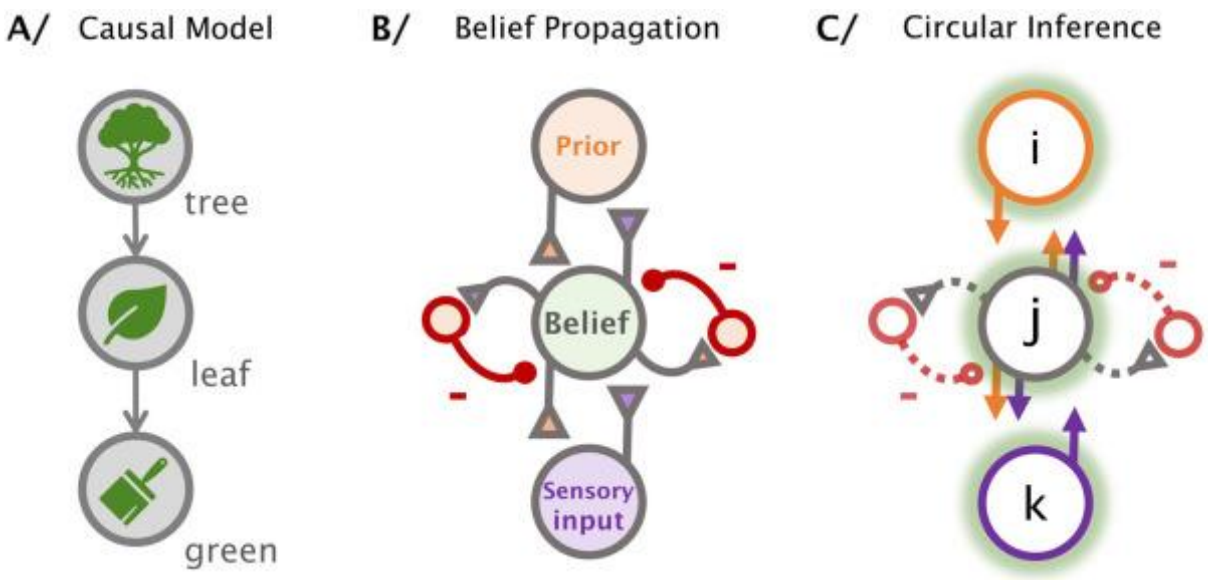


Figure 7 : Principles of belief propagation and circular inference, from (Bouttier et al., 2022). (A) In this example of inference processing, the goal of the system is to determine the probability of perceiving leaves in the environment combining prior belief of walking under trees and some sensory evidence for the color green. The sensory input for the color green is given at the bottom of the hierarchy, while the prior expectation of a tree is given at the top. **(B)** Representation of a possible implementation of this inference processing at the neuronal level. Prior expectations and sensory input information are propagated through the cortical hierarchy. This flow of information is controlled by inhibitory interneurons which remove redundant information from the messages. **(C)** In case of an inhibitory deficit, messages in the network are being reverberated uncontrollably, causing loops of information. This phenomenon referred to as *Circular Inference* can lead to belief rigidification.

The CI framework offers two substantial strengths over alternative models. First, it encompasses a "*domain general*" inference mechanism, validated in the cognitive domain, through probabilistic decision-making tasks (Derome et al., 2023; Jardri et al., 2017), but also in the perceptual domain, thanks to bistable tasks (Leptourgos, Notredame, et al., 2020). It thus provides a unified framework that can commonly account for rigid beliefs and altered perceptions in schizophrenia (i.e., delusions plus hallucinations). Second, it does not rely on an all-or-nothing approach; instead, it allows for the consideration and quantification of the strength of the information processing imbalance that would cause inference biases. Furthermore, it allows to qualitatively decipher the mechanisms underlying such imbalance (as reflected by the different parameters of the model). As a result, CI can not only account for varying levels of biased inference present across the normal-to-pathology continuum, but also distinguish between different inference profiles associated with various types of beliefs, effectively addressing the polymorphic nature of non-clinical inflexible beliefs.

This model was experimentally validated by Jardri and colleagues (Jardri et al., 2017), seminally using a decision-making task in an ambiguous situation, known as the fisherman task (**Figure 8**), which was derived from the classic probabilistic beads task, presented earlier. This behavioral task was designed to quantify the respective importance given to sensory and prior information during beliefs formation. Each trial began with the presentation of an a priori information: participants were shown two fisherman's baskets that were said to come from two different lakes. This information was then removed and had to be held in working memory. Subsequently, full sensory information was provided by presenting the fish distributions within the two lakes (in an inverse ratio of red and black fishes). Simultaneously, a supposedly “caught” red fish was presented below the lakes. Participants were asked to indicate their level of confidence using a semi-circular scale regarding the origin of the fish (i.e., whether the fish came from the right lake or the left lake). This variant of the beads task had the advantage of giving access to a direct measure of conviction when the participants had to reach a decision.

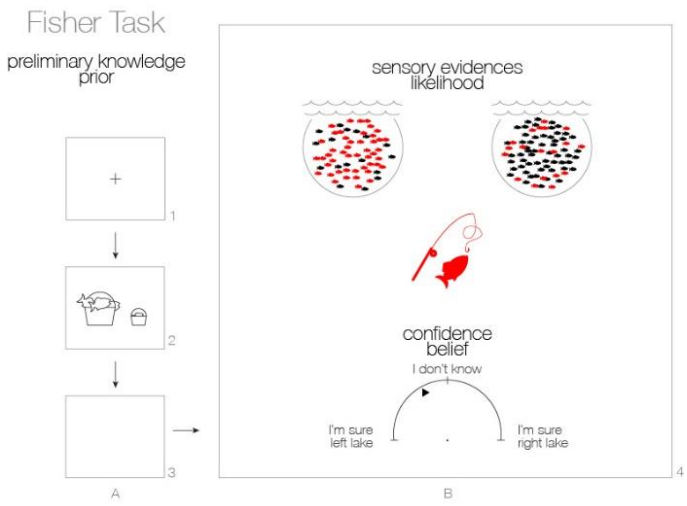


Figure 8 : The fisherman task (Jardri et al., 2017). In this experiment, participants are being presented with two lakes, each filled with black and red fishes. Participants are informed that a fisherman caught a fish in one of the two lakes and their task is to guess which one combining two information: (i) baskets of varying size representing the chance that the fisher caught a fish from the left or right lake (prior information) and (ii) the proportions of black and red fishes within both the left and right lakes (sensory information). Each trial can be decomposed into four steps. After a fixation cross was presented (1), the prior information is provided represented by the baskets (2). This prior is then removed (3), and the likelihood information was provided represented by the lakes (4). The participants are then asked to report their confidence that a red fish originated from one of the two lakes using a semi-circular scale.

In this experiment, individuals with schizophrenia tend to make hasty decisions based on less evidence and assign greater credibility to their decisions (a phenomenon we previously referred to as JTC bias). By applying the CI model to these results, the team showed that only a significant number of upward inferential loops caused by inhibitory deficits accurately reproduced the behavioral pattern observed in patients. A significant link was also found between this model parameter value and the severity of positive symptoms in the participants, as well as between the degree of inhibition impairment and the rigidity of beliefs. Recently, the same paradigm was administered to a sample of participants with low- and high-levels of schizotypy (Derome et al., 2023). Congruent with previous findings, individuals with strong schizotypal traits exhibited more confidence in their decisions. Fitting the model to the data, they showed that high schizotypy was associated with aberrant inference processing, where the sensory evidence was over-counted, while priors were under-weighted. These findings strengthen the validity of the CI model by accounting for non-clinical inference biases, observable at varying degrees across the schizotypal spectrum.

A second line of research aimed at validating the CI framework in the domain of visual perception, both in healthy controls and patients with psychosis. In a first study from Leptourgos and colleagues (Leptourgos, Notredame, et al., 2020), the authors used a modified version of the Necker Cube (NC) task. They manipulated prior expectations about the cube providing forged assumptions to the participants such as "Most people see the cube from above" but also the level of sensory evidence available to resolve uncertainty using contrast cues (**Figure 9**). The goal of the experiment was to decipher and quantify how participants use both information to form perceptual inferences. As expected, these manipulations significantly affected the way healthy participants perceived the NC which provides additional evidence advocating for a Bayesian account of visual processing. This paradigm is currently administrated to patients with schizophrenia and preliminary results suggest a sub-optimal integration of visual inputs in this population compared to healthy control (Leptourgos et al., *in prep*).

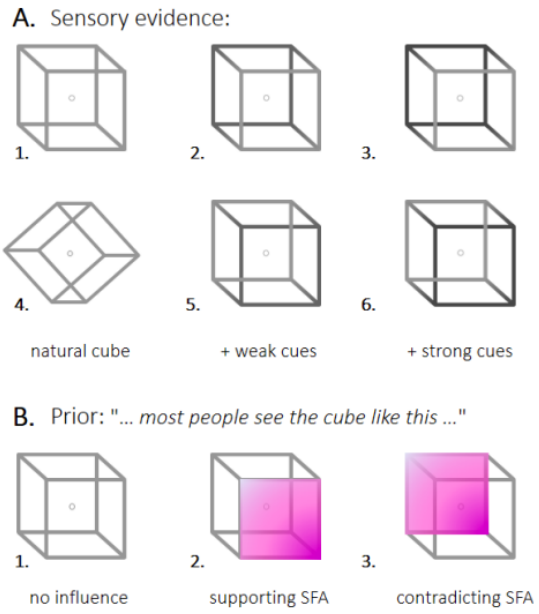


Figure 9 : The Necker Cube task (Leptourgos, Notredame, et al., 2020). The experimental procedure consisted in manipulating sensory evidence and prior information to bias perceptions of a Necker Cube (NC). **(A)** The NC could be presented as completely ambiguous **(1)**, or sensory evidence could be manipulated by adding visual cues in the form of contrasts or by tilting the cube **(2-3 and 5-6)**. The contrast was strong (3 and 6) or weak (2 and 5) and could support (2 and 3) or contradict (5 and 6) the implicit prior. Additionally, since people have an implicit preference for the cube seen from above (SFA interpretation), this preference was interpreted as an implicit prior. Such an implicit prior can be refuted by tilting the stimulus (4). **(B)** Prior was also manipulated by providing correct **(2)**, wrong **(3)** or no information **(1)** to the participants about the implicit prior either supporting or contradicting it.

Interestingly, the theoretical work conducted by the team in 2017 (Leptourgos et al., 2017) suggests that a slight redundancy in the circulation of prior information at the neuronal level would be necessary to the establishment of bistability and, more broadly, of a stable perceptual system (**Figure 10**). These findings indicate that a small amount of circularity would constitute a general mechanism underlying perceptive processing in healthy individuals. Elaborating further on these results, we started to wonder whether the strength of the inhibitory imbalance and subsequent message-passing redundancy could partially account for the normal-to-pathology continuum of beliefs. We thus theorized an inference processing mechanism where a minimum of loops is required to stabilize the system, but where, on the contrary, too many loops would generate instability and sub-optimal functioning of the perceptual system. Generalizing this perceptual functioning to cognitive inference would correspond to the establishment of coherent and flexible beliefs on the one hand, and rigid beliefs and perceptual anomalies observable on the schizotypal spectrum on the other hand. These theoretical considerations inspired the experimental work presented in the third chapter of the present thesis. Extending previous work from the team, we tested whether the CI model applied to perceptual behavior in non-clinical populations could account for rigid beliefs such as CTs.

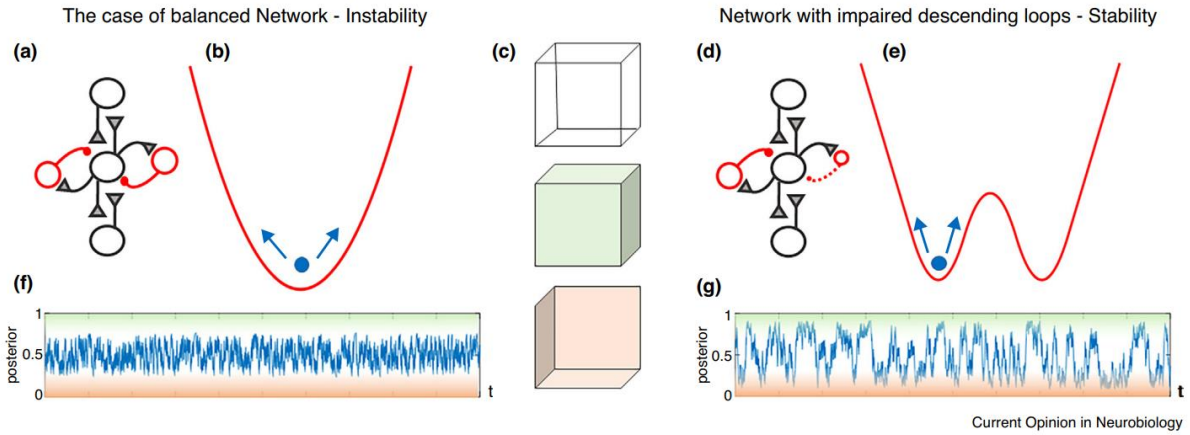


Figure 10 : Circular inference and the dynamics of bistable perception (Leptourgos et al., 2017). Representation of the perception of the Necker Cube (NC) over time **(c)** in the case of two possible neural implementation of belief-propagation networks: a balanced network without loops **(a)**, and a network with a small amount of descending loops **(d)**. NC Perception in a balanced network alternates between the two possible interpretations at a pace that appears unable to generate strong and persistent representations **(f)**. By contrast, the overcounting of priors in a network with descending loops lead to an increase in the strength and persistence of percepts able to generate bistability **(g)**.

Drawing on the idea that a message-passing redundancy at the neuronal level could elicit biased inference resulting in overconfidence, this theory was further extended at the level of social networks. We wondered whether a similar redundancy in the propagation of information across large-scale networks could account for the radicalization phenomenon increasingly reported among online communities. The last chapter of this dissertation aims at exploring this hypothesis.

Key concepts

- Circular Inference (CI) is a computational model that theorizes how neuronal excitatory and inhibitory imbalances may lead to cognitive and perceptual manifestations of schizophrenia.
- CI offers a unified framework to understand hallucinations and delusions, but also to quantify the strength of underlying information processing imbalances.
- The CI model can be used to fit decision-making processes and elementary perceptual inference equally.
- CI was seminally and experimentally validated in psychosis, but could constitute a well-suited framework to investigate non-clinical beliefs as well.

Table 1 : **Summary of the main cognitive and perceptual features discussed in the present thesis.**

Conspiracy Theory features	Supported by
Cognitive features (for a review, see Gagliardi, 2022)	
Low analytical/critical thinking	Swami et al., 2014
	Lantian et al., 2017
Reality-testing deficits	Lewandowsky et al., 2013
Intuitive processing	Drinkwater et al., 2012
	Georgiou et al., 2021
	Pytlik et al., 2020
Motivated reasoning	Wycha 2015
Confirmation bias	Kuhn et al., 2022
	McHoskey, 1995
Disconfirmatory bias	Georgiou et al., 2021
	Woodward et al., 2007
Jumping-to-conclusion	Kabengele et al., 2023
	Pytlik et al., 2020
Conjunction fallacy	Brotherton and French, 2014
	Dagnall et al., 2017
Perceptual features	
Illusory pattern detection	van Prooijen et al., 2018
	Hartmann and Müller, 2023
	Müller and Hartmann, 2023; *nuanced by:* Dieguez et al., 2015
Computational features	
heightened prior expectations for volatility	Suthaharan et al., 2021
Illusory pattern detection	Zhang et al., 2022
Reduced adaptation to changing contingencies	Zhang et al., 2022

Objectives of the thesis

We saw that CTs can emerge in non-clinical populations, notably during crises with uncertain issues. Several social, cognitive and computational factors, summarized in the introduction, were proposed to account for these rigid beliefs. This PhD work is an attempt to propose a coherent and integrated approach of CTs and its mechanisms through three experimental paradigms and one theoretical research.

The first experiment of this thesis will take an interest in a specific alteration of probabilistic input integration, called aberrant salience attribution, and its association with (i) belief conviction on one hand, and (ii) vaccine attitude on the other hand. This work has been accepted in *L'Encéphale* as a first-author publication.

The second experiment will try to decipher the link between the reduced ability to modulate likelihood outcome according to prior expectations about visual inputs and belief rigidity. It will also explore how this relation is modulated by stress and uncertainty levels. This work is still in preparation for publication as a first-author.

The third experiment will explore deeper the inferential mechanisms underlying the formation of conspiratorial beliefs under conditions of maximized uncertainty, using repeated-measures computational modeling. This research work has been presented in a national and two international conferences, and is currently under peer-review for publication as a first-author paper.

The last research from this thesis will focus on algorithmic simulations of larger communities and take an interest on the impact of a biased external source of information on belief formation and maintenance in social networks. This work has been submitted for publication as a second-author paper.

Chapter II

A cognitive approach of conspiracy beliefs: the roles of hypersalience and lack of control

Highlights

Uncertainty is not just a stressor; it's a game-changer that messes with our sense of control. When facing such uncertainty, our minds might lean into cognitive biases to make the world feel more predictable. Such biases would be at the roots of conspiracy theories. Among them, hypersalience has never been properly investigated. Furthermore, while some studies hint at a connection between control and conspiracy theories, replications have yielded mixed results. Well-controlled designs aiming at deciphering the respective contributions of stress and uncertainty on this potential correlation stay limited. To address these concerns, we thus conducted two complementary studies.

First, we used self-assessment measures of CTs, salience attribution and COVID-19 vaccine hesitancy to uncover how hypersalience fuels in the proliferation of conspiracy theories and anti-vaccination attitudes.

Second, we merged three critical elements: (i) experimental and self-reported indicators of lack of control, (ii) measures of conspiracy adherence, and (iii) stress ratings, both before and after the resolution of a tumultuous socio-political event. We demonstrate that the link between control threat and conspiracy adherence thrives in the midst of intense real-world uncertainty. We also unraveled that this effect was mediated by individual stress levels and was domain-specific, offering fresh insights into the intricate mechanisms weaving together lack of control and conspiracy beliefs.

II.1. Hypersalience is associated with conspiracy ideations and vaccine hesitancy[1]

Recent years saw a drastic surge in conspiracy theories, ranging from climate-change denial to suspicion about the real motives behind COVID-19 control measures. A better understanding of the cognitive roots of these unfounded and rigid beliefs is of crucial importance. We need to learn from this pandemic, overcome the population's rejection of science and support critical thinking at a societal level. Notably, conspiracy theories have been proposed not only as false beliefs, but also as a way of making sense of events occurring in a context of great social uncertainty (Pertwee et al., 2022). Interestingly, previous work underscored that perceiving some irrelevant elements of our environment as abnormally important, a phenomenon called *aberrant salience attribution*, was associated with the endorsement of rigid beliefs, such as delusional ideations (Kapur et al., 2005), or the acceptance of scientifically doubtful facts (Irwin et al., 2014). Surprisingly, despite some similarities in the unshakable nature of these belief categories (Suthaharan et al., 2021), the presence of aberrant salience among people embracing conspiracy theories has never been properly explored.

To remedy this situation, we ran an international online survey during the most critical period of the COVID-19 pandemic (2020 – 2022), in three Western countries chosen for their high degree of doubtfulness and polarization, i.e., the United States of America, the United Kingdom and France (https://www.visualcapitalist.com/polarization-across-28-countries/). A total of 699 adult participants, exempt of any current neurological or psychiatric disorder, were recruited via a dedicated platform (https://www.prolific.com/) to complete standardized validated questionnaires: the *Generic Conspiracist Beliefs* scale (GCB (Brotherton, CC French, et al., 2013)) and the *Aberrant Salience Inventory* (ASI (Cicero et al., 2010)), as well as a (0-10) *Visual Analog Scale* measuring trust in the COVID-19 vaccine in a single session. Complete anonymous data were collected for 691 participants (50% were female), aged 34 years ±12 on average. We found a strong positive association between ASI and GCB total scores (Pearson's **r=.43, p<.001**, **Fig.1a**). This effect

[1] Leclercq S., Szaffarczyk S., Jardri R. (*Encéphale*). Forged evidence and vaccine hesitancy during the covid-19 crisis (https://doi.org/10.1016/j.encep.2023.09.001).

appeared strongest when conspiracy theories were centered on health issues (**r=.42, p<.001**, **Fig.1b**). In contrast, ASI was found to be negatively associated with trust in the COVID-19 vaccine (**r=-.17, p<.001**, **Figure 11**).

We know that defiance in science can prevent patients from gaining access to the most appropriate care (Dubé et al., 2015). This phenomenon reached a climax during the COVID-19 outbreak during which a distrust in sanitary and medical authorities clearly emerged (Bierwiaczonek et al., 2022). For the first time, we confirmed that a particular cognitive bias, i.e., hypersalience, was also linked with a substantial increase in conspiracy theories, together with anti-vaccination attitudes. Even if the cross-sectional nature of the survey, as well as the use of correlational testing, do not allow us to address causality, longitudinal designs would help deciphering how hypersalience impacts the trade-off between (i) *exploring* new conspiracist explanations in a stressful context, versus (ii) *exploiting* prior beliefs (Kasper et al., 2023) assimilated to the prevailing view. Critically, this cognitive signature also paves the way for future assessment of educational and psychological programs, previously found effective in clinical populations (Sauvé et al., 2020), that could specifically target *aberrant salience attribution*, found associated with the rejection of pragmatic mitigation measures of one of the largest global health threats of the 21st century.

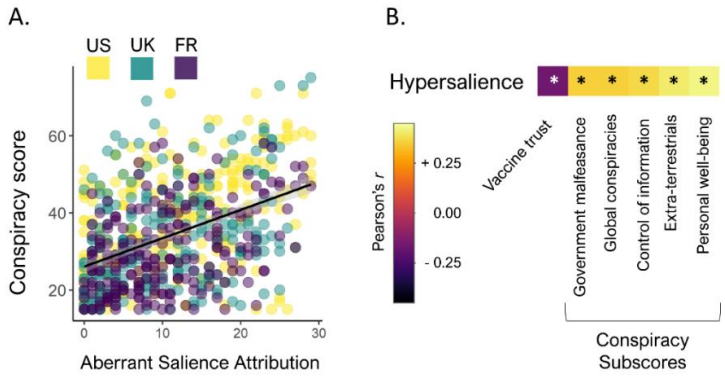


Figure 11 : Conspiracy theories and hypersalience. (A) Scatter plot showing the correlation between the scores on the *Aberrant Salience Inventory* (ASI) and *Generic Conspiracist Beliefs* (GCB) scales across the three tested samples (US stands for United States of America, UK for United Kingdom and FR for France): **r=.43, p<.001, n=691**. **(B)** Heatmap depicting the strength of the associations between ASI, trust in COVID-19 vaccines and GCB subscores for the whole sample (Pearson's correlations, corrected for multiple comparisons using the *false discovery rate* method, * stands for p<.001). While ASI is negatively associated with vaccine trust (**r=-.17**), it shows a positive association with GCB subscales (government malevolence: **r=.35**; malevolent global conspiracies, **r=.35**; control of information, **r=.37**; extraterrestrial cover-up, **r=.40**; and personal well-being **r=.42**). Interestingly, this last GCB subscore, exhibiting the strongest association, is focused on the deliberate spread of viruses or the use of technologies, including new drugs, reported to be able to mind-control people without consent (a theme compatible with conspiracist infodemics claiming a link between 5G and mRNA technologies).

II.2. Political distress mediates the association between lack of control and conspiracy ideations[2]

II.2.1. Introduction

When facing major life stressors with uncertain outcomes, like a serious illness or a societal crisis, the sense of personal control over one's life, or the unfolding of events, can be adversely affected. This deterioration is subject to various individual vulnerability factors, ranging from the personal need for structure (Noordewier and Rutjens, 2021) to the childhood environment (Mittal and Griskevicius, 2014). In turn, the subjective loss of perceived control in times of uncertainty was found to be associated with changes in beliefs and attitudes (Bukowski et al., 2017; Peluso and Pichierri, 2021; Thompson et al., 1993; Zhu et al., 2020).

More specifically, it has been suggested that, in the event of a global social crisis, blaming an external group could constitute a coping strategy for re-establishing a sense of personal control (Bukowski et al., 2017; Sullivan et al., 2010). Similarly, conspiracy theories are generally defined as beliefs assuming the existence of a group or organisation secretly plotting against the common good and for its own benefit. The *Compensatory Control Theory* (Kay et al., 2008) links these two notions, asserting that the adherence to conspiracy theories (**CTs**) can serve as a coping mechanism when confronted with a subjective loss of control over the course of events (**LoC**).

There is a growing, yet contrasted, evidence for such an association in the literature. While several studies support Kay's theory (Whitson and Galinsky, 2008), notably when control in the political arena is threatened (Kofta et al., 2020; Pantazi et al., 2022; Stojanov et al., 2022; van Prooijen and Acker, 2015), attempts at replicating these findings have not always been successful (Hart and Graether, 2018; Nyhan and Zeitzoff, 2018; Stojanov et al., 2020; van Elk and Lodder, 2018). The LoC-CTs association is likely driven by several common co-variables (Stojanov et al., 2020), including stress (Swami et al., 2016) and uncertainty (van Prooijen and Jostmann, 2013), that may have biased previous findings.

[2] Leclercq S., Szaffarczyk S., Leptourgos P., Yger P., Cachia A., Whatelet M., Denève S., Jardri R., (*in preparation*). Political distress mediates the association between lack of control and conspiracy ideations.

Inspired by this idea, considerable efforts have been made to ecologically assess the impact of real-world events fraught with uncertainty — such as political elections, natural disasters, or the COVID-19 pandemic — on LoC and simultaneous CT adherence (Šrol et al., 2021; Stojanov et al., 2022). However, experimental designs aimed at elucidating the role of uncertainty in this LoC-CTs association are rare (Dow et al., 2022). Furthermore, the widespread use of laboratory procedures provides only limited information on the actual real-world validity of the relationship between control threat and conspiracy ideations (van Prooijen and Acker, 2015), encouraging future studies to adopt an ecological approach to address the central claims of the compensatory control model (Stojanov et al., 2022).

To reach this goal, we assessed participants online regarding their levels of LoC, stress and conspiracy ideations around the resolution of socio-political events known to induce high uncertainty. We used self-report measures and introduced a complementary and more objective approach to control assessment, i.e., the participants' ability to cognitively force bistable perception towards a chosen interpretation. This ability is known to be reduced in psychosis, a condition associated with rigid beliefs and inferential biases (McBain et al., 2011). In line with previous work examining the conditions under which LoC are most likely to affect CTs (Noordewier and Rutjens, 2021), we assumed that the LoC-CT association would (i) be strengthened under conditions of heightened socio-political uncertainty, and (ii) be mediated by individual levels of political distress.

II.2.2. Methods

Population

Four hundred and eighty-five participants were recruited online via the dedicated Prolific© web-platform (https://www.prolific.co/). The same participants were assessed one month before and one month after a major political event known to induce stress. A first subset of 260 participants were recruited in the UK around the 2021 Brexit implementation, followed by a second subsample of 225 French participants around the 2022 Presidential elections. Targeted participants were of the nationality of the tested country. Participants were aged from 18 to 60 years old and had normal or corrected-to-normal vision. They reported having no known history of psychiatric or neurologic disorder and no ongoing treatment. From the initial samples,

11 participants were excluded based on failed attentional checks during the procedure or very-low reaction times (mean reaction time < 300ms), while 63 were lost longitudinally. Data from 411 participants were analysed in total (n_{UK}=225, n_{FR}=186). The Prolific© web-platform ensures data privacy following standards of the European and UK data protection law (i.e., General Data Protection Regulation (GDPR), transposed into UK law as the UK GDPR). Participants' sociodemographic characteristics were associated with their respective behavioral data through an anonymous ID randomly assigned at enrollment. The overall online survey complies with French regulations and ethics (*Comité de Protection des Personnes Nord-Ouest IV*).

Task apparatus

The protocol was implemented in PsychoPy v3, then exported and hosted online on the Pavlovia.org platform. For the perceptual control part of the experiment, participants were instructed to stand in total darkness, approximately 60 cm away from the screen and to adjust it to be perpendicular to the floor with their eyes aligned to the fixation cross displayed at the center of the screen. The perceptual task and the self-reported assessment of beliefs were administered in a randomized order.

Self-reported measures

Participants were asked to indicate their age and educational attainment as defined in the *International Standard Classification of Education* (**ISCED**) (UNESCO Institute for Statistics and Statistics, 2020). Socio-demographic features of the participants are displayed in **Table 2**. At each time-step, we assessed conspiracy ideations using the 15-item *Generic Conspiracist Beliefs Scale* (**GCB**) (Brotherton, CC French, et al., 2013) and its French translation (Lantian et al., 2016). Participants were also asked to rate their levels of individual distress related to the political event at play in their country using visual analogical scales ranging from 0 to 10 (**political distress**). The precise questions used are reported in the *Appendix A* of the *Supplementary Materials* section.

Subjective sense of control was assessed using the *Midlife Development Inventory* (**MIDI**) (M. E. Lachman and Weaver, 1998a, 1998b). To produce a French version of this scale, a French researcher translated the scale (with the instructions) from English to French. Then, two independent bilingual researchers (a French native speaker and an English native-speaker) back-translated this version from French to English. Finally,

the first (English-to-French) translator compared the back-translated version to the original. Remaining discrepancies were resolved in a discussion including the three researchers involved. This French version of the MIDI can be found in *Appendix B* of the *Supplementary Materials* section.

When Likert or visual analogical scales were used, the cursor was coded to return to the center of the screen after each question to avoid the answer being biased by previous ones. Attentional checks were integrated during the psychometric assessment to ensure that participants did not provide random answers. Among the scales used, five items were randomly added, regularly asking participants for a specific answer (example: "*This is an attentional check, please answer 'Not sure/cannot decide' to that question.*").

Perceptual control over bistable stimulus

A perceptual control score was defined as the amount of control a participant was able to exert over their perception of a bistable stimulus (i.e., force one of the two possible interpretations of the Necker cube - NC). Naturally, the interpretation of a 2-D NC projected from a 3-D space alternates between two possible configurations: a *seen from above* (SFA), or a *seen from below* (SFB) cube. To assess perceptual control, a task (**Figure 12-a**) was administered as follows.

Visual stimuli representing Necker cubes (NC) were displayed in the center of a black screen. The stimulus size was standardized across the participants using a matching method based on a standard credit card displayed on the screen that the participant was required to adjust in size before starting the experiment. Participants were instructed to stare at the target located in the middle of the screen to neutralize the potential effects of eye movements. The two possible interpretations of the NC (SFA, SFB) were explicitly mentioned and a training session ensured participant's comprehension of the task.

The block-design of the task was inspired by Mamassian and Goutcher's protocol (Mamassian and Goutcher, 2005). During each block, a NC was presented discontinuously. Using a forced-choice methodology, we asked participants to report their interpretation of the stimulus using their keyboard each time a new cube appeared on the screen. The cube disappeared after a pseudorandom duration (**ISI** ranging from 0.1 to 1.2 seconds). Each recorded response constituted a trial, and the experiment was divided into 2 blocks of 64 consecutive trials.

Participants were instructed to consciously bias their perception towards the SFA interpretation during one of the two blocks and conversely towards the SFB interpretation during the other block. This procedure was chosen to neutralize the potential effects of preference participants may hold towards one configuration or the other. The order between those two blocks was randomized and they were separated by a new instruction screen.

To assess perceptual control, each response was assigned a value: 1 for "SFA" responses and 0 for "SFB" responses. The average of these “mean resp” values (MR) can be interpreted as the overall probability of perceiving the "SFA" interpretation where 1 is the maximum probability for the "SFA" interpretation and 0 is the maximum probability for "SFB". A MR of 0.5 would characterize a purely stochastic system in which the two percepts are equiprobable. Perceptual control was equal to (MR) or (1-MR) in the blocks where perception had to be forced on SFA or SFB configurations respectively. As a result, perceptual control corresponded to a 0-to-1 score, where 0 means a complete lack of control over the stimulus, and 1 that the participant perceived only the instructed interpretation.

Statistical Analysis

A first series of Wilcoxon’s signed rank tests was run to assess the differences across samples for each variable. Normality of conspiracy beliefs distribution was assessed using a Shapiro-Wilk normality test.

Due to the non-normal distribution of conspiracy ideations (**W = .938, p < .001**), we assessed the link between (i) GCB scores and subscores and (ii) LoC, using both MIDI and cognitive control scores, with Spearman rank-correlation analysis, corrected for multiple comparisons with the *false discovery rate* (FDR) method. The same procedure was used at baseline and retest, before and after the resolution of uncertainty.

To check the role of distress on the statistically significant associations, we performed a mediation analysis using political distress as a mediator. Referring to the percentile method, we ran 500 simulations with non-parametric bootstrap confidence intervals and followed the Barron and Kenny’s steps (Baron and Kenny, 1986). To assess the dynamics of LoC, we finally compared MIDI and perceptual control at test and retest using a Wilcoxon signed-rank test for repeated measures.

II.2.3. Results

Table 2 : Features of populations at baseline

	Whole Sample (n = 411)	**UK (n = 225)**	**FR (n = 186)**
Sex (F /M)	212 / 199	119 / 106	93 / 93
Age (y.o.)	34.4 ± 11.2	38.2 ± 10.7	29.8 ± 9.90
Education	5.85 ± 1.35	5.52 ± 1.44	6.25 ± 1.12
pol. distress	5.74 ± 2.94	5.69 ± 3.32	4.87 ± 3.07
MIDI	56.9 ± 11.7	57.6 ± 12.0	56.1 ± 11.3
pCtrl.	.678 ± .220	.631 ± .255	.735 ± .150
GCB	31.8 ± 12.6	33.3 ± 13.7	30.0 ± 11
Control of information	8.05 ± 3.05	8.08 ± 3.05	8.02 ± 3.07
Government malfeasance	6.79 ± 3.04	6.98 ± 3.21	6.56 ± 2.82
Malevolent global conspiracies	6.37 ± 3.14	6.63 ± 3.40	6.06 ± 2.78
Personal well-being	5.48 ± 2.65	6.02 ± 2.86	4.82 ± 2.20
Extraterrestrial cover-up	5.09 ± 2.80	5.56 ± 3.06	4.51 ± 2.33

UK: United Kingdom; FR: France; F/M: female or male; y.o., years old; Education levels are provided according to the International Standard Classification of Education (ISCED); pol. distress: political distress; MIDI: Midlife Development Inventory; pctrl: Perceptual Control; GCB: Generic Conspiracist Beliefs Scale. The sex-ratio did not differ across samples (X^2 = .234, p = .628). UK participants were significantly older (W = 10932, p < 0.001, Cohen's *d* = .810) and FR participants reached a higher educational attainment (W = 27554, p < 0.001, Cohen's *d* = .560). UK participants demonstrated a higher level of conspiracy endorsement (W = 18411, p = .036, Cohen's *d* = .260) and but these differences were only observable for *Personal well-being* and *Extraterrestrial cover-up* subscales (W = 15601, p < 0.001, Cohen's *d* = .470 ; W = 16513, p < 0.001, Cohen's *d* = .380). French participants exerted higher levels of perceptual control (W = 25354, p < 0.001, Cohen's *d* = -.490) while political distress levels and subjective sense of control were consistent across samples (W = 20363, p = .640 ; W = 19441, p = .216).

Due to (i) similar UK/FR context of socio-political uncertainty at baseline , (ii) consistency of self-reported political distress (W = 20363, p = .640), (iii) of sense of control (W = 19441, p = .216) and (iv) of perceptual stability (W = 20594, p = .783) we merged the two sub-samples together for the remainder of the analysis (of note, perceptual stability was computed based on the procedure detailed in the "*Judgement criterion*" section of Chapter III).

A first series of analyses was conducted under conditions of maximized socio-political uncertainty (1st time-point, before the resolution of the political event). Partially replicating some findings of the literature, we found a negative link between MIDI and

GCB scores (**p = .012, ρ = -.121,** n = 428, **Figure 12-b**). We then tested whether the same effect could be detected using an objective measure of perceptual control instead of the subjective sense of control. Again, a negative association between control and GCB score was evidenced (**p = .013, ρ = -.123, Figure 12-c**). However, while the self-reported measure of control was significantly associated with political distress (**p < .001, ρ = -.185**), no association between perceptual control and stress could be detected (p = .190), suggesting a possibly different underlying mechanism.

Examining the various categories of conspiracy theories as outlined by the GCB scale (**Figure 12-d**), we discovered some intriguing correlations. Specifically, we observed that a reduced sense of subjective control was linked with higher-level beliefs such as *Government Malevolence* (**p = .011, ρ = -.013**) and *Control of Information* (**p = .050, ρ = -.011**). Conversely, actual perceptual control was found negatively correlated with personal or paranormal conspiracy themes, such as *Personal Well-Being* (**p = .006, ρ = -.015**) and *Extraterrestrials Coverup* (**p = .014, ρ = -.013**).

A second series of analyses aimed at elucidating the possible modulatory effects of stress and uncertainty on these associations. First, we wondered whether the CT-LoC association evidenced prior to the event resolution was influenced by individual levels of socio-political distress. To this end, we performed a mediation analysis using MIDI as a predictor of GCB and political distress as a mediator. To conduct this first mediation analysis, we performed a linear model to assess the impact of MIDI on GCB that revealed a trend (**F = 2.821, p = .0938, R^2= .00685**). Due to the strong theoretical background supporting the MIDI-GCB association and the significant correlation between the two variables previously highlighted we decided to pursue further the analysis following Shrout & Bolger recommendations (Shrout and Bolger, 2002). A second linear model then assessed the impact of the MIDI on Political Distress, which was significant (**F = 16.75, p< .001, R^2= .0393**). A third linear model comprising both MIDI and political distress as predictors of GCB also achieved statistical significance (**F = 10.11, p < .001, MIDI : p = 0.0127; political distress : p < 0.001, R^2= .0472**). The mediation analysis was the final step of this pipeline confirming an influence of individual stress levels on the MIDI-GCB association (**ACME = .044, p < .001, Figure 12-e**). A second mediation analysis investigating the GCB-perceptual control association was run following the same steps as the first but did not confirm an effect of individual stress levels on the association between objective control and GCB.

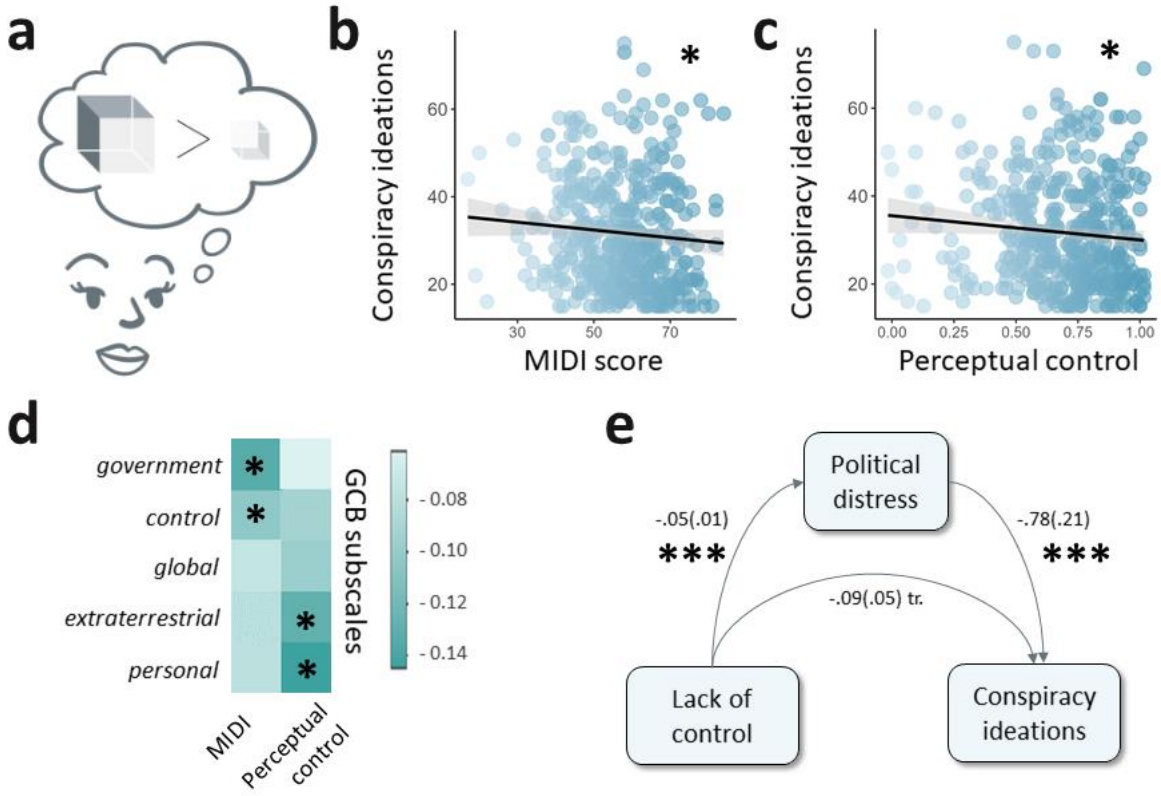


Figure 12 : Sense of control is associated with conspiracy theories. (a) The objective control score was computed based on a perceptual task during which participants were asked to consciously bias their perception towards one interpretation of the Necker Cube. **(b)** Scatter plot showing the correlation between conspiracy ideations measured with the *Generic Conspiracist Beliefs Scale* (GCB) and the *Sense of Control Scale from the Midlife Development Inventory* (MIDI) scores (spearman correlation, **ρ = -.121**). **(c)** Scatter plot illustrating the correlation between GCB and perceptual control scores (spearman correlation, **ρ = -.123**). **(d)** Heatmap depicting the strength of associations between sense of control and domain-specific conspiracy ideations (GCB subscales: control of information, government malfeasance, malevolent global conspiracies, personal well-being and extraterrestrial cover-up ; Pearson's correlations, corrected for multiple comparisons using the *false discovery rate* method). MIDI scores were found to be significantly associated with government malfeasance and control of information themed conspiracy theories (**ρ = -.013**; **ρ = -.011**) while perceptual control was found more associated with personal well-being preoccupations (**ρ = -.015**) and extraterrestrials related beliefs (**ρ = -.013**). * stands for p<.05, *** stands for p<.001. **(e)** The association between subjective sense of control and conspiracy ideations is mediated by political distress (ACME = .044).

Investigating the dynamics of LoC, we found that the perceptual control measure at the time of testing increased between the first and second assessment (**W = 37323, p =.021, Cohen's d = -.209**) suggesting that sense of control was restored after resolution of uncertainty. However, echoing the literature discrepancy, we did not replicate at retest, the GCB-MIDI, GCB-perceptual control or MIDI-stress correlations evidenced at baseline (p = .321; p = .228; p = .116), suggesting that the latter was mainly driven by political uncertainty.

II.2.4. Discussion

Significant life stressors can alter individuals' sense of control. This perceived loss of control in the face of uncertainty has already been associated with changes in beliefs and attitudes (Bukowski et al., 2017; Peluso and Pichierri, 2021; Thompson et al., 1993; Zhu et al., 2020). The *Compensatory Control Theory* (Kay et al., 2008) further suggested that CTs might help to cope with perceived losses of control in stressful situations, even if some findings were sometimes difficult to replicate. These mixed results were proposed linked with the limited use of objective measures of control (Dow et al., 2022), as well as a lack of real-world settings (Stojanov et al., 2022; van Elk and Lodder, 2018). In the present paper, we addressed these concerns by combining experimental and self-reported LoC before and after the resolution of socio-political events. We measured CTs and individual levels of socio-political distress online at each time-step. This approach allowed us to assess the effects of stress and uncertainty on potential LoC-CTs association in a naturalistic way.

We were able to show that in the case of high uncertainty, there is a negative association between conspiracy endorsement and the sense of control, both subjectively and experimentally measured. Our results are consistent with those suggesting that in people facing a major and uncertain stressor, such as a cancer diagnosis, perceived LoC is associated with irrational beliefs (Thompson et al., 1993). These results also bolster Kay's theory of *Compensatory Control* (Kay et al., 2008) which has been supported by consistent evidence across the literature (Whitson and Galinsky, 2008), especially when the threat of control pertains to the socio-political domain (Kofta et al., 2020; Pantazi et al., 2022; Stojanov et al., 2022; van Prooijen and Acker, 2015). Interestingly, Stojanov and colleagues (Stojanov et al., 2020) highlighted the fact that several covariates common to LoC and CTs, notably stress (Swami et al., 2016) and uncertainty (van Prooijen and Jostmann, 2013), could explain this association. In line with this idea, we were able to evidence that the subjective feeling of control - conspiracy association was mediated by the individual's vulnerability to stress.

Furthermore, echoing discrepancies reported in the scientific literature and the idea that the link between CT and LoC could be transient (Stojanov et al., 2020), the LoC-CTs correlation initially evidenced at baseline resolved at retest, suggesting that

this association might be driven by political uncertainty. This result also confirms the idea that inducing a threat of control may not be sufficient to reinforce CTs (Stojanov et al., 2020). The low level of evidence from experiments which explicitly measured the impact of control modulation on CTs could be explained by the insufficient level of stress and uncertainty induced by these manipulations in unnatural contexts. On the contrary, a chronic or systemic control threat might be more likely to elicit CTs endorsement (Bilewicz, 2022).

Another feature of the LoC-CTs association that can make it difficult to reveal lies in its domain-specific nature i.e., the fact that a control threat in one domain is more likely to predict CTs related to the same topic (Stojanov et al., 2020, 2023). Congruent with the idea that different types of control threat are related to different conspiratorial themes, Oleksy and colleagues (Oleksy et al., 2021) notably evidenced that individual or collective LoC can predict different CTs. Our results seem consistent with these recent findings. More explicitly, we showed that the subjective sense of global control was associated with endorsement of CTs related to socio-political concerns, while actual perceptual control of the NC, a factor less influenced by socio-political distress, was more associated with non-human related conspiracy themes (infectious agents, new technologies, other life forms…). These results suggest that different types of CT-LoC associations could rely on different cognitivo-perceptual mechanisms.

Overall, these results provide valuable insights for understanding previous contradictory findings and suggest key directions for future research. Highlighting the transient nature of the association between LoC and adherence to CTs during periods of political uncertainty bolsters Stojanov and colleagues' recommendation for naturalistic designs (Stojanov et al., 2022), exploring the impact of real-world events. Furthermore, by demonstrating the significant influence of uncertainty and stress vulnerability on the LoC-CTs relationship, we highlight the importance of carefully controlling for these factors and other potential mediators in future research. We also address the question of the core mechanisms underlying the LoC-CTs association by proposing the use of an objective measure of perceptual control, which has proven to be less sensitive to these confounding variables and associated with less social types of CTs. Moreover, in line with previous findings (Oleksy et al., 2021; Stojanov et al., 2020, 2023), we confirm the domain-specific nature of the LoC-CTs association and

recommend the use of dimension-based metrics designed to assess adherence to various conspiracy-theory themes.

This study is not exempt from limitations. First, due to the nature of conspiracy theories and the controversy and potential stigma associated with these beliefs, we may wonder whether our participants were hesitant to provide honest answers during the self-report parts of the research. However, we believe that the sense of anonymity offered online might have overcome this hesitancy. In addition, we think that the use of a low-level perceptual task might have provided a good proxy of control assessment, unlikely prone to social biases such as interviewer compliance. Considering the recent development of web-based eye-tracking tools, future research endeavors might however integrate more sophisticated apparatus to explore perceptual control such as the use of ocular temporal windows (Polgári et al., 2020), even if still difficult to properly implement online.

A second potential limitation concerns the representativeness of our sample. Indeed, we only tested two *Western Educated Industrialized Rich and Democratic* (W.E.I.R.D.) countries that share a similar geographical location and are therefore subject to close cultural influences and climatic challenges. We know that internalized trauma history along with low political power and education can be intertwined with chronic feelings of powerlessness and lack of control capable of modulating the LoC-CTs association. It would be interesting for future research to replicate these findings in a larger and diversified panel of countries/populations.

We believe this work could pave the way for future research that could explore the LoC-CTs link (i) in more natural contexts (ii) controlling for the role of other covariates shared by LoC and CTs, such as chronic feelings of helplessness and anxiety and (ii) in terms of specific CTs. Overall this paper, which provides new insights into the mechanisms underlying the control-conspiracy association, might be part of a broader field of research aimed at studying the psychological and cognitive processes underlying belief formation. Gaining better understanding of individual vulnerability factors that influence the spread of conspiracist and pseudoscientific beliefs, such as the response to stress and uncertainty, might help us to develop interventions capable of targeting them efficiently.

Chapter III

Circular Inference accounts for conspiracy beliefs and perceptual inference in times of uncertainty[3]

Highlights

Sociopolitical crises causing uncertainty have accumulated in recent years, providing fertile ground for the emergence of conspiracy ideations. Computational models constitute valuable tools for understanding the mechanisms at play in the formation and rigidification of these unshakeable beliefs. Here, the Circular Inference model was used to capture associations between changes in perceptual inference and the dynamics of conspiracy ideations in times of uncertainty. A bistable perception task and conspiracy belief assessment focused on major sociopolitical events was performed on large populations from three polarized countries. We show that when uncertainty peaks, an overweighting of sensory information is associated with conspiracy ideations. Progressively, this exploration strategy gives way to an exploitation strategy in which increased adherence to conspiracy theories is associated with the amplification of prior information. Overall, the Circular Inference model sheds new light on the possible mechanisms underlying the progressive rigidification of conspiracy theories when individuals face highly uncertain situations.

[3] Leclercq S., Szaffarczyk S., Leptourgos P., Yger P., Fakhri A., Wathelet M., Bouttier V., Denève S., Jardri R (*under review*). Conspiracy beliefs and perceptual inference in times of political uncertainty (preprint available with this DOI : 10.31234/osf.io/x3fc6).

III.1. Introduction

Conspiracy theories (CTs) are appearing with increasing frequency in our modern societies, with criticism of mainstream knowledge and scientific evidence at center stage. CTs are commonly defined as beliefs assuming the existence of a secret group or organization operating maliciously and for its own benefit. Adherence to multiple unrelated CTs that contradict each other is common (Drinkwater et al., 2012; Goertzel, 1994; Swami et al., 2011; Wood et al., 2012), suggesting common underlying mechanisms by which belief in CTs arises.

Interestingly, a first line of research revealed that these rigid beliefs often crystallize around highly polarizing societal or political events (van Prooijen and Douglas, 2017) and may serve as coping mechanisms for stress and loss of control when uncertainty increases sharply (Dow et al., 2022; Farias and Pilati, 2023; Sullivan et al., 2010; van Prooijen and Acker, 2015; Whitson and Galinsky, 2008). Although CTs can induce widespread misconceptions - as it has been observed during the COVID-19 pandemic - they also constitute intuitive explanations for complex issues (e.g., simple cause-effect relationships), that can meet the need to restore predictability (Douglas et al., 2019) at the cost of suboptimal reasoning.

A second line of research focused on the role of reasoning biases in CT emergence (Brotherton and French, 2015; Georgiou et al., 2021; Wycha, 2015). According to this framework, it is thought that conspiracists bias the weight they attribute to certain stimuli to reduce uncertainty (Kabengele et al., 2023; Pytlik et al., 2020), sometimes leading people to jump to conclusions (JTC) when probabilistic decisions must be made. Conspiracy ideations have also been associated with a more intuitive thinking style (Drinkwater et al., 2012; Georgiou et al., 2021) than the common analytical approach. This tendency toward fast, preconscious and spontaneous processing could be based on specific reality-testing deficits in people endorsing CTs (Lewandowsky et al., 2013).

These results have not always been replicated, leading some authors to wonder whether CTs could mainly be traced back to social constructs (Hartmann and Müller, 2023; Müller and Hartmann, 2023; Raihani and Bell, 2019). However, others suggest that social learning depends on broader associative mechanisms responsible for the detection of predictive relationships in every natural domain (Heyes, 2012); thus, Bayesian methods could be a complementary approach to addressing the existing link

between CTs and uncertainty. This framework assumes that cognitive and perceptual factors are rooted in a common probabilistic mechanism (Helmholtz, 1948). Surprisingly, only a few attempts have been made to investigate the potential links between perceptual inference and conspiracy ideations in a controlled experimental setting.

Some results from the CT literature appear compatible with a probabilistic formalism. Dagnall and colleagues (Dagnall et al., 2015) explored the link between CTs and a wide range of cognitive-perceptual factors. They showed that such factors, including hallucination proneness, often conceptualized as false inferences (Fletcher and Frith, 2009), were associated with CTs. Additionally, conspiracy ideations were found to be associated with illusory visual pattern detection (Müller and Hartmann, 2023; van Prooijen et al., 2018), a phenomenon regularly explored through the prism of Bayesian theory (Geisler and Kersten, 2002).

Very few papers have directly fitted computational models to behavioral data in nonclinical samples with some noticeable exceptions exploring paranoia and/or conspiracy ideations (Barnby et al., 2022; Suthaharan et al., 2021). Purely theoretical papers also confirmed that computational approaches could help to better understand the spreading of CTs on simulated or social-media data (Cook and Lewandowsky, 2016; Madsen et al., 2017). Crucially, a more personalized computational lens (Rigoli, 2022), and a study of CTs in their ecological environment (Stojanov et al., 2022) seem required to decipher the respective contributions of sociopolitical factors and information weighting in CTs' emergence.

Thus, combining the strength of normative and ecological research during uncertain societal crises appears necessary to establish a bridge between CT and inference quantification. In the present paper, we referred to *Circular inference* (CI), a Bayesian framework that has proven effective in capturing both perceptual suboptimality in nonclinical populations (Leptourgos, Notredame, et al., 2020) and JTC in patients with psychosis (Jardri et al., 2017; Simonsen et al., 2021). We hypothesized that by fitting the CI model to a simple bistable task (which maximizes ambiguity at the perceptual level), we could benefit from an ideal setup to challenge the potential links between (i) the inferential mechanisms at play under conditions of extreme uncertainty, and (ii) the dynamics of conspiracy ideations in large populations exposed to natural sociopolitical stress.

III.2. Results

III.2.1. Measuring multilevel inference before and after stressful political events

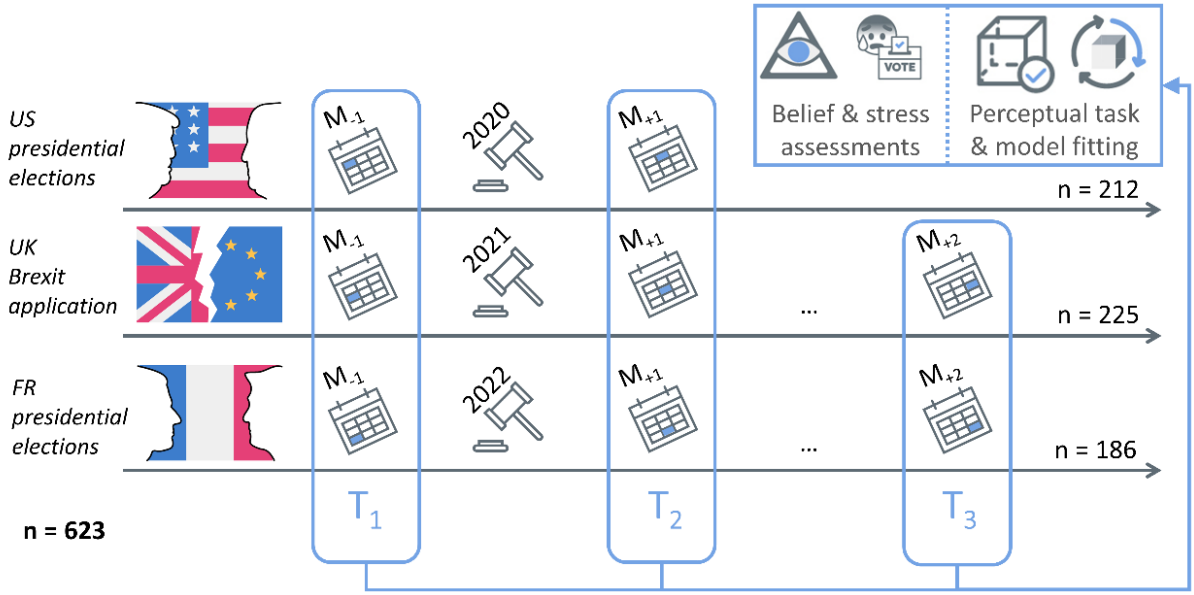


Figure 13. A repeated-measures design framing stressful political events in 3 different countries. Conspiracy ideations, political distress and perceptual stability were measured in the same participants (n = 623) via an online procedure, before and after the occurrence of a polarizing political event in three Western countries (M stands for month): the 2020 presidential election in the *United States of America* (n = 212, US), BREXIT implementation in the *United Kingdom* (n = 225, UK) and the 2022 presidential elections in *France* (n = 186, FR). We used T1 and T2 measures in the main model, while T3 was used in control analyses (see *Supplementary Material section: Controlling for experimental design biases*).

Because we assumed that periods of great sociopolitical uncertainty lead to significant increases in individual levels of distress and favor inferential biases such as conspiracy endorsements, we explored rigid beliefs and perceptual stability around polarizing political events in three independent Western countries (see **Figure 13**): the *United States of America* (US, 2020 presidential elections), the *United Kingdom* (UK, 2021 BREXIT implementation) and *France* (FR, 2022 presidential elections). At each time point, healthy participants were instructed to rate their level of distress related to the ongoing event in their own country (later referred to as ***political distress***, see *Methods* and *Supplementary Material section: Self-reported measures*).

Table 3. Description of the populations at baseline.

	Sex (F /M)	**Age** (y.o.)	**Education**	**pol. distress**	**GCB**	**stability**
Whole Sample (n = 623)	310 / 311	33.0 ± 10.9	5.64 ± 1.41	5.07 ± 3.48	33.8 ± 13.3	.572 ± .178
US (n = 212)	98 / 112	30.4 ± 9.83	5.22 ± 1.43	4.60 ± 3.88	37.7 ± 13.9	.585 ± .173
UK (n = 225)	119 / 106	38.2 ± 10.7	5.52 ± 1.44	5.69 ± 3.32	33.3 ± 13.7	.566 ± .178
FR (n = 186)	93 / 93	29.8 ± 9.90	6.25 ± 1.12	4.87 ± 3.07	30.0 ± 11	.565 ± .183

US: United States of America; UK: United Kingdom; FR: France; F/M: female or male; y.o., years old; Education levels are provided according to the International Standard Classification of Education (ISCED); pol. distress: political distress; GCB: Generic Conspiracist Beliefs Scale; stability: fitted stability score (see *Methods section: Judgment criterion*). The sex-ratio did not differ across samples ($X^2 = 1.68$, $p = .431$). UK participants were significantly older ($F(2.408) = 44.255$, $p < 0.001$, $\eta^2 = 1.29e^{-19}$) and FR participants reached a higher educational attainment ($F(2.411) = 35.458$, $p < 0.001$, $\eta^2 = 3.76e^{-13}$) than the other samples. UK participants demonstrated a higher level of political distress ($F(2.408) = 5.8388$, $p = .00316$, $\eta^2 = 2.82e^{-3}$), while stability was consistent across samples ($F(2.405) = 0.81828$, $p = .442$).

III.2.2. Necker cube experiment

At each time point, the 623 enrolled participants performed an online bistable perception task based on the Necker cube (NC). The interpretation of the two-dimensional NC projected from a three-dimensional space naturally alternates between two possible configurations: a *seen from above* (SFA), or a *seen from below* (SFB) cube (**Figure 14-a**). A perceptual ***stability score***, ranging from 0 to 1, was estimated at the participant level. This score corresponds to the probability of switching from one interpretation to the other (0 means total instability, while 1 reflects a perceptive rigidity where the participant only sees one interpretation of the two, see the *Methods* section). Assuming a universal mechanism at the roots of belief formation, we merged the 3 samples after ensuring their comparability in terms of perceptual stability at baseline (**Table 3Table 3. Description of the populations at baseline.; Figure 14;** see also *Supplementary Material section: Controlling for experimental design biases*). Importantly, perceptual stability was tested for *in lab/online* within-subject reproducibility on a pilot independent sample before running the final online experiment (**Figure 14-d,e**). We also ensured that dynamic changes in stability between the different time points were not due to a simple training effect between the sessions (see *Supplementary Material section: Controlling for experimental design biases*).

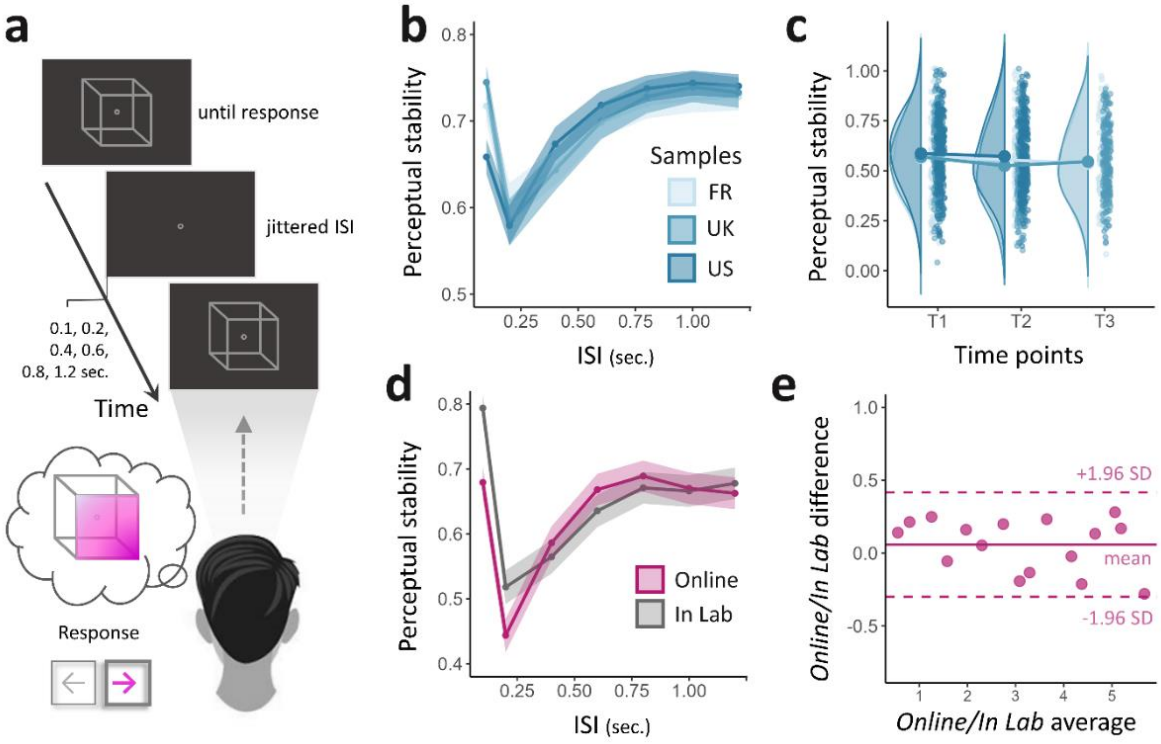


Figure 14. The Necker cube (NC) task: procedure and validity. (a) The experimental procedure consisted of serial NC presentations. Each trial was decomposed into three steps (see *Methods* section). After a fixation cross of pseudorandomized duration (ISI) (1), the Necker cube was presented (2) until participants reported their interpretation of the stimulus: 'seen from above' (SFA) or 'seen from below' (SFB), using the right or left arrow of their keyboard, respectively (3). **(b)** Perceptual stability as a function of the interstimulus interval (ISI) for each national sample. US (United States of America, mean Stability Score= .587, s.d.= .172), UK (United Kingdom, mean Stability Score= .570, s.d.= .176) and FR (France, mean Stability Score= .565, s.d.= .1472). **(c)** Averaged stability scores at each time point for the three national samples. **(d)** Perceptual stability as a function of ISI, for online (mean stability score= .441, s.d.= .190) and in-lab methods (mean stability score= .500, s.d.= .140). **(e)** Bland-Altman plot of the agreement between online and in-lab methods comparing stability scores obtained in each condition for the same participants (n = 16). The x axis represents the average scores of the two methods. The y axis represents the mean difference between online and in-lab stability scores. The limits of agreements (LoA, pink dotted lines) are defined as the mean difference computed on all participants (pink line) ± 1.96 s.d., and each dot represents a participant. As all participants are included in the LoA, the methods are considered to be in agreement and may be used interchangeably.

III.2.3. Conspiracy adherence measures

The participants were instructed to self-rate their level of adherence to CTs, completing the *Generic Conspiracist Beliefs Scale* (**GCB**, see *Methods* section) at each time step. Replicating previous findings, we showed that conspiracy ideations were not normally distributed across the tested participants (**$W = .954$; $p = .440e^{-12}$**, **Figure 15-a; Supplementary Figure 1-a**), suggesting that only a subpart of the general population commonly endorses such beliefs. The distribution of the total GCB scores differed across the three samples (**$\chi^2 = 31.5$, $p < .001$, $\eta^2 = .348e^{-07}$**) despite a similar

pattern across subscales (**Supplementary Figure 1-a-b, Supplementary Table 1**), notably demonstrating a common preoccupation for information control.

Looking more precisely at the sociodemographic features associated with conspiracy endorsement, we replicated previous findings from the literature (see *Supplementary Material section: Sociodemographic features of conspiracy theories*), notably showing that despite an absence of a link with the sex of participants (**Figure 15-b**), GCB scores significantly differed as a function of age (**$F(2,620) = 3.10$, $p = .046$, $\eta^2 = .039$**, **Figure 15-c**) or education (**$F(2,620) = 13.5$, $p < .001$, $\eta^2 = .395e^{-05}$**, **Figure 15-d**). Thus, we retained those variables as covariates for later analyses.

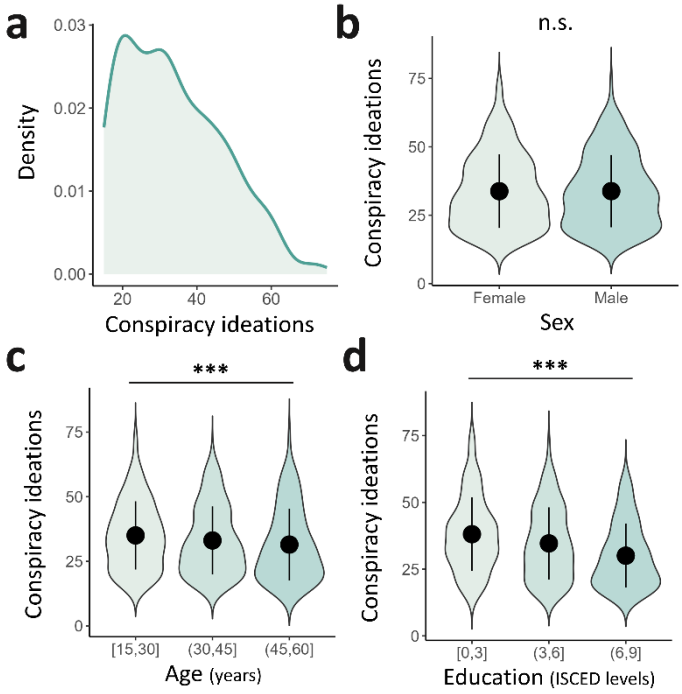


Figure 15 : Sociodemographic features associated with conspiracy theories at baseline. **(a)** Left-skewed distribution of GCB scores across the entire international sample (N = 623). **(b)** Mean conspiracy scores in females (n = 310, mean = 33.8, s.d. = 13.5) and males (n = 312, mean = 33.8, s.d. =13.2). The between groups difference was not significant. **(c)** Mean conspiracy scores according to age level. *Young* participants (n = 310, age = [18;30]) displayed higher GCB scores (mean = 35, s.d. = 13.2) than the *adults* (n = 210, age = (30;45], mean = 33.1, s.d. = 13.2) and older *adults* (n = 103, age = (45;60], mean = 31.5, s.d. = 13.8). **(d)** Mean conspiracy scores according to educational attainment levels. *The low education* group (n = 86, ISCED = [0;3]) scored significantly higher on GCB (mean = 38.2, s.d. = 13.7) than the *medium education* (n = 363, ISCED = [3;6], mean = 34.7, s.d. = 13.5) and the *high education* groups (n = 179, ISCED = [6;9], mean = 30, s.d. = 12).

III.2.4. Stress correlates at baseline

We assume that some participants might adopt information-processing strategies that can reduce the uncertainty induced by the framed political event. Notably, we expect that the search for stability would translate into high levels of confidence measurable at different levels of processing, from perception to conspiracy beliefs. Since belief in CTs has been proposed to be a coping strategy able to reduce the stress elicited by uncertainty, we also expect an association between great levels of confidence and low levels of distress. We first checked for associations between

political distress at baseline (i.e., when uncertainty peaked) and: (i) perceptual stability on the one hand, and (ii) conspiracy endorsement on the other hand (**Figure 16-a**). Political distress was found to be negatively linked with both levels of inference (**p = .028, ρ = -.120** and **p = .007, ρ = -.094** respectively). We further confirmed these findings by splitting the sample into two subsamples according to stress: (i) a 'low stress' (**LS**) and (ii) a 'high stress' group (**HS**). Comparing these two groups at baseline, we confirmed a significant difference in both stability (**U = 41385, p = .002, Cohen's *d*= .140**) and GCB scores (**U = 43411, p = .023, Cohen's *d* = .110**), such as the LS group scored higher in both (**Figure 16-d,e**).

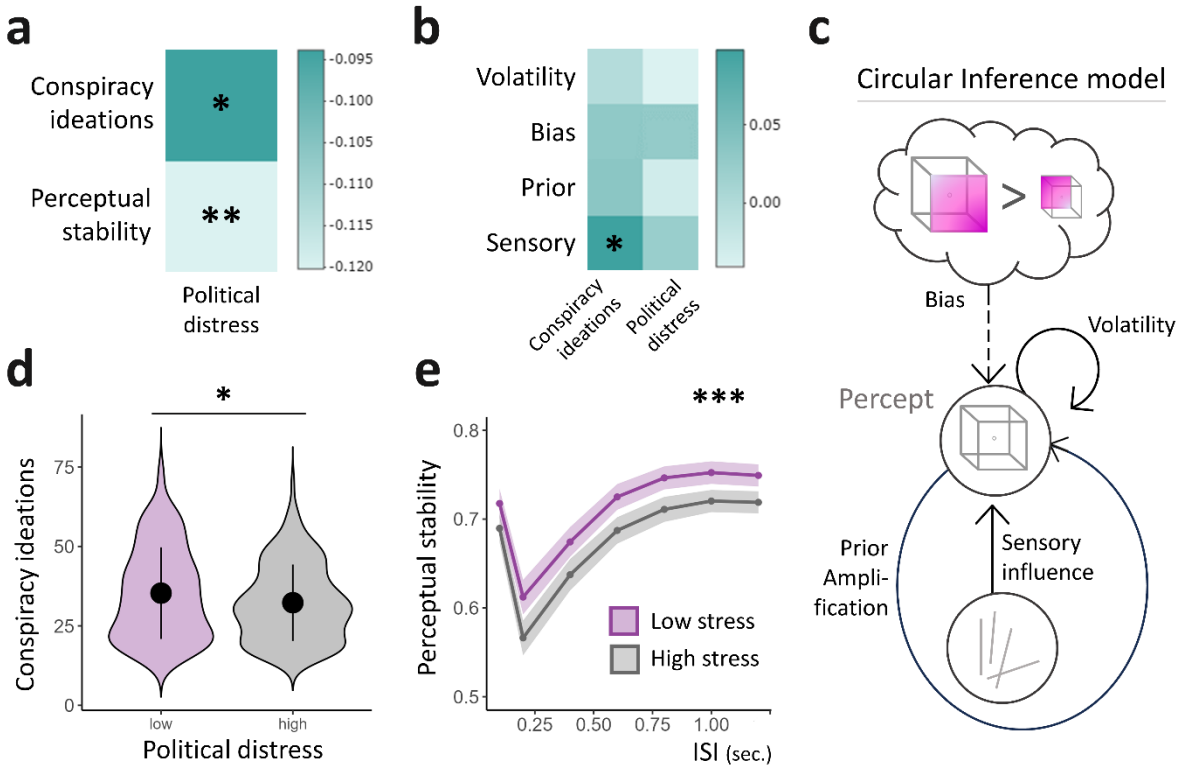


Figure 16 : Cognitive and perceptual inference correlates at baseline. **(a)** Heatmap depicting the strength of associations at baseline between political distress, conspiracy ideations measured with the *Generic Conspiracist Beliefs Scale* (GCB) and perceptual stability (Pearson's correlations, corrected for multiple comparisons using the *false discovery rate* method, FDR). Political distress was negatively associated with GCB **(ρ = -.094)** and perceptual stability scores **(ρ = -.120)**. **(b)** Heatmap illustrating the strength of associations at baseline between GCB scores, political distress and Circular Inference parameters (sensory weight (w), prior amplification (a), bias ($r_{on} - r_{off}$) and volatility (r_{on})). Pearson's correlations were corrected for multiple comparisons using FDR. GCB scores were significantly associated with sensory overweighting **(ρ = .098)**. **(c)** The Circular Inference (CI) model relies on 4 parameters: the overall sensory gain (sensory, w), the descending loops (prior, a), the transition rate (volatility, r_{on}) and the configuration preference (bias, ($r_{on} - r_{off}$)) (see *Methods*). **(d)** Mean conspiracy scores in the *low stress* (LS) group (n = 310, mean political distress = 2.10, s.d. = 2.06; mean conspiracy score = 35.3, s.d.= 14.4) and *high stress* (HS) group (n = 313, mean political distress = 8.01, s.d.= 1.58; mean conspiracy score = 32.3; s.d.= 12.0). GCB scores were significantly higher in the LS group (**Cohen's *d* = .110**). **(e)** Perceptual stability plotted as a function of inter-stimulus-interval (ISI) in LS (mean stability score = .597; s.d.= .176) and HS (stability score = .548; s.d.= .177) groups. Perceptual processing was found significantly more rigid in the LS group than in the HS group (**Cohen's *d* = .140**;). * stands for p<.05, ** for p<.01 and *** for p<.001.

We then looked at the influence of age, education, political distress and perceptual stability on GCB scores (**F(4,618) = 11.1, p <.001, adjusted R^2 = .061**). Again, we found that age (estimate = -.125, **p = .009**), and education (estimate = -1.88, **p <.001**) were significantly associated with CTs, further confirming that political distress (estimate = -.396, **p = .009**) had a significant impact on conspiracy endorsement, even after controlling for those sociodemographic factors.

III.2.5. Fitting the Circular Inference Model

Because we conceptualized perception as an inferential process, we also fitted the *Circular Inference* model to the Necker cube data (**Figure 16-c**). We have previously found that CI can explain NC data better than other pure Bayesian models (Leptourgos, Notredame, et al., 2020). This approach allowed us to quantify four model parameters contributing to the perceptual decision: sensory weight, prior amplification, bias, and volatility. We checked whether these CI parameters could capture the effects of political distress and conspiracy adherence. Sensory weight was the only parameter positively associated with GCB scores at baseline (**p = .030, ρ = .098, Figure 16-b**), supporting the idea that participants more prone to CTs at baseline rely more on sensory evidence when asked to make a decision in a highly ambiguous environment. We confirmed this GCB-sensory weight association (**estimate = 1.20, p = .051**) even after controlling for the effects of age, education and political distress (**F(4,618) = 11.86, p <.001, adjusted R^2 = .065**).

III.2.6 Measured changes after political event resolution

We then assessed changes in political distress, conspiracy ideations and perceptual stability over time (**Table 4**). We confirmed an overall stress reduction at T2 compared to that at baseline (**W = 100834, p <.001, Cohen's *d*= -.250** ; **Figure 17-a**), despite some heterogeneity in the participants. Meanwhile, GCB scores significantly increased (**W = 73048, p = .017, Cohen's *d* = .068**), while stability scores decreased (**W = 114427, p <.001, Cohen's *d* = -.139**) – this tendency toward destabilization was observed in each national sample (see also **Supplementary Figure 3**).

Table 4 : Population description at each time-step: scores, and CI parameters.

	pol. distress	GCB	stability	sensory	prior	bias	volatility
T1	5.35 ± 3.33	33.78 ± 13.33	.57 ± .18	1.70 ± .85	1.85 ± .91	.59 ± .07	-2.04 ± 1.50
T2	4.54 ± 3.16	34.70 ± 13.56	.55 ± .18	1.65 ± .85	1.86 ± .93	.59 ± .06	-2.20 ± 1.40

Pol. distress: political distress; GCB: Generic Conspiracist Beliefs Scale; stability: estimated stability score (see also *Methods section: Judgment criterion*); Sensory: sensory overweighting (w); Prior: prior amplification (a).

To account for the heterogeneity in stress evolution, we split the sample into two subgroups according to their trajectories: a first subsample with decreasing stress (**Dec**, n=330) and a second subsample with increased stress between T1 and T2 (**Inc**, n=227; **Figure 17-b**). Considering that the Dec group should have adopted the most efficient coping strategies, we checked how the CI parameters and degree of conspiracy ideations changed over the same period in these two subsamples (**Supplementary Figure 3**).

Figure 17 : Computational and cognitive features associated with changes in political distress over time. **(a)** Illustration of political distress scores over testing sessions. Political distress decreased between baseline (T1, mean political distress = 5.35, s.d. = 3.33) and retest (T2, mean political distress = 4.54, s.d. = 3.16; **Cohen's *d* = -.250**). **(b)** Political distress over time according to stress trajectories. Some participants showed decreased political distress after uncertainty resolution (Dec, n = 330, mean Δpolitical distress = -2.78, s.d. = 2.45), while another subsample showed increased stress (Inc, n = 220, mean Δpolitical distress = 2.01, s.d. = 2.14). **(c)** Left: Illustration of the evolution of prior amplification (ΔPrior) over time for the 'Dec' (mean = .081, s.d. = 1.11) and 'Inc' (mean = -.121, s.d. = 1.14) groups. The intergroup difference was significant (**Cohen's *d* = .180**). Right: Evolution of conspiracy ideations (ΔConspiracy ideations) over time for the 'Dec' (mean = 1.29, s.d. = 10.3) and 'Inc' (mean = -.305, s.d. = 11.6) groups. The intergroup difference was significant (**Cohen's *d* = .145**) **(d)** Scatter plot showing the correlation between ΔPrior and ΔConspiracy ideations in the 'Dec' group (p = .035, ρ = .116; Spearman correlation, **ρ = .116**). * indicates p<.05, *** indicates p<.001.

A delta measure for each CI parameter was computed (parameter value at retest minus value at baseline), such as a positive delta indicated a gain in the parameter value, while a negative delta reflected a decrease in this parameter. The Dec group showed increased reliance on prior information in the bistable task between T1 and T2 (mean ΔPrior=.0811, s.d=1.11), while the Inc group showed decreased use of priors in the same period (mean ΔPrior=-.121, s.d.=1.14). This difference was statistically significant (**t(460.65) = 2.07, p = .039, Cohen's *d* = .180** ; **Figure 17-c-left**). We found no differences in the 3 other CI parameters (**Supplementary Figure 3**).

We also computed a composite ΔGCB score corresponding to GCB at retest minus GCB at baseline, such that a positive delta corresponded to an increase in conspiracy adherence while a negative delta resulted in a decrease. We observed a trend for conspiracy strengthening in participants with decreased stress in comparison with that observed in the rest of the sample (t(429.52)=1.65, p = .099, Cohen's *d* = .145). Because conspiracy ideations were proposed to act as a coping mechanism when facing uncertainty, we next ran an oriented test to confront that hypothesis which reached significance (**t(429.52)=1.65, p = .050, Cohen's *d* = 0.145 ; Figure 17-c-right**). This finding supports a gain in the GCB score for the Dec group compared to the Inc group. To confirm the idea that the GCB score increase was directly associated with an increase in Prior in the Dec group, we compared ΔPrior and ΔGCB in this specific subsample; these measures were found to be positively associated (**p = .035, ρ = .116, Figure 17-d**).

III.3. Discussion

A surge in CTs has been observed in recent years, and CTs have been proposed to act as coping strategies for the stress and perceived lack of control generated by global uncertainty (Dow et al., 2022; Farias and Pilati, 2023; Sullivan et al., 2010; van Prooijen and Acker, 2015; Whitson and Galinsky, 2008). CTs offer intuitive and easy-to-understand explanations to unsolved problems (van Prooijen, 2017). Links have already been established between conspiracy endorsement and some inference biases (Brotherton and French, 2015; Drinkwater et al., 2012; Georgiou et al., 2021; Wycha, 2015). However, very few studies have primarily focused on low-level perceptual aspects of conspiracy (Dagnall et al., 2015; Hartmann and Müller, 2023; Müller and Hartmann, 2023; van Prooijen et al., 2018), and limited efforts

have been made to delve into the potential mechanisms of information processing that may convey such associations.

To address these concerns, we combined online assessments of bistable perception in large international samples with Bayesian modeling. This approach allowed us to quantify perceptual inference mechanisms and to test their links with conspiracy ideations during periods of great sociopolitical uncertainty. We were able to capture the rigidification of conspiracy beliefs in nonclinical populations. Specifically, using the Circular Inference (CI) model, we highlighted a significant association between conspiracy endorsement and the overweighting of sensory information when ambiguity reaches a climax, later followed by a selective increase in prior reliance in those who subsequently decreased their stress levels.

Several attempts at modeling the features of conspiracy beliefs can be found in the literature. However, most of these models have either focused on the network scale (Peruzzi et al., 2019) or remained purely theoretical, without experimental testing (Rigoli, 2022). Recent findings highlighted the added value of a computational framework to account for the emergence of rigid beliefs during the COVID-19 pandemic (Suthaharan et al., 2021) and the protective aspect of CTs against distress in a social context (Suthaharan and Corlett, 2023). These studies used high-level cognitive tasks and mainly focused on paranoia, a condition sharing some phenomenological features with CTs but also considered significantly different (Greenburgh and Raihani, 2022), further justifying specific explorations. The quantitative approach proposed in the present work nicely completes these initiatives, adding the testing of low-level inference, together to measurements of conspiracy beliefs' emergence and rigidification.

Here, we provide the first evidence for an association between sensory information overweighting in ambiguous contexts and a high level of conspiracy endorsement. This finding suggests that when uncertainty peaks, a subpart of the population, more vulnerable to stress, is prone to embracing conspiracy explanations based on intuitive reasoning. Motivated by the need to cope with uncertainty, these participants first adopt an "exploration" strategy, seeking explanations in their direct environment to make their perceptual decisions. Interestingly, such a mechanism accounts for perceptual and inferential biases previously found to be associated with conspiracy ideations, such as illusory pattern detection (Hartmann and Müller, 2023;

Müller and Hartmann, 2023; van Prooijen et al., 2018), intuitive thinking (Binnendyk and Pennycook, 2022; Swami et al., 2014) and the JTC phenomenon (Kabengele et al., 2023; Pytlik et al., 2020).

We also explored the dynamic changes in model parameters after stress resolution by using a pre/post design surrounding the political events. We shed light on the association between prior knowledge amplification in perceptual decisions and the enhanced adherence to CTs in those who showed reduced stress level. This finding suggests that some participants coped with uncertainty by embracing conspiracy-oriented explanations, secondarily shifting to an "exploitation" strategy (**Supplementary Figure 5**), validating their newly established view and reinforcing their own beliefs. This second mode appears compatible with findings showing confirmation biases (Brotherton, 2015) and reality testing deficits (Lewandowsky et al., 2013) in people with CTs, making these beliefs more resilient to counterevidence.

These results can also be compared with models of the emergence and maintenance of clinical beliefs, such as delusional ideations. Indeed, prior research conceptualized delusion formation as the result of impaired associative learning processes driven by excessive prediction error (Corlett et al., 2007), a framework that was later extended to account for delusion persistence as aberrant reinforcement of previously leaned associations (Corlett et al., 2009). Our results also add to previous work showing that parametric changes might mimic behaviors observed during the transition to psychosis (Denève and Jardri, 2016). It was shown using CI-based simulations that the seminal amplification of sensory information involved in the integration of aberrant causal relationships (during the transition to psychosis) subsequently constituted strong priors proposed as responsible for the stability of delusional contents from one psychotic episode to the next. Both approaches (predictive coding and Bayesian modeling) are congruent with (i) the idea that conspiracy endorsement is associated with the establishment of aberrant causal relations between random events (Whitson and Galinsky, 2008), and (ii) that conspiracy could be rooted in the self-reinforcement of previously integrated suboptimal beliefs.

While the endorsement of CTs may serve as an effective short-term coping strategy, it also appears to pave the way for the long-term rigidification of suboptimal beliefs (beliefs that would be computed through mechanisms deviating from Bayes'

rule), making it maladaptive for stress-regulation overall. The social implications of gaining a better understanding of this phenomenon are vast. Humankind has experienced repeated periods of heightened uncertainty throughout history, ranging from civilizational collapses or wars to economic crises. In extending the well-established association between political distress and the endorsement of CTs (van Prooijen and Douglas, 2017), our model also explains the recent rise in extremism and populism observed since the beginning of the XXIst century in a global context of the pandemic, terror attacks and climate change.

We must acknowledge some limitations of this work. First, although significant, some results exhibit small effect-sizes (i.e., Cohen's *d* around 0.2). Of note, small effect-sizes were previously found to still have substantial significance when studies were conducted on large populations (McNeish and Stapleton, 2016). It is also important to remember that small effects were expected because we attempted to capture an association between a low-level inference process (bistable perception) and a more complex cognitive process (conspiracy). These findings still constitute an important proof-of-concept demonstration that the CI model can capture small variations in nonclinical populations' perceptual decisions, paving the way for promising advancements in deepening our understanding of the mechanisms underlying belief rigidification.

A second limitation is that we cannot rule out that some participants may have felt hesitant in honestly reporting their views about CTs, due to the controversy and potential stigma surrounding conspiracy thinking. However, we think that our experimental design offers two advantages in the valid assessment of conspiracy endorsement. First, its online nature ensured anonymity and encouraged freedom of speech, as frequently observed on the internet and digital social media. Second, the joint use of a low-level perceptual task, the NC, provided access to a proxy of inference processing that is rarely prone to social biases, such as interviewer compliance.

A third limitation is the representativity of the sample: we indeed chose to recruit participants from three Western educated countries, known for their high degree of polarization (Fletcher et al., 2020). Although our sample may not represent the world population, we argue that the phenomenon we are investigating follows some universal rules. First, links between sociopolitical uncertainty and the resurgence of conspiracy beliefs have already been observed at various times and locations, dating back to the

Roman Empire (Boddington, 1963). Second, while the GCB total scores were distributed differently among our three samples (**Supplementary Figure 1-a**), their qualitative distribution across GCB subscales followed the same pattern (**Supplementary Figure 1-b**). Third, the main results and trends (i.e., sometimes not reaching significance due to the reduced statistical power) appear consistent across the 3 samples when tested separately (**Supplementary Figure 2**).

For the same reasons, we focused on the level of distress related to specific political events in the countries where the tests took place. Importantly, we did not consider other types of individual stress levels. Instead, we concentrated on the broader phenomenon of sociopolitical uncertainty.

Overall, this study highlights the potential of the Circular Inference model in examining subtle variations in inference processing associated with high-level cognitive beliefs. This model has already proven effective in accounting for the positive symptoms of schizophrenia (Jardri et al., 2017; Leptourgos et al., 2017; Simonsen et al., 2021) and schizotypal traits (Derome et al., 2023); however, this breakthrough opens up new avenues for applying quantitative approaches to dynamically explore subjective beliefs in nonclinical populations. By applying this computational framework, we delved deeper into the mechanisms underlying the emergence and maintenance of conspiracy beliefs, shedding light on their societal impact and providing insights that could be valuable for developing interventions aimed to counter the influence of CTs during highly uncertain periods.

III.4. Methods

III.4.1. Participants

Three independent samples were recruited using the Prolific© web-platform: 212 US citizens, 225 British citizens and 186 French citizens. The same protocol was administered 1 month before and 1 month after a major stressful political event: the 2020 US presidential election, the 2021 UK BREXIT implementation and the 2022 French presidential election (**Figure 1**). The targeted participants were aged between 18 and 60 and had normal-to-corrected vision. They were from the nationality of the country of interest for each sample and regularly used social media. Th exclusion criteria were a history of psychiatric or neurological disorder, strabismus, or eye

surgery. From the initial sample (N = 755), 30 participants were excluded based on failed attentional checks (see *Supplementary Material section: Controlling for experimental biases*) or very-low reaction times (mean reaction time < 300ms), while 102 were lost longitudinally.

The Prolific© web-platform (https://www.prolific.co/) ensures data privacy following standards of the European and UK data protection law (i.e., General Data Protection Regulation (GDPR), transposed into UK law as the UK GDPR). Participants' sociodemographic characteristics were associated with their respective behavioral data through an anonymous ID randomly assigned at enrollment. The overall online survey complies with French regulations and ethics (*Comité de Protection des Personnes Nord-Ouest IV*).

III.4.2. Apparatus

The protocol was implemented in PsychoPy v.3, exported and hosted online on the Pavlovia.org platform. For the perceptual part of the experiment, participants were instructed to stand in total darkness, approximately 60 cm away from the screen and adjust it to be perpendicular to the floor with their eyes aligned to the fixation cross displayed at the center of the screen. The NC task and the self-reported assessment of beliefs were administered in a randomized order (see also *Supplementary Material section: Controlling for experimental biases*).

III.4.3. The Necker Cube Task

Stimuli:

Visual stimuli representing Necker cubes (NC) were displayed in the center of a black screen. The stimulus size was standardized across the participants using a matching method based on a standard credit card displayed on the screen that the participant was required to adjust in size before starting the experiment.

Procedure:

The block-design of the task was inspired by Mamassian and Goutcher's (Mamassian and Goutcher, 2005) protocol. During each block, a NC was presented discontinuously. Referring to a forced-choice methodology, we asked participants to report their interpretation of the stimulus using their keyboard each time a new cube

appeared on the screen. The cube disappeared after a pseudorandom duration (**ISI** ranging from 0.1 to 1.2 seconds). Each recorded response constituted a trial, and the experiment was divided into 10 blocks of 64 consecutive trials (i.e., 640 NC presentations per run), providing a discontinuous sample of the participant's perceptual dynamics. A 10-second black screen display separated each block to minimize the influence of the previous block on later responses (**Figure 14-a**).

Participants were instructed to stare at the target located in the middle of the screen to neutralize the potential effects of eye movements. The two possible interpretations of the NC (SFA, SFB) were explicitly mentioned, and subjects were asked to look at the cube passively, without attempting to orient or force their perception. A short training session was performed beforehand to give participants the opportunity to become familiar with the stimulus and the task while ensuring that the instructions were well understood.

Judgment criterion

Various parameters can be used to understand and describe the phenomenon of bistable perception. We chose to focus on **perceptual stability** because we were interested in its dynamical dimension, i.e., how the system could stabilise and destabilise.

Perceptual stability is defined as the probability that a percept persists from one trial to the next. According to Markovian modeling, the current percept (one of the two interpretations SFA or SFB) depends on the previous percept and its updating by sensory observation. This implies a circularity in the integration of information where the percept at time *t* becomes the prior information at time *t+1*. A value was thus assigned to each trial "i": 0 if the response was different from the response to trial "i-1" and 1 if the response to trial "i" was identical to the response to trial "i-1". The average SP was thus calculated for all trials and separately for each interpretation (SP0 and SP1 for SFA and SFB, respectively). Overall, the SP was interpreted as the general probability that the system remains stable from one trial to the next, where 1 corresponds to a system with no perceptual change and 0 to a system governed by maximum instability.

A previously proposed way to assess perceptual stability is by computing stability curves representing SP as a function of different ISI values. Such a curve usually consists of an initial “destabilization” portion corresponding to a drastic drop in

perceptual stability, and a “stabilization” portion reaching a “ceiling threshold”, considered a good proxy of perceptual stability (**Figure 14-b,d**). This second portion of the curve was fitted to a reversed exponential function, and we considered the parameter corresponding to the last point of the curve as the **stability score** for each participant.

III.4.4. Self-reported measures

A sociodemographic form and some psychometric assessments were then conducted/collected on the Prolific© platform. Participants specified their age and educational attainment as defined in the *International Standard Classification of Education* (**ISCED**)(UNESCO Institute for Statistics and Statistics, 2020). The participant demographics are shown in **Table 3**. When Likert or visual analogical scales were used, the cursor was coded to return to the center of the screen after each question to avoid the answer being biased by previous ones. Adherence to CTs was assessed using the 15-item *Generic Conspiracist Beliefs Scale* (**GCB**)(Brotherton, C French, et al., 2013) and its French translation (Lantian et al., 2016). The GCB scores and subscores for each sample are shown in **Table 3** and **Supplementary Table 1**. Participants were also asked to rate with a 10-point visual analogical scale how distressed they were regarding the target event in their country (**political distress**). The precise questions used are shown in the *Supplementary Material section: Self-reported measures*.

III.4.5. Data Analysis and Statistics

Characteristics of conspiracy adherence

The normality of the distributions was tested using the Shapiro-Wilk test. If nonnormally distributed, further analyses were run using nonparametric statistics. Notably, we compared GCB scores between the three US-UK-FR samples using a Kruskal-Wallis test. We compared GCB scores between males and females using a Mann-Whitney test, while GCB scores across ISCED levels of education and across different age levels were compared using Welch ANOVAs.

The correlates of stress at baseline

We conducted a series of model-free analyses to confirm the association between political distress, stability score, and GCB. Again, due to the non-normal distribution of the GCB scores, we referred to Spearman rank correlations to explore

linear associations, corrected for multiple comparisons based on the *false discovery rate* (FDR) method. These analyses were conducted on the whole sample, and on subsamples generated through a median split on the political distress score: the 'low stress' (**LS**, n=310) and 'high stress' (**HS**, n=313) subgroups. We used Mann-Whitney tests to assess the difference between these two subgroups regarding stability scores or GCB scores. We also used a linear regression model to confirm the association between political distress and GCB, adding age and education level as covariates to control for the effect of these sociodemographic factors.

Model fitting

To better understand the association between conspiracy theories and stress, we fitted a dynamical *Circular Inference* model (CI) to the data (for more details, see (Leptourgos, Bouttier, et al., 2020)). Applied to the NC task, CI describes the process through which participants combine prior expectations about the visual appearance of three-dimensional (3D) objects and (illusory) depth cues to compute a 3D interpretation of the two-dimensional (2D) NC : seen from above (SFA) or seen from below (SFB). Belief updating in CI can be formalized as follows:

$$\frac{dL}{dt} = -\Phi(L) + aL + wS$$

This equation describes how the posterior belief about the ambiguous figure L changes over time (positive/negative L corresponds to SFA/SFB beliefs), under the influence of 3 driving "forces": dynamics ($\Phi(L)$), descending loops (aL) and sensory noise (wS).

Function $\Phi(\ )$ describes the (Markovian) dynamics of the system and is equivalent to a leak term. It captures the intuition that in the real-world, objects are not eternal and can appear, disappear or change abruptly. Markovian temporal statistics can be reduced to 2 parameters, r_{on} and r_{off} (corresponding to the probability of switching from SFB to SFA and from SFA to SFB respectively). This term pushes L toward its prior value ($\log\left(\frac{r_{on}}{r_{off}}\right)$). By making r_{on} greater than r_{off}, we can implement an implicit SFA bias.

The second term describes the auto-amplification of priors due to descending loops (parameter a). According to CI, prior information can be reverberated and

counted several times (Jardri and Denève, 2013). This overcounting of priors is akin to a positive feedback that strengthens and stabilizes currently held perceptual beliefs, resulting in bistable perception (Leptourgos, Notredame, et al., 2020).

Finally, the third term describes the sensory noise that drives switches between the 2 interpretations. For simplicity, we assume that S is sampled from a normal distribution with 0 mean and variance equal to 1. Furthermore, w is a free parameter representing the overall sensory weight (sensory weight and climbing loops are mathematically indistinguishable, so they are both included in w).

In summary, this model of perceptual dynamics can be reduced to 4 free parameters: the overall gain of sensory inputs w (sensory), the descending loops a (prior), the transition rate r_{on} (volatility) and the bias ($r_{on} - r_{off}$).

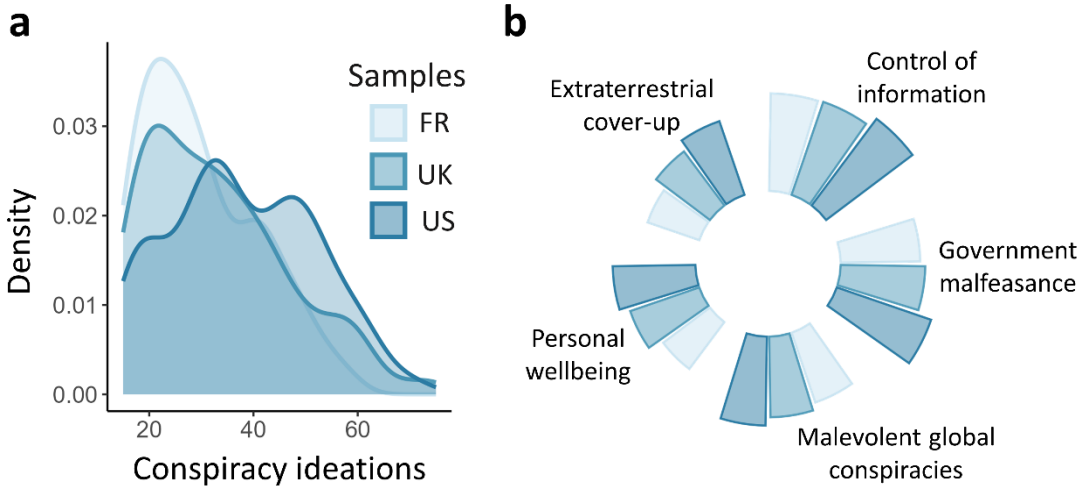
Changes after political event resolution

We assessed the evolution of political distress, stability scores and GCB scores over time using Wilcoxon signed-rank tests for repeated measures. We then split our sample into two groups: **Dec** and **Inc** comprising individuals who showed decreased or increased stress, respectively, between the two time points. We computed a delta measure for each parameter that corresponded to the parameter's value at retest minus that at baseline. A positive value indicated a gain in the parameter, while a negative value indicated a decrease. Due to the normal shape of distributions in these composite scores and our sample size, we referred to Welch tests for group comparisons.

The same procedure was used to compare the two groups regarding the gain in GCB (ΔGCB). We successively performed a two-tailed Welch's test, followed by Welch's test for the oriented hypothesis that the **Dec** subsample would significantly increase its GCB score compared with the **Inc** subsample. Finally, a Pearson correlation test was used to check for an association between ΔAlpha and ΔGCB in the **Dec** subgroup.

III.5. Supplementary material

III.5.1. Sociodemographic features of conspiracy theories



Supplementary Figure 1 : Conspiracy beliefs across tested populations. (a) Distribution of GCB scores across each national sample i.e., US, *United States of America*; UK, *United Kingdom*; FR, *France*. **(b)** National samples displayed a similar pattern across GCB subscales distribution (control of information, government malfeasance, malevolent global conspiracies, personal well-being and extraterrestrial cover-up).

Supplementary Table 1 : Conspiracy ideations in tested populations at baseline.

	Control of information	Government malfeasance	Malevolent global conspiracies	Personal well-being	Extraterrestrial cover-up
Whole Sample (n = 623)	8.35 ± 3.10	7.28 ± 3.19	6.69 ± 3.22	5.94 ± 2.85	5.53 ± 3.01
US (n = 212)	8.92 ± 3.12	8.23 ± 3.27	7.31 ± 3.30	6.82 ± 3.21	6.40 ± 3.21
UK (n = 225)	8.08 ± 3.05	6.98 ± 3.21	6.63 ± 3.40	6.02 ± 2.86	5.56 ± 3.06
FR (n = 186)	8.02 ± 3.07	6.56 ± 2.82	6.06 ± 2.78	4.82 ± 2.20	4.51 ± 2.33

US: United States of America; UK: United Kingdom; FR: France; columns correspond to the 5 subscales of the *Generic Conspiracist Beliefs Scale*.

As mentioned in the *Results* section, we replicated previous findings showing that conspiracy ideations are not normally distributed across nonclinical populations (Bronstein et al., 2022), suggesting that this type of belief is not commonly endorsed by most of the population (**Supplementary Figure 1**). Of note, the scale we used assesses a general degree of adherence to CTs, suggesting that even if a large part of the population could deem certain CTs believable to a certain degree, rigidity of adherence to such beliefs is an isolated phenomenon only represented by an extreme fringe of the tested sample.

Conversely, we did not replicate previous findings suggesting that males were more prone to endorsing CTs (Freeman and Bentall, 2017). We compared GCB scores between males and females using a Mann-Whitney test and found no significant difference (W = 47832, p = .868, Cohen's *d* = .0017, **Figure 15-b**). Of note, six participants preferred not to specify their sex and were excluded from this specific analysis (this sample was too small to be considered in itself).

Congruent with the literature, we found that a higher degree of conspiracy endorsement is associated with lower educational attainment. We compared GCB scores according to ISCED levels of education using Welch ANOVA and found a significant difference between the *low*, *medium* and *high* education groups (**F(13.477) = 13.477, p <.001, η^2=.395 e^{-05}**, **Figure 15-d**). This effect could be partially explained by multiple intertwined factors associated with education, such as analytical thinking and belief in simple solutions for complex problems(van Prooijen, 2017).

Similarly, we observed an association between a higher degree of conspiracy endorsement and younger age using Welch ANOVA (**F(3.1015) = 3.1015, p = .046, η^2=.039**, **Figure 15-c**). Initially, we speculated that this effect might be driven by the educational factor mentioned earlier. However, the linear model testing for the age x education interaction only showed a trending variation in GCB scores. Therefore, we argue that education cannot fully account for this relationship, particularly as it does not explain the observed decrease in conspiracy endorsement after 30 years of age, when individuals typically stop pursuing institutional education. This finding contradicts previous results showing a positive relationship between age and belief in conspiracy theories (Romer and Jamieson, 2020). This discrepancy might be due to a difference in methodology. While the aforementioned authors measured a percentage of believers in specific COVID-19-related CTs in different age groups, we evaluated a general degree of adherence to CTs. Thus, while the tendency to deem health-related CTs believable might increase with age – perhaps due to an increased feeling of health-related threat that have been found associated with conspiracy ideations (Federico and Malka, 2018) – there might be a general tendency for CT adherence to decline over the life course. In particular, this interpretation is in line with existing literature involving a similar approach to measuring conspiracy beliefs (Wagner-Egger et al., 2022).

III.5.2 Controlling for experimental design biases

Pilot data

We tested the validity of the online perceptual stability test before the main experiment. An independent sample of 16 participants performed the NC task twice: (i) online and (ii) supervised in a laboratory setting. The order between these two conditions was counterbalanced. Details about the online administration of the task can be found in the *Methods* section. We kept the same design for the in-lab version. We ensured that participants were installed 60 cm from the screen and that their eyes were aligned with the middle of the screen using a chin-strap. They were placed in the dark and received the same instructions as the online sample but were told that they could ask the investigator for further explanations if the instructions were unclear.

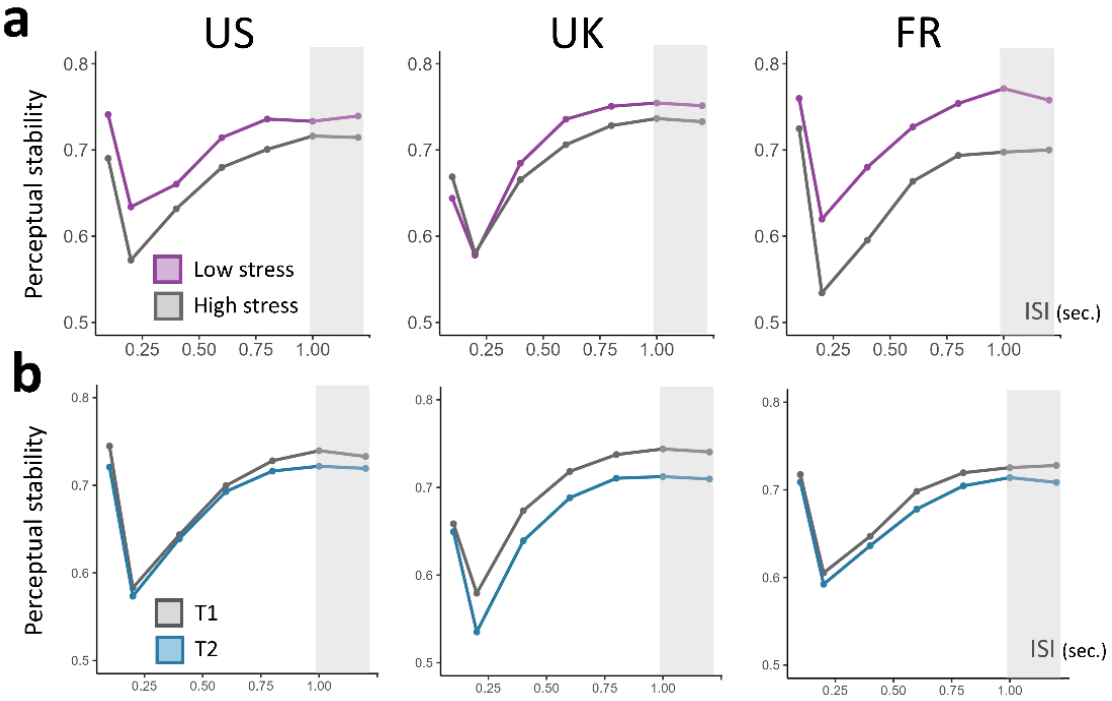
We checked for differences in stability scores between the two methods using a Wilcoxon signed-rank test for repeated measures and found no significant difference ($W = 92$, $p = .231$, **Figure 14-d**). We then assessed the range of agreement between the two methods using a Bland Altman test and found a nonsignificant mean-difference bias of .0586 (**Figure 14-e**). All individual differences were contained in the *limits of agreement* ([-.300 ; .417]), supporting a good agreement between online and in-lab stability measures in the same participants.

To control for a potential training effect on the perceptual task induced by this pilot repeated-measures design, we checked for changes in stability scores between the first and second assessment in the pilot sample using a Wilcoxon signed rank test for repeated measures. The values at baseline (mean stability score = .474, s.d. = .179) did not differ from those at retest (mean stability score = .467, s.d. = .160 ; $W = 67$, $p = .980$) suggesting that any test-retest difference further observed cannot be attributed to training effects.

We added attentional checks during the psychometric assessment to ensure participants did not provide random answers. Among the scales used, five items were randomly added, regularly asking participants for a specific answer (example: "*This is an attentional check, please answer 'Not sure/cannot decide' to that question.*").

Experimental samples

To further strengthen the validity of our method, we compared stability scores at baseline between the three national samples (US, UK and FR) using Welch's ANOVA and found no significant difference ($F(2,620) = .81828$, $p = .4419$). Furthermore, we observed the same patterns of association between (i) political distress at baseline and perceptual rigidity (**Supplementary Figure 2-a**) and (ii) perceptual stability decrease between the first and second measurement (**Figure 14-c, Supplementary Figure 2-b**). Finally, we investigated whether the uncertainty induced by the bistable task could influence adherence to conspiracy beliefs, or whether activation of conspiracy ideations prior to the behavioral task could affect perceptual stability. The NC task and the self-reported assessment of beliefs were administered in a randomized order to control for those biases. We compared both groups of participants randomly assigned to an order using Mann-Whitney tests and found no difference in perceptual stability ($U = 47550$, $p = .767$) or GCB scores ($U = 45526$, $p = .230$).



Supplementary Figure 2 : Perceptual stability in each national sample. (a) Perceptual stability as a function of interstimulus interval (ISI) in *the low stress* (LS) and *high stress* (HS) groups across the 3 samples (US: *United States of America*; UK: *United Kingdom* and FR: *France*). **(b)** Perceptual stability as a function of interstimulus interval length over time across the 3 samples (T1: baseline; T2: retest).

To ensure that this phenomenon of global destabilization was specifically associated with the dynamics of sociopolitical uncertainty, we tested stability a third time on the UK and FR samples one month after the second acquisition (**Figure 14-c**). We compared stability at baseline and during the third test using Wilcoxon signed-rank tests for repeated measures. We found no significant difference whether we tested the whole sample ($W = 30618$, $p = .141$) or the 3 national subsamples independently (UK: $W = 9952$, $p = .247$; FR: $W = 5729$, $p = .137$), excluding a possible learning effect across the sessions.

III.5.3. Self-reported measures

At each time step, participants were instructed to rate their level of distress related to the ongoing event in their country using 10-point visual analog scales. Political distress scores were obtained by computing the mean of these ratings. We asked the following questions:

Political distress assessment at baseline:

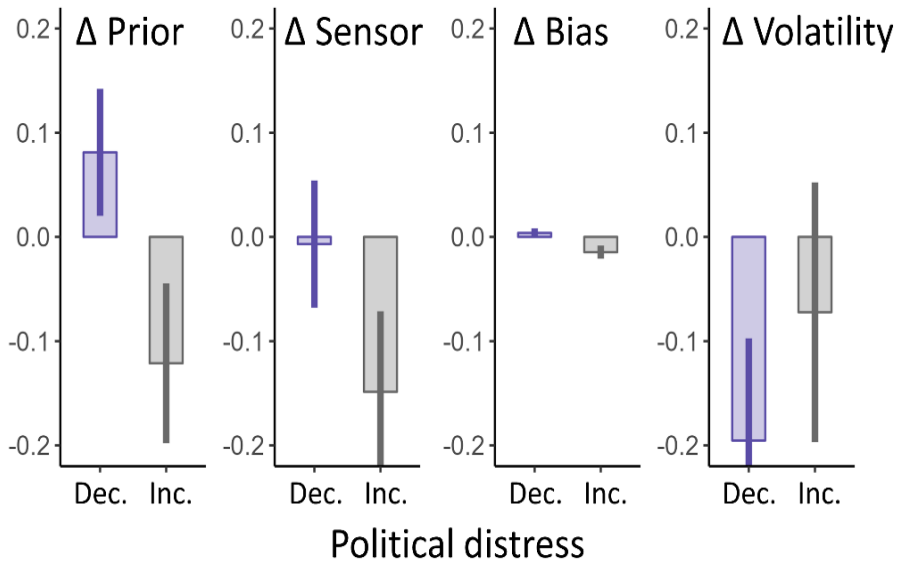
- US : How distressed were you during the week preceding the 2020 presidential elections?
- UK : How distressed are you since the announcement of Brexit?
- UK : How distressed are you regarding the establishment of BREXIT that will come to pass?
- UK : How distressed are you regarding the consequences of Brexit?
- UK : How distressed are you regarding the possibility of a no-deal?
- FR : À quel point êtes vous stressé(e) par l'approche des élections présidentielles ?
- FR : À quel point êtes vous stressé(e) par les conséquences de l'élection présidentielle à venir ?

Political distress assessment at retest:

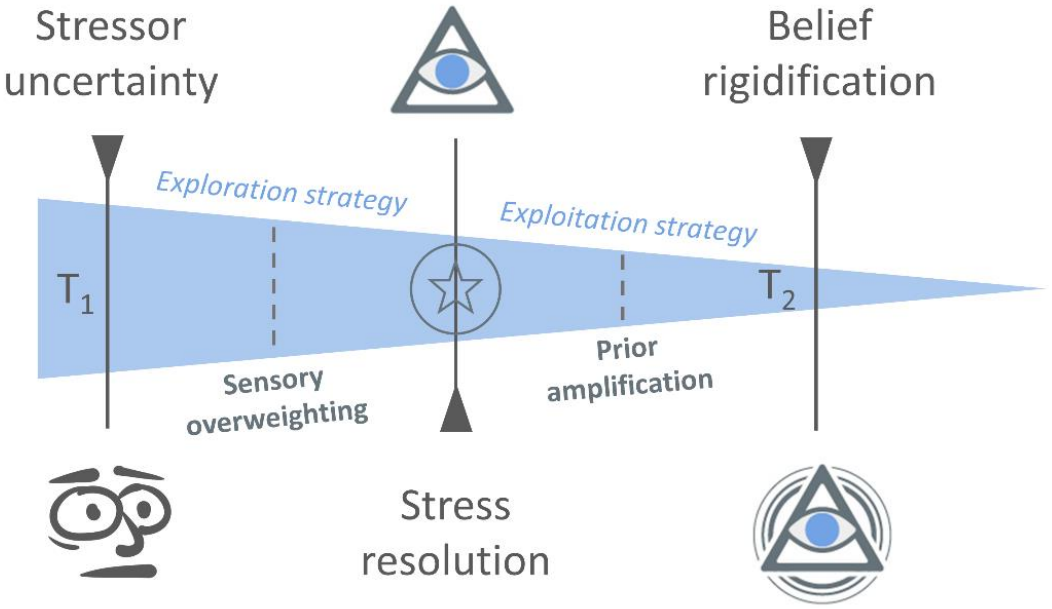
- US : How distressed are you since the announcement of the 2020 presidential election outcome in the media?
- UK : How distressed are you since the UK left the EU?
- UK : How distressed are you regarding the consequences of Brexit?
- FR : À quel point êtes vous stressé(e) par le résultat des élections présidentielles ?
- FR : À quel point êtes vous stressé(e) par les conséquences de l'élection présidentielle ?

III.5.4. Supplementary figures

Supplementary Figure 3 : Evolution of the Circular Inference parameters over time according to stress dynamics.



Some participants demonstrated decreased political distress after uncertainty resolution (Dec, n = 330) while others showed increased stress in the same period (Inc, n = 330). From left to right: evolution of prior amplification (ΔPrior) over time for 'Dec' group (mean = .081, s.d. = 1.11) and 'Inc' group (mean = -.121, s.d. = 1.14); evolution of sensory overweighting (ΔSensor) over time for the 'Dec' group (mean = -.007, s.d. = 1.11) and 'Inc' group (mean = -.149, s.d. = 1.15); evolution of bias (ΔBias) over time for the 'Dec' group (mean = .004, s.d. = .079) and 'Inc' group (mean = -.015, s.d. = .093); evolution of volatility (ΔVolatility) over time for the 'Dec' group (mean = .195, s.d. = 1.78) and 'Inc' group (mean = -.072, s.d. = 1.85).



Supplementary Figure 4 : The shift in the balance of Circular Inference parameters is associated with the evolution of conspiracy ideations after the resolution of uncertainty. A possible hypothesis is that, motivated by the need to cope with uncertainty, distressed participants first adopt an "exploration" strategy, seeking simple and intuitive explanations in their environment to make their perceptual decisions; they secondarily shift to an "exploitation" strategy, validating their newly established view and reinforcing their own beliefs.

III.5.6. Supplementary table

TABLE S3. Conspiracy and Circular Inference parameters over time according to stress dynamics.

	ΔGCB	**Δsensory**	**Δprior**	**Δbias**	**Δvolatility**
Dec (n = 330)	1.29 ± 10.28	-.01 ± 1.11	.08 ± 1.11	.004 ± .08	.19 ± 1.78
Inc (n = 220)	-.31 ± 11.61	-.15 ± 1.15	-.12 ± 1.14	-.02 ± .09	-.07 ± 1.85

Dec, 'Dec' group corresponding to participants who decreased their stress over time; Inc: 'Inc' group corresponding to participants who increased their stress over time; ΔGCB: evolution in *Generic Conspiracist Beliefs Scale* over time; Δsensory: evolution of sensory overweighting over time; Δprior: evolution of prior amplification over time; Δvolatility: evolution of volatility over time; Δbias: evolution of bias over time. (see also *Methods*).

Chapter IV

Extreme beliefs in online community can be corrected by Circular Belief propagation[4]

Highlights

In recent decades, the massification of online social connections has made information globally accessible in a matter of seconds. Unfortunately, this has been accompanied by a dramatic surge in extreme opinions, without a clear solution in sight. Using a model performing probabilistic inference in large-scale loopy graphs through exchange of messages between nodes, we show how circularity in the social graph directly leads to radicalization and the polarization of opinions. We demonstrate that these detrimental effects could be avoided by actively decorrelating the messages in social media feeds. This approach is based on an extension of Belief Propagation (BP) named Circular Belief Propagation (CBP) that can be trained to drastically improve inference within a cyclic graph. CBP was benchmarked using data from Facebook© and Twitter©. This approach could inspire new methods for preventing the viral spreading and amplification of misinformation online, improving the capacity of social networks to share knowledge globally without resorting to censorship.

[4] Bouttier V., Leclercq S., Jardri R., Denève S., (*submitted*). A normative approach to radicalization in social networks (preprint available at the following DOI : 10.48550/arXiv.2309.00513).

IV.1. Introduction

Online social networks have great benefits and advantages. They allow for the quasi-instantaneous exchange of up-to-date information and give access to persons around the world with different backgrounds, experiences and opinions. They also create communities with sizes well beyond the usual social constraints, and perhaps even beyond cognitive ones (Dunbar, 2016). Nevertheless, the constant increase in network size and complexity may introduce more information than we can normally process (Rodriguez et al., 2014), as well as promoting passionate (and sometimes extreme) debates. Beyond the initial excitement these networks provided, the regular polarization of positions on social media appears worrisome. For example, it promotes severe conflicts between communities expressing opposite beliefs, while also making social networks particularly vulnerable to manipulation or propaganda, for instance, by bots accused of interference with presidential elections (Ferrara, 2020).

Solutions need to be found, but without sacrificing the advantages of worldwide information access or impoverishing social interactions. In our view, the problem goes far beyond the propagation of fake news, which is a symptom as much as a cause of polarization. More than the content of one's belief, the issue seems to revolve around overconfidence and excessive trust (or distrust) in information confirming (or contradicting) these beliefs. Many of the most polarizing issues discussed on the internet may not even have a universally defined, knowable, or absolute answer (this is the case for societal questions such as immigration policies but also questions beyond these such as the existence of extraterrestrial intelligence). For these issues, *radicalization* can be defined as people reaching unreasonably confident and monolithic beliefs based on multifaceted, biased or untrustworthy data (Cinelli et al., 2021). Additionally, the emergence of two or more radicalized groups with opposite, irreconcilable beliefs results in *polarization*.

To capture these phenomena in a simplified, mathematically grounded but intuitive framework, we treat large-scale opinion sharing in social networks as a form of probabilistic inference. People's beliefs are modeled as the probability of giving an answer to a particular question (e.g. *Should abortion be legal or not?*). Rather than just deciding "yes" or "no" once for all (a binary choice), someone could have a graded confidence level represented with a probability, close to 100% or 0% for high confidence or equivalently strong opinions but approaching 50% if the person is

uncertain. Agents embedded in a social network derive their beliefs both from external or private sources of evidence (direct experience, expertise, news articles, religious values, etc.) and from the expressed opinions of people they are connected to or communicate with (see **Figure 18-a**). Through communication, that is, the propagation of messages within a social network, each person's opinion should ideally become as informed as possible, integrating the knowledge and experience from all the network members. In other words, we work under the hypothesis of normativity, according to which the purpose of communication is to ensure that individual opinions converge to a consensus corresponding to the posterior probability of the correct answer given all the external evidence. This "ideal" situation, well defined mathematically, represents a benchmark against which various message propagation schemes can be compared, while significant deviation (such as systematic overconfidence) could be considered irrational.

Unfortunately, the structure of social networks renders simple message passing schemes fatally flawed as an inference mechanism (see **Figure 18-b-d**). In particular, every loop in a social graph forms an *echo chamber* where opinions can reverberate *ad infinitum* and be artificially amplified (Baumann et al., 2020; Santos et al., 2021) (see **Figure 18-d-e**). We thus confront both the strengths and weaknesses of the massification of social media: social networks could (ideally) make local information globally available as never before. However, they also tend to aberrantly amplify confidence, leading to radicalization and polarization and, as we will see, severely limiting their true information sharing capability.

The goal of this paper is twofold. First, we provide a simple account of echo chambers using a probabilistic inference framework (BP) applied to realistic social graphs and systematically study their consequences. Second, we propose a method (CBP) that limits these detrimental effects by trying to achieve normality, bringing the confidence levels generated in the network closer to informed rationality. We demonstrate the efficiency of this algorithm in both toy graph-models and more realistic graph structures borrowed from popular social networks.

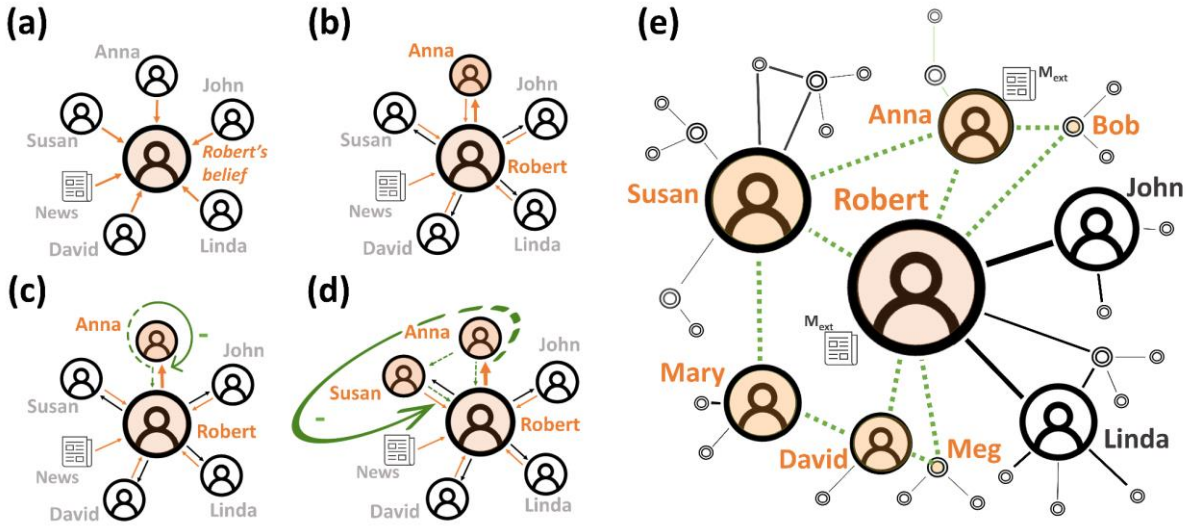


Figure 18 : A normative account of message-passing in social networks. **(a)** Let us consider a social network in which an agent, for instance Robert, communicates with other agents one-to-one and reads the news. In this context, Robert's belief can be computed as the sum of all the information he receives: internal messages from other agents and external messages (Mext). **(b)** What information should Robert communicate to Anna? An initial answer could be all the information Robert received, including information that came from Anna (the mean-field implementation). This approach is suboptimal since redundant information is exchanged and thus counted several times. **(c)** To address this overcounting, a second implementation (Belief Propagation) would consist of sending Anna all the information received by Robert except the information coming from her (dashed green arrow). This message cancellation is indicated by the green arrow. **(d)** However, the problem becomes much more complex when Anna and Robert have common friends like Susan. In this case, Robert's belief is corrupted by what Susan knows from Anna. Therefore, an extra correction has to be applied to control the flow of circular messages in the social graph. The Circular Belief Propagation (CBP) algorithm implements such a correction. **(e)** Without proper control, the highlighted problem becomes much more serious when the social graph is highly cyclic, when adding new friends/followers such as Mary, Meg and Bob (see green connections).

IV.2. Results

IV.2.1. A brief overview of message passing schemes

In our simplified social network model, each agent is a node of the network, and edges of the network represent the social circle of agents (the people they directly communicate with, that is, friends or followers). We assume that each agent *i* estimates the exact probability distribution $p(x_i)$ of a binary variable x_i ($x_i =$ "yes" or "no", for instance to the question *Should abortion be legal or not?*) with $b(x_i)$ the estimate probability distribution, given diverse sources of external information (web-search, articles, books, TV programs, or even direct evidence such as that coming from field journalists or researchers) and the agent's preference. In the following, we will refer to $p(x_i = \text{"yes"})$ as the *true probability* (formally the true *marginal posterior probability* of x_i to be "yes") and to $b(x_i = \text{"yes"})$ as the *estimate probability*. For convenience, we also

define the *belief* B_i of agent *i* (that x_i is “yes” versus “no”) such that $b(x_i = \text{"yes"}) = \text{sigmoid}(2B_i) = 1/(1 + e^{-2B_i})$; see **Supplementary Figure 1**.

The sign of B_i describes the agent’s opinion about the binary question: if for example $B_i > 0$, the agent believes that the answer to the corresponding question is more likely to be “yes” than “no”. Additionally, the absolute value $|B_i|$ quantifies the *confidence* of agent i. The higher the confidence, the more certain the agent is about the answer, while $|B_i| = 0$ implies complete uncertainty.

To determine the value of his/her belief B_i, the agent has to combine two types of information:

- External sources of information received by this agent, or agent’s preference, grouped together and quantified as the *external message* $M_{\text{ext}\to i}$. Such a message (mathematically defined as a log-likelihood ratio) is negative if it supports “no”, and positive if it supports “yes” ; The amplitude of this external message indicates its assumed reliability.
- Information provided by the opinions broadcasted by members of the agent’s social circle (called *internal messages* in the following). $M_{i\to j}$ denotes the message sent from agent i to agent j.

An agent's core belief is the sum of all the internal and external messages it receives (see **Figure 18-a**): $B_i^t = \sum_j M_{j\to i}^t + M_{\text{ext}\to i}$.

Meanwhile, the message $M_{j\to i}$ depends on the belief of agent j, and the amount of trust (Liu et al., 2018) between the two agents i and j. In the simplest possible message passing scheme, called *variational message-passing* (Winn and Bishop, 2005), the message corresponds to the sender’s belief modulated by trust: $M_{j\to i}^{t+1} = f_{ji}(B_j^t)$, where $f_{ij} = f_{ji}$ is a sigmoidal function which depends on the (mutual) amount of trust between the two agents. This naive method of communication assumes that agents systematically broadcast their opinion to their entire social circle, and in turn combine internal and external messages to update their own beliefs (see **Figure 18-b**). This message-passing algorithm corresponds to what was proposed in previous models of opinion dynamics in social networks (Baumann et al., 2020, 2021; Gray et

al., 2018) with slight differences in the precise form of the sigmoidal function (see *Methods*). However, the above mean-field scheme is highly suboptimal at performing inference in a graph. In particular, it creates a (potentially uncontrolled) reverberation of messages between each connected pair of nodes: agent j influences i, who influences j, etc. Humans probably never communicate this way; for instance, we only tell our friends things they presumably do not already know.

A less naive communication method, which we hypothesize to be our model for communication, ensures that messages do not include the messages sent previously in the opposite direction (see **Figure 18-c**). Messages are updated iteratively as follows: $M_{j\to i}^{t+1} = f_{ji}(B_j^t - M_{i\to j}^t)$. The resulting message passing scheme corresponds to a widely used inference algorithm called *Belief Propagation* (BP) (Pearl, 1988) (see *Math Appendix*). Despite its simplicity, this algorithm is surprisingly powerful as an (approximate) inference method (Bishop, 2006). In fact, BP is even exact in graphs without cycles, that is, it converges to the true posterior probabilities. However, in the presence of cycles, messages can still be reverberated and artificially amplified, leading to overconfidence, shown schematically in **Figure 18-d**. Unfortunately, social networks contain a large number of such loops (see **Figure 18-e**). As a result, we will see that BP, considered as a model of social communication, systematically leads to radicalization and polarization in cyclic social graphs.

As a society, we urgently need to find solutions that can preserve the global knowledge sharing capabilities of social networks, while suppressing the detrimental effects of loops or echo chambers. When integrating information from someone, one should in theory consider all the indirect ways the content has been brought to him or her (through a common friend for instance) in order to not take into account the same piece of information twice. With this goal in mind, we introduce an adaptation of the Belief Propagation algorithm called *Circular Belief Propagation* (CBP) (Bouttier, 2021) which aims at actively removing redundancies between messages introduced by loops and amplification of messages through cycles. The resulting message passing scheme can be written as follows:

$$M_{j\to i}^{t+1} = f_{ji}(B_j^t - \alpha_{ij} M_{i\to j}^t)$$

where beliefs are defined by:

$$B_i^t = \kappa_i \Big(\sum_j M_{j \to i}^t + M_{\text{ext} \to i} \Big)$$

In contrast to BP, CBP contains two types of control parameters: a gain κ_i applied to each node, and a loop correction term $\alpha_{ij} = \alpha_{ji}$ applied to each link. The idea the first equation is to subtract more than once the opposite message $M_{i \to j}$ from the belief of agent j. This is based on the fact that agent j is not only influenced directly by i, but also indirectly by any person k (all messages $M_{k \to j}$ might contain some part of $M_{i \to j}$ as i might influence k). Intuitively speaking, the loop correction term "$-\alpha_{ij} M_{i \to j}$" subtracts the predictable "redundant" part from incoming messages, which is the result of the reverberation of the outgoing message through all the graph's loops. Similarly, the gain κ_i in the second equation prevents the amplification of beliefs due to excess correlations between all incoming messages as introduced by loops (agents influence themselves, as messages travel back).

Note that these control parameters need to be adjusted to the specific graph structure of the social network, thus posing an additional challenge. Here we will consider two methods of finding a good set of control parameters: (a) a supervised learning method, that can only be used in extremely small graphs, and (b) a local unsupervised learning rule that is less optimal but applicable to graphs of arbitrary sizes (see *Methods*). All control parameters can be trained in an unsupervised manner by ensuring that incoming and outgoing messages remain as decorrelated as possible when they contain no meaningful information.

To model opinion formation in a social network, we iterate 100 times the BP/CBP message passing scheme simultaneously in all the nodes, to let the information provided by the external messages propagate in the entire graph (at which stage beliefs and messages usually reach a stable state). Further details are provided in the *Methods* section and the pseudo-code is given in the *Math Appendix*.

IV.2.2. Performance of the bp and cbp algorithms

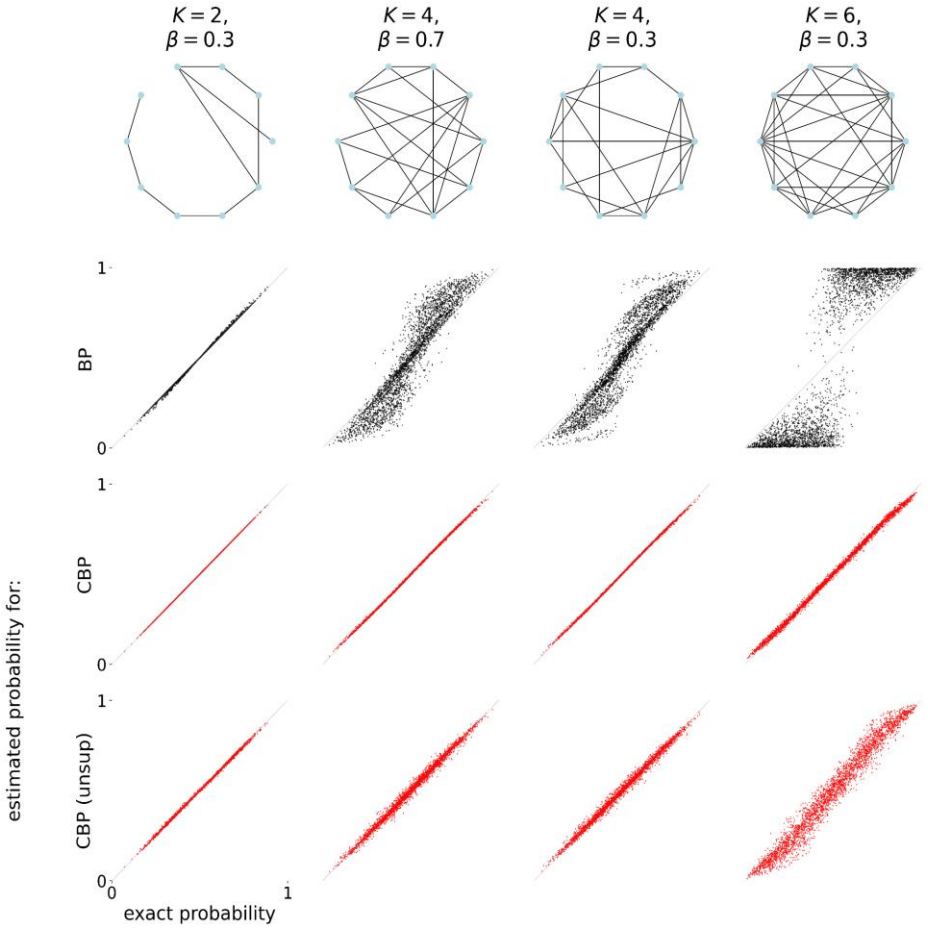


Figure 19 : Performance of message passing algorithms in 10-node toy examples. **Top row**: Example of Watts-Strogatz small-world graphs with 10 nodes and different characteristics: the average degree (K) and the probability that a local connection is replaced by a long range connection (β). **Second row**: Comparison between the marginal posterior probability computed by BP $b(x_i = \text{“yes”})$ and the exact posterior probability $p(x_i = \text{“yes”})$, for the different types of networks. Each dot shows the probability of a node, in one of 100 trials (using random external messages), in one of 30 random graphs with the structural properties (K and β) shown above. **Third row**: Same for the marginal posterior probability computed by CBP, with control parameters obtained by minimizing the distance between exact and approximate posteriors on a distinct training set (supervised learning)..**Fourth row**: Results for CBP with control parameters learned in an unsupervised way. See the *Methods* section for more details.

We first tested the performance of BP and CBP in small graphs (toy examples with $n = 10$ nodes) where running exact probabilistic inference is still practical, as well as supervised learning (see **Figure 19**). This way, the resulting exact posterior marginal probabilities can be used as a benchmark (an ideal result) for comparison with BP or CBP. The graphs were generated to have a Watts-Strogatz *small-world* structure, as the latter share some features with social networks (Vespignani, 2018). Such a graph structure is controlled by two parameters: the mean degree (number of connections)

for each node, called K, and the probability that a connection is “long range” as opposed to local between neighbors, called β (see **Figure 19-a** for example structures).

We compared the approximate posterior probability solutions $b(x_i = \text{“yes”})$ found by different message passing schemes with the exact posteriors, $p(x_i = \text{“yes”})$, for given external messages $M_{\text{ext}\to 1}, M_{\text{ext}\to 2}, \ldots, M_{\text{ext}\to n}$. To ensure the generality of the results, this comparison was performed for several randomly generated graph structures (30 graphs for each setting of the structural parameters) and in response to numerous sets of randomly generated external messages (“trials”). Each dot in **Figure 19** corresponds to an approximate posterior probability for one node in a given trial. As the density of the network increased, the performance of BP degraded (**Figure 19-b**). In particular, all beliefs became too extreme, resulting in a condensation of approximate posteriors close to 100% or 0% even when the external evidence did not justify such confidence (for instance if the true posterior was in fact close to 50%).

Next, we tested CBP after learning the control parameters using a supervised learning method (provided in *Methods*). The parameters κ_i and α_{ij} were chosen to minimize the distance between the approximate and true posterior (on a training set independent from the test set shown in the figure). After optimization, CBP matched the exact inference very closely, with no sign of overconfidence (**Figure 19-c**).

Such supervised optimization is only possible in networks with a relatively small number of nodes. In larger networks, and any realistic social graph, exact probabilistic inference is impossible because it scales exponentially with the number of nodes. Fortunately, CBP parameters can also be trained without any knowledge of the true posteriors. Using purely local learning rules, the control parameters can be trained to remove correlations between incoming and outgoing messages and to suppress redundancies between incoming messages (see *Methods*). Despite the heuristic nature of these learning rules, the approximate posteriors remain close matches to the true posteriors (**Figure 19-e**).

This toy example demonstrates that CBP can alleviate the overconfidence problem associated with BP in cyclic graphs, resulting in more rational beliefs. Since we move on in the next section to larger graphs where exact inference is intractable, it is assumed that the parameters of CBP were trained for each graph structure using the proposed local, unsupervised learning rules, rather than with supervised learning.

IV.2.3. Towards greater realism: larger graphs

The next step was to investigate how these effects generalize to more realistic social graph structures. First, we investigated larger (but still simplistic) Watts-Strogatz graphs with 200 nodes. By systematically varying K and β, we explored the impact of the number of connections per node and long-range connections. These findings will be useful for explaining more complex behavior in "realistic" social graphs (see the next section).

Figure 20 shows an example graph with moderate degree ($K = 30$) and proportion of long-range connections ($\beta = 0.12$). We provided unreliable external messages that did not strongly support a "yes" or "no" answer. More specifically, in each "trial", each external message was sampled from a Gaussian distribution with zero mean and standard deviation $\sigma_{\text{ext}} = 0.1$.

To understand how opinions are formed, it can be useful to visualize the belief trajectory during the deliberation process, that is, while messages are still being propagated. **Figure 20-a**, top row, examines the case of BP. Starting from complete uncertainty ($B_i = 0$ for all agents *i*), the beliefs in the different nodes evolve over the iterations of the BP message passing scheme until they stabilize at constant levels, representing the *opinions* generated by BP. Differences in opinions among the nodes are induced by random variations in local graph structures and in the external messages the nodes receive. Each new trial generates a different set of opinions (left and right panels of **Figure 20-a**). Note that the beliefs converge to very large values (either positive or negative), most agents being at least 99% confident in having a correct answer (see **Figure 20-c** for the relationship between beliefs and probabilities).

While it is not tractable to compute the exact posteriors, we can estimate an upper bound on "rationality" (the dashed line). This corresponds to the belief of a universal observer summing all the external messages directly: $B_{\text{univ}} = \sum_i M_{\text{ext}\to i}$[5]. Beliefs larger than B_{univ} (in absolute value) are necessarily overconfident, since they go beyond the total external evidence. As we can see, BP results in severe overconfidence for most nodes in the graph, despite the true unreliability of external messages. Note that B_{univ} is an upper bound, not an exact posterior. In fact, if inference

[5] In practice, this would correspond to the beliefs of all nodes if exact inference were performed in a network with full connectivity and infinite trust (in which case, all the x_i values would collapse to a single binary random variable and all external messages would be noisy evidence for this shared variable). Since our network has limited connectivity and trust, each node can only achieve a lower confidence level, at least if it remains rational.

were exact, the agents would have significantly lower confidence than the universal observer, for two reasons: the nodes do not trust each other completely (there is always a chance that your friends are wrong…), and not all of them are connected to all other nodes.

In contrast to BP, the CBP algorithm leads to far more moderate opinions (see **Figure 20-a,b**, bottom panels), with no sign of radicalization or polarization. The final beliefs are narrowly distributed around a consensus value, which is itself close to zero (low confidence). The beliefs always remain below the universal observer, as would be expected from a rational deliberation process and are in agreement with the completely uninformative nature of the external messages chosen for these trials.

The BP-generated opinions are represented graphically on the top row of **Figure 20-b**, illustrating how opinions can be distributed as a function of the proximity (inverse path length) between two nodes. Only two possible outcomes were observed in those graphs. In the first scenario, the entire population reaches the same extreme opinion, either for or against (**top-right panels, Figure 20-a,b)**. We interpret this phenomenon as a *radicalization* of the entire population. In the second scenario, two populations with opposite but similarly extreme opinions emerge. These populations are separated into 2 or more local clusters within the graph (***top-left panels***, **Figure 20-a,b**). We interpret this as *polarization*. We quantify the level of *radicalization R* as the mean absolute value of the beliefs and the level of *polarization P* as their mean standard deviation (computed within a single trial). These definitions correspond to (or are highly similar to) the ones used in other studies (Banisch and Olbrich, 2019; Baumann et al., 2020, 2021; Lee, 2016). The left panels in **Figure 20-a,b** have both high radicalization and high polarization, while right panels have high radicalization but low polarization. Note that the only thing differing between the two panels are the external messages (two sets sampled from the same distribution).

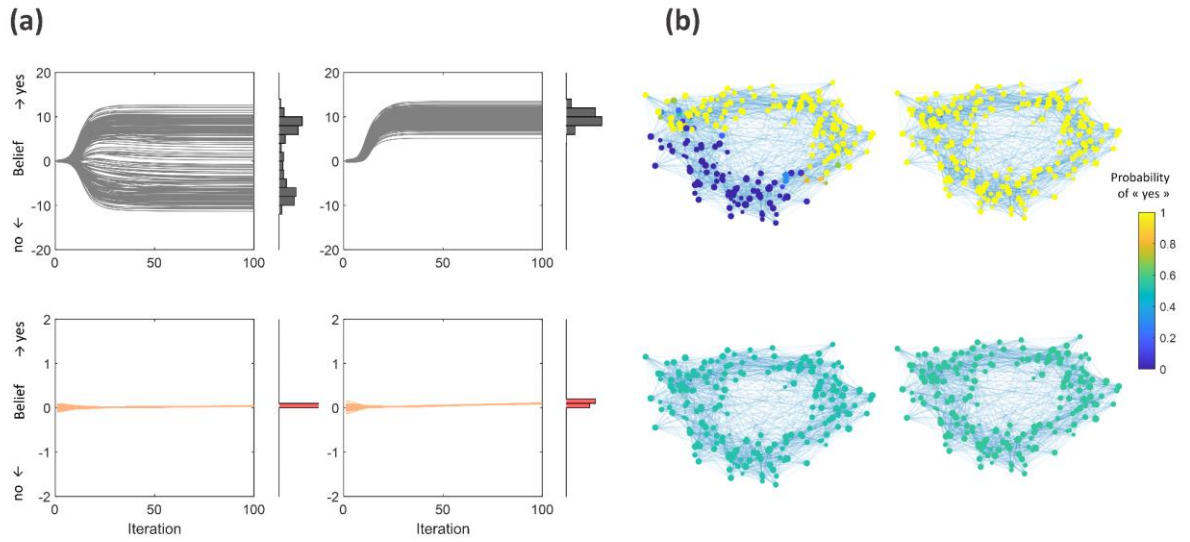


Figure 20 : Response of an example 200-node small-world network (K= 20, β= 0.12) to uninformative external messages (random and unbiased). **(a)** Temporal evolution of the belief under the message passing algorithm (that is, while internal messages propagate in the network). By convention, the final beliefs are those obtained after 100 iterations. Two example trials are provided. **Top row**: BP leads to two possible outcomes depending on whether the whole population behaves similarly or separates into two groups with opposite, extreme beliefs. This is interpreted as polarization (**left**) or radicalization (unimodal in the **right** panel, bimodal in the **left** panel). **Bottom row**: In contrast, CBP leads to beliefs narrowly distributed around a weak consensus. This consensus varies from trial to trial but remains close to zero, reflecting higher uncertainty. **(b)** Final beliefs of the 200 nodes, visualized in the whole graph. The nodes (dots) are arranged topographically according to the path lengths (separation within the graph). The size of a dot represents the node's degree, and its color represents the marginal posterior probability estimate ($b(x_i = \text{“yes”})$), abbreviated as the “Probability of yes”. Thin lines are connections. The two trials shown here are the same as in (a), with the top and bottom rows corresponding to BP/CBP. The relationship between the belief and the “probability of yes” is illustrated in Sup.Fig.1.

The radicalization or polarization due to BP and the suppression of these characteristics by CBP are very general results that are independent of the specific network structure, as illustrated in **Figure 21** . In the case of BP (**Figure 21-a**), the severity of polarization and radicalization systematically depends on the two structural parameters: radicalization increases quasi-linearly with K (***left panel***), while polarization decreases with β (right panel). Interestingly, polarization is strongest in a sweet spot with a moderate K and a small value of β. This sweet spot corresponds to a high probability of echo chambers, which corresponds to local clusters of highly interconnected nodes that are relatively isolated from the rest of the graph (due to the predominance of short-range connections). **Figure 21-b** examines in more detail the belief distributions resulting from BP at the level of the population (combined over many

trials and several random graphs) when increasing K. Note that the distribution has two distinct modes, whose separation increases with K.

These features are completely suppressed by CBP. Radicalization and polarization are eliminated (**Figure 21-c**), and beliefs are no longer separated into two distinct modes. Instead, the distribution presents a single mode, centered at zero, with a variance increasing with K (**Figure 21-d**).

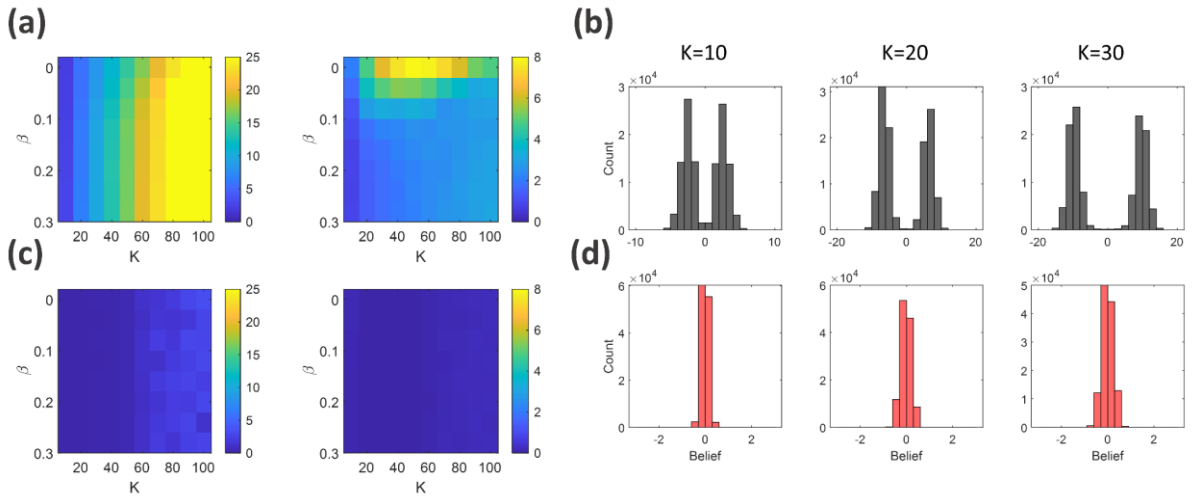


Figure 21 : Response of an example 200-node small world network to uninformative external messages as a function of their structural properties K and beta. **(a)** Mean radicalization R **(left panel)** and polarization P **(right panel)** as a function of K and β when using the BP algorithm. **(b)** For BP, the distribution of beliefs over multiple trials, for networks with β = 0.12 and increasing values of K. **(c)** Same as (a) for CBP. **(d)** Same as (b) for CBP.

Preventing radicalization and polarization is not sufficient per se (for instance, a trivial way of achieving this result would be to set all gains at $\kappa_i = 0$: this way, all beliefs would have been equal to zero). One must also ensure that the message passing scheme operates properly when external messages are actually informative, that is, when they globally provide more support for one option than the other. That is why we now consider a situation in which there is a true answer supported by evidence (such as "*vaccination decreases the risk of severe outcomes following a COVID-19 infection*"). The task of the network is now twofold. First, there should be as many agents as possible whose beliefs point in the direction supported by the evidence (in this case, that most agents believe that COVID vaccines work). Second, confidence

levels should increase in proportion to the true strength of this evidence (for instance, skepticism about a therapeutic approach is desirable as long as it has not been validated by rigorous clinical studies).

We generated informative external messages (inputs to the graph) by sampling them from a biased distribution (with a positive mean if the true answer is “yes”). This bias was small compared to the variance of the distribution, such that many individual nodes received misleading external messages (fake news). Moreover, these external messages were injected into only a small portion of the nodes, while others received no external messages (in a situation where the majority of people have no expertise on vaccines). If the network allows all users to share information optimally, every agent should believe in the answers supported by the highest amount of evidence (the sign of the sum of all external messages B_{univ}) even if their private external message points in the opposite direction (that is, even people exposed to fake news would eventually be convinced, through their social contacts, that vaccines are effective).

In investigating inference in the presence of informative messages, we found an interesting dissociation in performance when considering people’s choices or their confidence levels. People’s *choice* would correspond to their answer to a survey with only two possible options (such as “*Do you think that the COVID vaccine works? yes/no*”). Presumably, they would choose the answer they believe the most, that is, answer “yes” if their belief is positive. In contrast, people’s confidence would correspond to the absolute value of their beliefs (for example, “*How confident are you that covid vaccines work/do not work, on a scale from 1 to 10?*”).

Let us first consider choices. In a strongly connected network ($K = 40$, $\beta = 0.2$) with small mean path length between nodes (1.9 here), the portion of nodes with the “correct choice” after running either BP or CBP increases similarly to the proportion of informed nodes (the proportion of nodes receiving external messages). Moreover, this increase is perfectly predicted by a universal observer summing all the external messages together, whose belief is $p(B_{\text{ideal}} > 0)$ (**Figure 22-a**, left panel). In a network containing less long range connections ($K = 20$, $\beta = 0.08$) with a longer mean path length (2.6 here), both BP and CBP perform worse than the universal observer, reflecting the limitations introduced by the more indirect communication between nodes. However, CBP now clearly outperforms BP (**Figure 22-b**, left panel). To intuitively understand why a smaller number of long range connection results in poorer choices, consider an

extreme scenario: a network with no long range connections at all ($\beta = 0$), in which case all nodes are organized on a fat ring, with subpopulations at opposite ends having no direct connections. They can only influence each other indirectly by changing the beliefs of intermediate nodes, which is not possible if those nodes are radicalized (as is the case of BP). By keeping beliefs graded, CBP restores long range communication within the network.

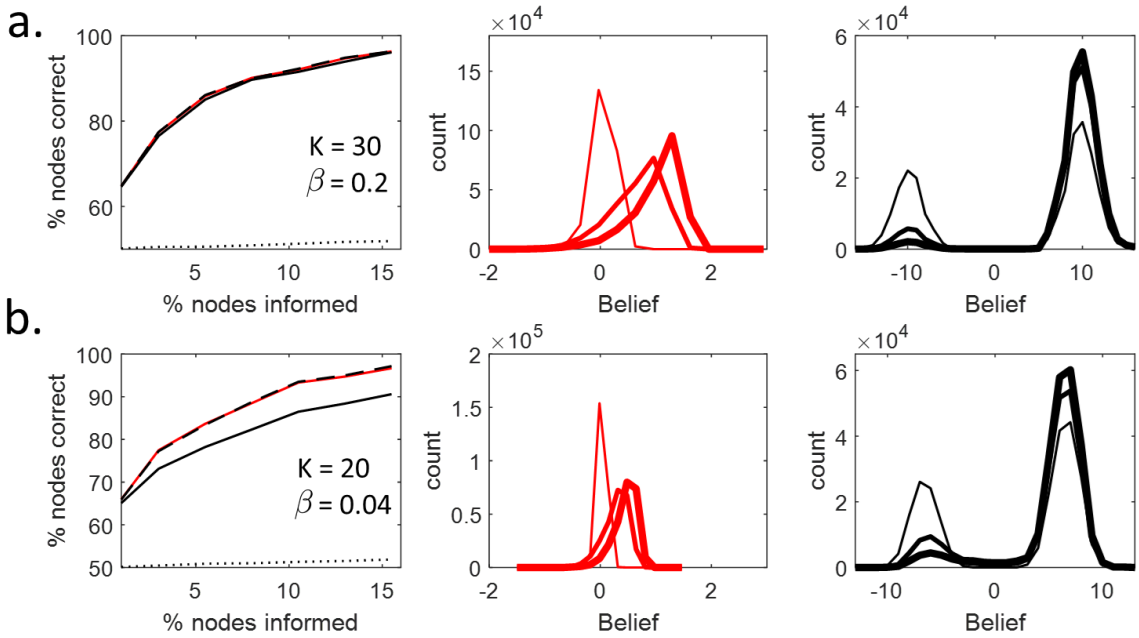


Figure 22 : Responses of the 200-node small world graphs to informative (biased) external messages. **(a)** Average of 6 graphs with K= 30, β = 0.2. **(b)** Same as (a) but for K = 20, β = 0.08. **Left panels**: choice performance of the different message passing schemes. Black: BP, red: CBP, dashed: universal observer, dotted: based on external messages only, without taking into account the internal messages. "% Node informed" is the percentage of nodes receiving non-zero external messages "% Nodes correct" is the percentage of nodes with a belief whose sign points to the true answer (defined by the bias in the distribution of external messages). **Middle panels**: Belief distribution over all nodes for increasing amounts of external information as a result of CBP. Thin line: 1% nodes informed, normal line: 5% of nodes informed, thick line: 10% of nodes informed. **Right panels**: Same as the middle panels, but for BP.

Where BP and CBP most strikingly differ is in their confidence levels (that is, answering with not only a yes/no reply but on a scale - **Figure 22-a,b**, middle versus right panels). In the case of BP, beliefs are always distributed in two extreme modes, leaving no room for uncertainty (**Figure 22-a,b**, right panels). As more evidence arrives in support

of a positive choice (more nodes are informed), the proportion of belief in the positive mode (making the “right choice”: positive beliefs) increases, but the nodes that are still in the negative mode (making the “wrong choice”: negative beliefs) remain equally overconfident (**Figure 22-a,b**, right panels). When a node finally changes its mind, it can only switch between these two extremes, with no intermediate stage of uncertainty. Such phenomena could have potentially deleterious societal consequences: people convinced of their correctness could reject the vaccine at any cost and become impervious to information campaigns and contrary evidence; even if they change their minds, one form of extremism could lead to the opposite one. In contrast, with CBP, the beliefs are far less extreme, and their unimodal distribution gradually shifts toward the positive side as more evidence is provided (**Figure 22-a,b**, middle panels). In other words, the stronger the evidence, the more confident the correct nodes are of being right. Conversely, the incorrect agents become less confident, as should occur following a rational consensus building process.

IV.2.4. Real social-network examples

Finally, we tested BP and CBP on large online social network structures taken from open-access Facebook© and Twitter© data (McAuley and Leskovec, 2012, see **Figure 23-to-29**). The results are globally consistent with what was observed in toy examples. As before, BP generates aberrantly strong beliefs, even in response to completely uninformative messages. More realistic social graphs contain cliques of highly connected nodes separated by relatively sparse long-range connections. As a result, polarization within local clusters (as opposed to general radicalization of the whole population) was by far the most likely outcome in response to uninformative external messages (see examples in **Figure 23-a,b** top-row). In contrast to BP, CBP generated moderate confidence levels, with no obvious radicalization or polarization (**Figure 23-a,b** bottom row). Small correlations of beliefs within cliques can still be observed, but they are to be expected even if inference is close to exact because of the predominance of short-range connections.

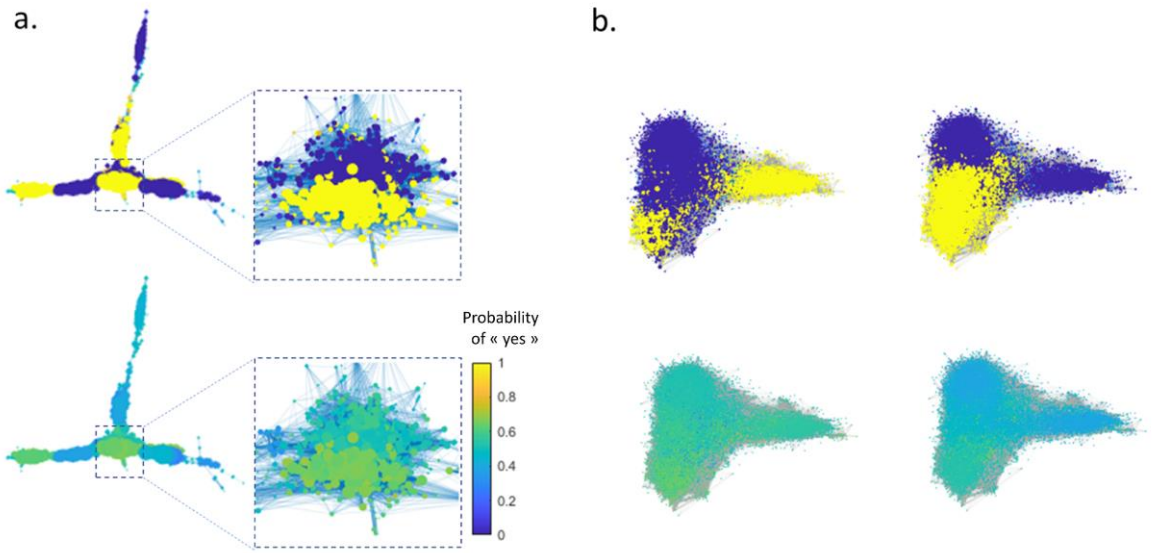


Figure 23 : Social network responses to uninformative external messages. **(a)** Response of the Facebook© subnetwork with 3959 nodes and 84243 connections, for one example trial, with enlargements of the central part. **(b)** Twitter© subnetwork with 81306 nodes and 1.34 million connections, tested on two different trials (columns). The **top row** represents beliefs computed with BP, and the **bottom row** with CBP. Same legend as in **Figure 24-b**.

The more complex structures of these networks made it possible to investigate in more detail the relationships between local graph structures, control parameters (for CBP), and beliefs (**Figure 24**). In the case of BP, overconfidence is directly proportional to the degree of the node being considered (**Figure 24-a,b**, ***black dots***). Thus, the most connected nodes (agents interacting with many people) develop more extreme views. In contrast, CBP results in far more moderate beliefs and globally weakens (but does not completely remove) the relationship between confidence and node degree (**Figure 24-a,b**, ***red dots***). CBP achieves this control by learning to decrease the gains (κ_i) and increase the loop corrections (α_{ij}) in nodes of larger degree (**Figure 24-c,d**). In other words, CBP needs to exert stronger controls on nodes that are most massively connected to the rest of the network (influencers) and are thus at the largest risk of becoming radicalized.

To investigate the information sharing capabilities of these networks, we tested them with informative messages provided to small subsets of the nodes, as previously done. In these larger and more modular networks (mean path length 5.5 for Facebook©, 4.9 for Twitter©), both BP and CBP unsurprisingly perform worse than a

universal observer (**Figure 24, left panels**). However, CBP strongly outperforms BP. As before, BP exhibits extreme overconfidence, regardless of whether the nodes are correct (B>0) or incorrect (B<0) in their choices (**Figure 24**, right panels). While the distribution of BP-generated beliefs appears unimodal, it is in fact a consequence of the naturally wide distribution of node degrees in social graphs. If only nodes of similar degrees (e.g., between 20 and 50) are combined, the distribution of belief once again becomes bimodal (**Figure 24-c**, left panel), and the separation between the two modes increases with the degree exactly as in **Figure 21-b** (for example, imagine measuring the distribution of black dots in a vertical slice in **Figure 24-a**). In contrast, the beliefs generated by CBP remain unimodal at all degrees and moderate but with a marked shift and extension toward larger confidence levels as more external information is provided (**Figure 24**, middle panels). In other words, correct nodes become more confident, while incorrect nodes become less so. Finally, for both BP and CBP, nodes are more likely to be correct and confident if their degrees are larger, that is, if they directly collect messages from a larger portion of the network. This is why the CBP belief distribution not only shifts but also extends to the right as evidence increases.

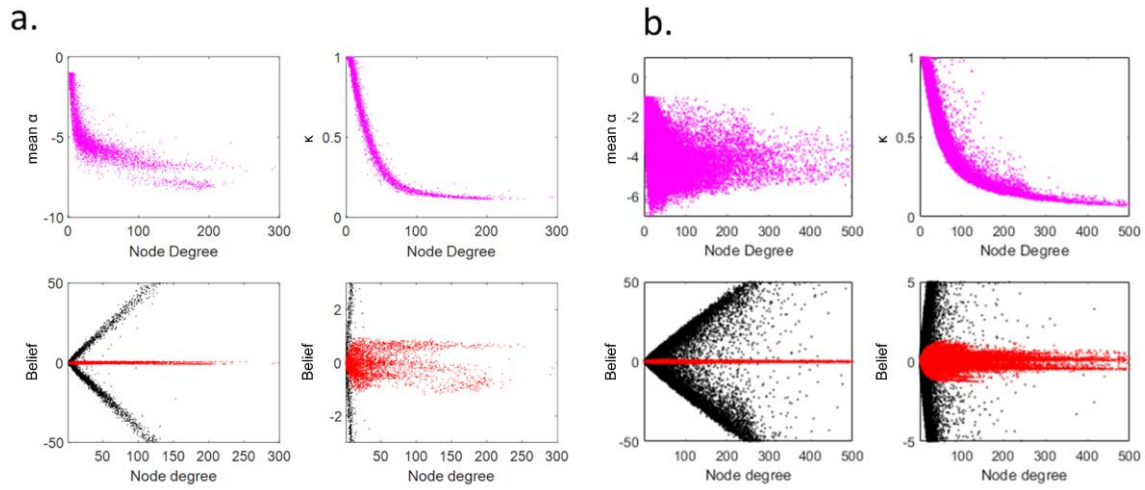


Figure 24 : Social network response to uninformative external messages and learned control parameters as a function of node degree. **(a)** Facebook© network, **(b)** Twitter© network. Top row: Learned control parameters κ_i, α_{ij} as a function of node degree. Each dot represents the control parameters for a single node, and the loop correction term is averaged over all incoming connections. Bottom row: Beliefs from BP (black dots) and CBP (red dots) at two different scales of the y axis. The response to a representative trial is shown, with each dot corresponding to one node.

We can predict from these results what would be the consequences of willfully spreading fake news on people's choices and confidence. Both BP and CBP are relatively resilient when it comes to choices: they integrate all the external messages. Fake news would have to overwhelm "real news" to cause a global change in people's choices. However, the most detrimental effect by far is the potential creation of a small number of extremely polarized nodes, with contrafactual but unshakable beliefs (under BP). This does not take place when reverberation in echo chambers is controlled (CBP). In this case, fake news decreases the mean confidence level but without causing the emergence of extremism (see **Figure 24-c, right vs middle**).

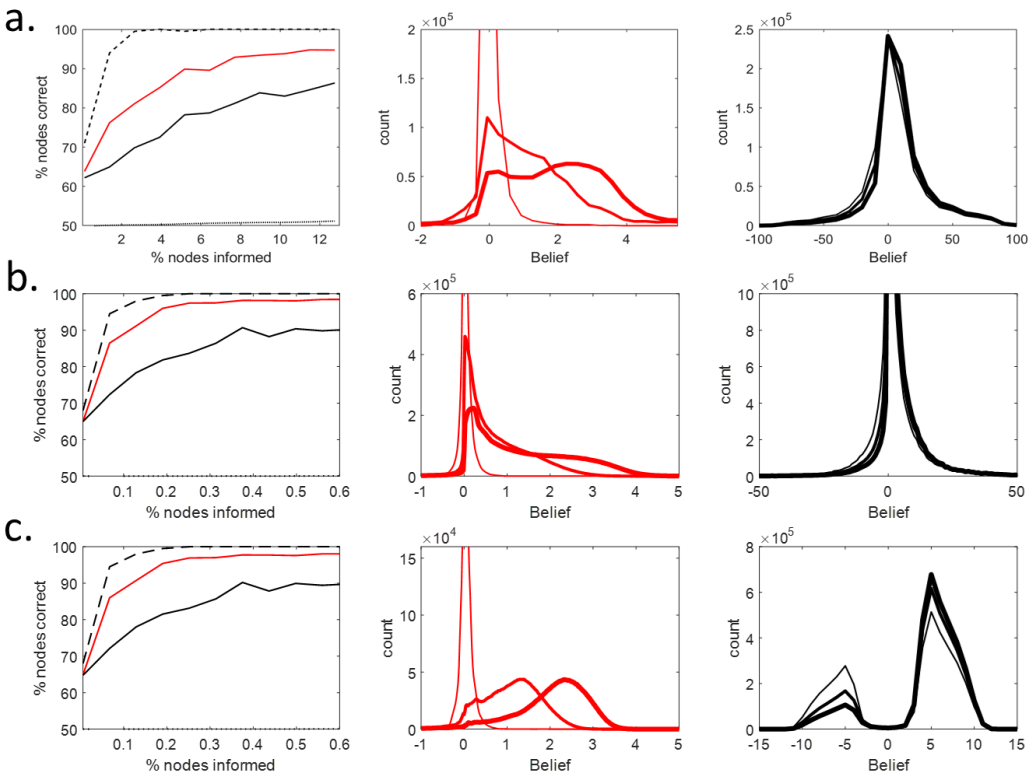


Figure 25 : Response of social networks to informative external messages (same legends as in **Figure 22**). **(a)** Facebook©. **(b)** Twitter©. **(c)** Twitter© network, but only for nodes with degrees between 20 and 50. The % of nodes informed is unchanged, but the % of nodes correct is now computed only for this subpopulation. Left panels: Choice performance of the different message passing schemes. Black: BP, red: CBP, dashed: universal observer, dotted: external messages only. Middle panels: Belief distribution over all nodes for increasing amounts of external information as a result of CBP. Right panels: same as middle panel, but for BP. Thin line: 1% of nodes informed, normal line: 5% of nodes informed, thick line: 10% of nodes informed.

IV.3. Discussion

Social media networks have always been the theater of repeated questioning and reassessment of ideas that were previously considered unshakable. On the one hand, they have repeatedly demonstrated their invaluable power in bringing together thousands of people to support common important causes of the 21st century. This was notably illustrated by the *MeToo* hashtag, which is famous today for denouncing sexual harassment and abuse, allowing the empowerment of survivors and often forcing societal actions against perpetrators. On the other hand, the way social networks shape public debate has also been exacerbated by populist parties and supporters of conspiracy theories (Tollefson, 2021). This last phenomenon appears to directly benefit from real-world uncertainty (Lee, 2016; Suthaharan et al., 2021) as well as from the viral spreading of information that may reinforce a monolithic (and often extreme) view (Franceschi and Pareschi, 2022). Beyond simply modeling echo-chambers in social networks (here, with the Belief Propagation model of communication), Circular Belief Propagation allows for a solution to moderate overconfidence, going against the effects of echo-chambers and current recommendation systems.

Numerous other theoretical models have been proposed to describe opinion formation in social networks (Barrat and Weigt, 2000; Baumann et al., 2020; Castellano et al., 2000; Deffuant et al., 2000; Galam et al., 1982; Girvan and Newman, 2002; Newman and Park, 2003; Peralta et al., 2021; Santos et al., 2021; Tokita et al., 2021). Some, like the famous voter model, are simple enough to allow complete mathematical analysis (Castellano et al., 2000). Others have used Ising models (Galam et al., 1982) or investigated the specificity of small world social graphs (Newman and Park, 2003). While many previous models have used binary opinions, others represented them on a continuum of belief (Deffuant et al., 2000) such as our model. However, all these models fundamentally differ from ours, not only in their mathematical details, but more importantly in their starting point and objective. All previously cited models take *descriptive* approaches: their starting point is the description of how agents locally interact and their goal is to understand the emergence of collective dynamics. On the contrary, our approach is *normative*: its starting point is a functional hypothesis about the purpose of communication - more precisely, that our opinions are formed optimally when considering the whole external information and the levels of trust between individuals (see Karamched et al., 2020 for another example of a normative approach,

based this time on collective evidence accumulation). Finally, the objective of our approach is to find strategies to achieve this function or come close to it.

We propose that the root of the radicalization problem is not disinformation or cognitive biases per se (although they certainly play a role). Rather, online message reverberation leads to systematic overconfidence, as information is unknowingly amplified in *echo chambers*. By using a normative approach of opinion formation in a social graph and exploring graphs of progressive complexity, we quantified these phenomena and demonstrated its generality. In popular online social networks (Facebook© and Twitter©), the resulting strength of convictions will largely exceed the available evidence and irremediably lead to the emergence of incompatible world views in different communities (subparts of the network). Confidence levels may become so extreme that opinions are virtually unshakable, remaining the same regardless of the amount of contradictory evidence.

Borrowing from variational methods of approximate inference in graphs, we proposed that the *Circular Belief Propagation* algorithm (CBP) can alleviate these detrimental effects. This algorithm learns to suppress messages according to how predictable (redundant) they truly are. In small graphs, we showed that CBP achieves close to optimal performance: the social network generates confidence levels that go hand-in-hand with the amount of available evidence. In larger graphs and realistic social networks, CBP avoids radicalization and polarization and ensures that beliefs remain rational.

We are social beings who have exchanged information for millions of years. It would be surprising if we did not have inbuilt cognitive and social strategies to deal with echo chambers in local communities. What may have changed recently is rather the scale and speed of social communication compared to what was previously possible (Dunbar, 2016). The worsening trend (increase in radicalization and polarization) appears to be domain-general and applies to many different fields. Beyond the political scene that is often taken as an illustration (Zmigrod and Tsakiris, 2021), the scientific community is not immune to overconfidence, particularly since scientific debates have spread from polite but limited academic circles to social media. As a recent example, the results of a trial on the clinical and brain effects of psychedelics in depressive disorders (Daws et al., 2022) were vividly discussed online

with unusual levels of passion even for a scientific debate (see, for instance, this blog entry relating the dispute (Love, 2022)).

Interestingly, problems that are naturally associated with highly interactive social communication may have their counterpart in the maze of our brain cells, another example of large-scale cyclic graphs. Indeed, the CBP algorithm used here was originally proposed in the context of hierarchical brain structures to investigate reverberations in feedforward/feedback circuits, causing so-called *circular inferences* (Deneve and Jardri, 2016). Controlling for reverberations in the brain could involve ubiquitous neural mechanisms such as enforcing the excitatory-to-inhibitory balance (Bouttier et al., 2022) and account for puzzling perceptual phenomena such as bistable perception (Leptourgos, Bouttier, et al., 2020).

Our simplified model of how communication changes people's opinions does not incorporate numerous aspects of social media communication. For instance, messages are not systematically broadcasted and connections are not necessarily symmetrical (for instance, messages propagate more often from influencers to followers than the reverse). Incorporating this new element into the model would limit polarization. Besides, while we considered stable states after unlimited message exchanges, temporal aspects were ignored. In real life, a piece of news is only propagated for a limited amount of time before becoming obsolete, and our beliefs are constantly updated as new information arrives. On the other hand, we also did not incorporate phenomena that could amplify the severity of echo chambers, such as biased information access (e.g., AI-powered chatbots fastening the spread of fake news (Giachanou et al., 2022)), past individual history and priors, or recommendation systems based on preferences (Santos et al., 2021) that might be used by social media to reinforce the weight of past online activities. Additionally, the present model considers fixed and positive connections, while individuals tend to communicate only with people having similar convictions (Sasahara et al., 2021), may distrust others (Proskurnikov et al., 2016), and may even actively distrust people with opposite convictions (Dubé and MacDonald, 2020). This last phenomenon would favor polarization. Lastly, the model only tackles communication over one particular topic, although people form opinions on many questions, and discussion about a subject influences our thoughts on related subjects (Baumann et al., 2021).

Despite these theoretical limitations, going towards an experimental validation of the model would be a giant leap forward. Simple online or offline experiments have been proposed and could potentially be modeled with either BP or CBP.

Future work will have to determine how the change brought by our proposed model (CBP compared to BP) could be implemented or promoted in real life, as this proposed solution remains theoretical for now. One way would be to inform people by displaying a measure of local polarization caused by the structure of their local interaction graph. This could make users integrate information differently, possibly in a CBP manner. Another way would be to act on recommendation systems by designing them to promote open-mindedness, which could help break echo chambers. This could mean reordering posts on social feeds to propose content according to their unpredictability for the user. This reordering could be monitored by users, for instance through a novelty scale.

IV.4. Methods

Here we describe how to reproduce the simulation results. For the theoretical foundation of BP and CBP equations, see the *Supplementary Material - Math Appendix*.

IV.4.1. Social graph models

Social graphs were formalized as Ising models with coupling strengths $\{J_{ij}\}$ and biases corresponding to the external messages $\{M_{\text{ext}\to i}\}$. Watts-Strogatz small-world graphs were generated as follows. First, a ring network was constructed by connecting each node to its K neighbors on the right and left. Next, with a probability β, this local connection was transformed into a long range connection between two randomly selected nodes. The structure of the realistic social networks were obtained from open source data, https://snap.stanford.edu/data/ego-Facebook.html for data from Facebook© and https://snap.stanford.edu/data/ego-Twitter.html for data from Twitter©. We assumed that coupling strengths were positive (since we communicate with others we trust). For each graph, coupling strengths were selected from a uniform distribution between 0 and J_{max}. We chose J_{max} = 0.6 for 10 node graphs, 0.36 for 200 node graphs, and 0.18 for realistic social graphs.

IV.4.2. Generation of external messages

Uninformative messages (used for training the control parameters and for measuring radicalization/polarization in the absence of meaningful evidence) where sampled independently from a Gaussian distribution with mean equal to 0 and standard deviation $\sigma_{ext} = 1$ (for small graphs with 10 nodes as in **Figure 19**) or $\sigma_{ext} = 0.1$ (for bigger graphs with 200 nodes and for realistic social graphs).

Informative external messages (see **Figure 22** and **Figure 25**) were sampled independently from a Gaussian distribution with mean +/- 0.05 (the sign defines the "correct choice") and equal variance $\sigma_{ext} = 0.05$. These informative external messages were provided sparsely to only a portion of the nodes m/n, where n is the number of nodes in the graph and m corresponds to the number of nodes receiving non-zero external messages (the proportion of informed nodes is $100\frac{m}{n}$). For each value of m, we generated 200 sets of informative external messages, each sampled independently from the same Gaussian distribution. Each time, these messages were fed to a different random selection of m nodes. After running the BP or CBP algorithm, we measured the final percentage of nodes with $B_i > 0$, which we called "percentage of correct nodes". This percentage was averaged over the 200 trials. In the case of the 200 node toy models (see **Figure 22**), this was also averaged over 6 different random networks generated with the same structural parameters K and β.

IV.4.3. Message passing algorithms

After being initialized at $M_{i\to j} = 0$, messages were propagated according to a damped version of the update equation provided in the Results section (see also *Math Appendix*):

$$M_{i\to j}^{t+1} = (1-\tau)M_{i\to j}^{t} + \tau f_{ij}(B_i^t - \alpha_{ij}M_{j\to i}^{t})$$

All messages were updated simultaneously for a total of 100 iterations, using $\tau = 0.2$ (the volatility, or rate of forgetting the old information).

The coupling function used in CBP is:

$$f_{ij}(x) = \tanh^{-1}\left(W_{ij}\tanh(x)\right)$$

where $W_{ij} = \tanh(J_{ij}) \in [0;1]$ since coupling strengths J_{ij} were taken to be positive. Note that this function closely relates to the one used in other models (which all consider $\alpha_{ij} = 0$) [5,10] : $g_{ij}(x) = W_{ij}\tanh(x)$. f_{ij} is bounded between $-J_{ij}$ and $+J_{ij}$ and has a sigmoidal shape.

IV.4.4. Parameter optimization

Control parameters for CBP were adjusted to the specific graph structure in order to improve inference as compared to BP. We considered two methods, supervised learning or unsupervised learning.

In *supervised learning optimization* (applied in this work exclusively to graphs with 10 nodes), the exact marginals $p_i(x_i)$ were computed using the junction tree algorithm. The control parameters were optimized by minimizing with supervised learning the mean squared error between the exact marginals $p_i(x_i)$ and the ones from CBP $b_i(x_i) = e^{(x_i+1)B_i}/(1+e^{2B_i})$ (where $x_i = 1$ for "yes", $x_i = -1$ for "no") over a set of 300 training examples (trials with uninformative messages):

$$[\alpha^*, \kappa^*] = \arg\min_{\alpha,\kappa} \sum_{\text{Trial}} \sum_i \sum_{x_i} (b_i(x_i) - p_i(x_i))^2$$

To propose *unsupervised learning* rules (applied to graph with 10 nodes or more), we noted that when the BP algorithm runs on a non-cyclic graph (in which case it performs exact probabilistic inference), messages in opposite directions $M_{i\to j}$ and $M_{j\to i}$ come from completely disjoint parts of the graph and are therefore uncorrelated. The same is true for different incoming messages to the same node (such as $M_{i\to j}$ and $M_{k\to j}$). When external messages (inputs to the graph) are uninformative - and thus uncorrelated -, these internal messages also remain uncorrelated. In contrast, in a cyclic graph, BP results in undue correlations of these opposite messages, which is a direct signature of information reverberation and overcounting in the graph (Ihler et al., 2005).

We thus used (unsupervised) learning rules on control parameters that aim at suppressing these detrimental correlations and ensure that they did not result in spurious belief amplification. We generated 2000 training trials with uninformative external messages. After being initialized as their default BP values $(\alpha, \kappa) = (1, 1)$, control parameters were updated after each trial as follows:

$$\Delta\alpha_{ij} \propto M_{j\to i}(B_i - \alpha_{ij}M_{j\to i})$$
$$\Delta\kappa_i \propto -(B_i^2 - \sum_j M_{j\to i}^2 - M_{\text{ext}\to i}^2)$$

The learning rates were adjusted to ensure that control parameters properly converged within the training window.

Because coupling weights are positive, the (anti-Hebbian) learning rule for α enforces uncorrelated incoming and outgoing messages M_{ij} and M_{ji}. The learning rules for κ gain-modulates beliefs according to how strongly incoming and external messages are correlated with each-other, and therefore fights against spurious belief amplifications. Importantly, we checked that these learning rules applied to an acyclic graph converge to $(\alpha, \kappa) = (1, 1)$ which corresponds to the BP algorithm (which is optimal for exact inference in acyclic graphs).

This purely heuristic approach results in suboptimal inference (see **Figure 19**) but nevertheless, can suppress polarization while improving the information sharing ability of the model social networks.

IV.4.5. Measures of radicalization of polarization

In **Figure 21**, radicalization was computed by averaging the mean absolute belief $|B_i|$ over all nodes, test trials and network structures (6 randomly generated networks were tested for each combination of K and β). Polarization was measured as the standard deviation of the beliefs, computed over nodes within a single trial, and then averaged over trials and network structures.

IV.5. Supplementary Material

IV.5.1. Relationship between probability and belief

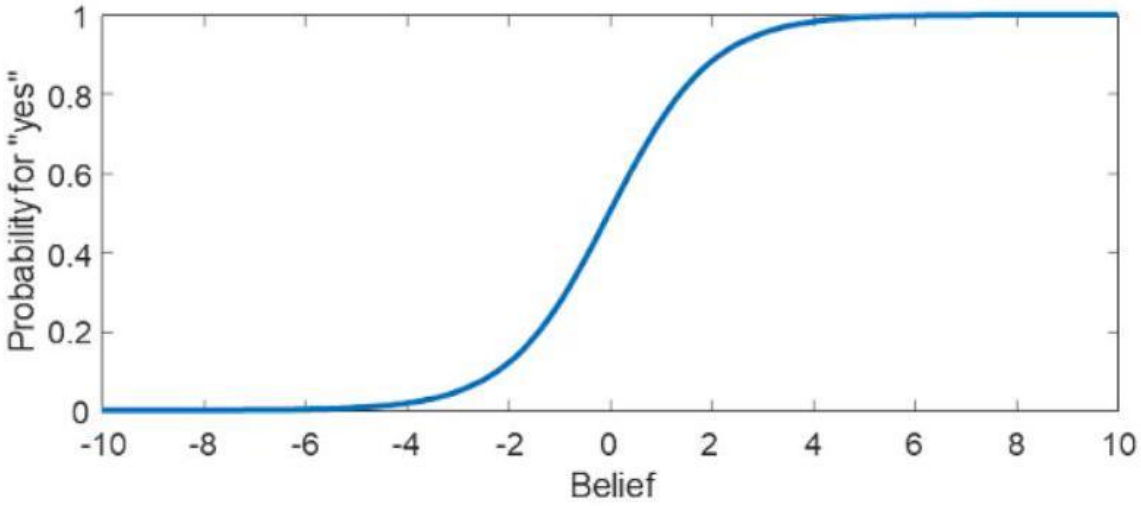
In the main text, the belief B_i is defined as a function of the estimate (marginal) probability of x_i to be "yes" $b(x_i = \text{"yes"})$:

$$b(x_i = \text{"yes"}) = \text{sigmoid}(2B_i) = \frac{\exp(2B_i)}{1 + \exp(2B_i)}$$

Note that this is equivalent to:

$$B_i = \frac{1}{2}\log\left(\frac{b(x_i = \text{"yes"})}{b(x_i = \text{"no"})}\right)$$

meaning that the belief of agent i represents half of the log-likelihood ratio associated to variable x_i.



Supplementary Figure 5 : Relationship between the belief B_i and the "Probability of yes" $b(x_i = \text{"yes"}) = \sigma(2B_i)$. Note that for a belief above 3 , there is a quasicertainty that the proper answer to the question is "yes", with no room left for doubt or changing one's mind.

IV.5.2. Mathematical appendix

Pairwise graph models

We model a social network as a pairwise, binary, and a-directed graph. Each member of the social network (each node) is associated with a binary random variable, $x_i \in \{-1; +1\}$. The external information received by this node, playing the role of a prior, is entered as a uni-modal factor $\psi_i(x_i)$. Meanwhile, social interactions (links in the graph) correspond to pairwise factors $\psi_{ij}(x_i, x_j)$. This factor describes the strength of the coupling between x_i and x_j, or equivalently, how strongly the belief of agent i influences the belief of agent j. The graph as a whole represents a joint probability distribution, $p(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, factorizing into a product of conditionally independent factors (Koller and Friedman, 2009; Wainwright and Jordan, 2008):

$$p(\boldsymbol{x}) \propto \prod_{(i,j)} \psi_{ij}(x_i, x_j)\psi_i(x_i)$$

Ideally, an agent's opinion would be obtained by computing the marginal posterior probability $p_i(x_i) \equiv \sum_{\boldsymbol{x} \setminus x_i} p(\boldsymbol{x})$. Typically, this (very large) sum is intractable. Local message-passing schemes as those described below can only compute an approximation for these marginal posteriors, called beliefs $b_i(x_i) \approx p_i(x_i)$.

Inputs to the social network are provided as external messages $M_{\text{ext} \to i} \equiv \log\left(\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)}\right)$, or equivalently, $\psi_i(x_i) \propto \exp\left(\frac{1}{2} M_{\text{ext} \to i} x_i\right)$. $\propto$ indicates a proportionality (factors have to be normalized to be interpreted as a probability). Note that positive external messages favor $x_i = +1$, and viceversa for negative external messages. For simplicity, we assume symmetrical factors: $\psi_{ij}(x_i, x_j) \propto \exp(J_{ij} x_i x_j)$ with $J_{ij} = J_{ji}$. With this formulation, the social network corresponds to an Ising model (also known as a Boltzmann machine). J_{ij} modulates the trust existing between the two agents. For example, large positive values for J_{ij} imply that the pairwise factor is very large for $x_i = x_j$, but very small if $x_i \neq x_j$; Thus, one agent will strongly influence the other's opinion in the direction of their own. In contrast, if J_{ij} is close to zero, the two agents do not believe (do not influence) each other. Since people tend to communicate preferentially with whom they trust, we limited ourselves to strictly positive interactions in the study, that is, $J_{ij} > 0$ for two connected agents, albeit the framework can generalize to any type of interaction.

Belief Propagation

The Belief Propagation algorithm (Pearl, 1988) is a variational inference method which performs approximate inference on a probabilistic graph. It is a message-passing algorithm: it approximates the marginals of the distribution by making variable nodes x_i share all the probabilistic information available with the rest of the network by sending messages to other variable nodes. The algorithm consists of running iteratively the following update message equation on the graph, where we consider here pairwise factor graphs or Markov networks:

$$m_{i \to j}^{\text{new}}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j) \psi_i(x_i) \prod_{k \in \mathcal{N}(i) \setminus j} m_{k \to i}^{\text{old}}(x_i)$$

where $\mathcal{N}(i)$ is the set of neighbors of node x_i in the graph. Messages are for instance initialized uniformly over the nodes $\left(m_{i \to j}(x_j) = 1/\mathcal{N}(j)\right)$. Once messages have

converged (or after some given number of iterations), approximate marginal probabilities (or beliefs) are computed as:

$$b_i(x_i) \propto \psi_i(x_i) \prod_{k \in \mathcal{N}(i)} m_{k \to i}(x_i)$$

A crucial feature of the BP algorithm is the message exclusion principle ($k \in \mathcal{N}(i) \setminus j$ in Equation (S4)): to compute $m_{i \to j}$, all messages coming to node x_i are taken into account and combined, except the message in the opposite direction $m_{j \to i}$.

Belief Propagation is exact when applied to acyclic probabilistic graphs, but can perform very poorly otherwise.

In the particular case of binary variables, BP takes a particularly simple form:

$$\begin{cases} M_{i \to j}^{\text{new}} = f_{ij}\left(B_i - M_{j \to i}\right) \\ B_i = \sum_{j \in \mathcal{N}(i)} M_{j \to i} + M_{\text{ext} \to i} \end{cases}$$

where $M_{i \to j} \equiv \frac{1}{2} \log\left(\frac{m_{i \to j}(x_j = +1)}{m_{i \to j}(x_j = -1)}\right)$ represents the information about variable x_j brought by variable x_i, and $B_i \equiv \frac{1}{2} \log\left(\frac{b_i(x_i = +1)}{b_i(x_i = -1)}\right)$ is by definition the log-odds (odds is a synonym for likelihood ratio or probability ratio). The approximate marginal probabilities are given by $b_i(x_i = \pm 1) = g(\pm B_i)$ where $g(x)$ is the standard logistic function.

Since we consider symmetrical graphs, function f_{ij} takes a simple form (see also Mooij and Kappen, 2007):

$$f_{ij}(x) = \phi^{-1}\left(\phi(J_{ij})\phi(x)\right)$$

where ϕ is the hyperbolic tangent tanh. f_{ij} is therefore a sigmoidal function of x, bounded between $-J_{ij}$ and $+J_{ij}$.

Circular Belief Propagation

The Circular Belief Propagation (CBP) algorithm is an extension of Belief Propagation (BP) which improves the quality of probabilistic inference in cyclic graphs compared to BP (Bouttier, 2021). CBP is based on the generalization of the Bethe approximation (Wiegerinck and Heskes, 2002).

The update message equation for Circular Belief Propagation is:

$$m_{i\to j}^{\text{new}}\left(x_j\right) \propto \sum_{x_i} \psi_{ij}\left(x_i, x_j\right)\left(\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}\right)^{\kappa_i}$$

and the beliefs are computed using:

$$b_i(x_i) \propto \left(\psi_i(x_i) \prod_{k\in\mathcal{N}(i)} m_{k\to i}(x_i)\right)^{\kappa_i}$$

BP corresponds to the particular case $(\boldsymbol{\alpha}, \boldsymbol{\kappa}) = (\mathbf{1}, \mathbf{1})$. The main conceptual difference with BP is the fact that the message in the opposite direction $m_{j\to i}$ is partly taken into account for the computation of $m_{i\to j}$ (more specifically, through the term $m_{j\to i}(x_i)^{1-\alpha_{ij}/\kappa_i}$). Other than that, κ acts as a reweighting factor of beliefs. Note that α_{ij} is assigned to the undirected edge (i, j), while κ_i is assigned to the variable node x_i. In this case of an Ising model, CBP in the log-domain takes a form almost as simple as BP:

$$\begin{cases} M_{i\to j}^{\text{new}} = f_{ij}\left(B_i - \alpha_{ij} M_{j\to i}\right) \\ B_i = \kappa_i \left(\sum_{j\in\mathcal{N}(i)} M_{j\to i} + M_{\text{ext}\to i}\right) \end{cases}$$

In practice, to implement the message-passing scheme, we used a damped version of the algorithm:

$$M_{i\to j}^{\text{new}} = (1-\tau) M_{i\to j} + \tau f_{ij}\left(B_i - \alpha_{ij} M_{j\to i}\right)$$

with $\tau = 0.2$. This avoids numerical issues, in particular when applying the algorithm to very large graphs. A similar damping mechanism was applied for BP, for which damping is well known to help (Pretti, 2005).

The CBP algorithm used in our numerical experiments is provided below as pseudo-code (where the maximum iteration number was set at 100, and the initial value for messages was set at 0):

Algorithm 1 Damped Circular Belief Propagation in an Ising model

for all directed edges $i \to j$ **do**
  $M_{i\to j} \leftarrow$ some initial value
end for
repeat
  for all directed edges $x_i \to x_j$ **do**
    Update the messages $M_{i\to j}^{\text{new}}$ using Eq. (S11) and (S10b)
  end for
  $M \leftarrow M^{\text{new}}$
until convergence or maximum iteration number
for all nodes x_i **do**
  Compute the final beliefs B_i using Eq. (S10b)
end for

Theoretical foundation for the BP and CBP algorithms

It has been shown that the BP algorithm minimizes the Bethe free energy (Heskes, 2002). The Bethe free energy is defined as the KL divergence (up to a constant) between $p(\boldsymbol{x})$ and an approximate distribution $b(\boldsymbol{x})$ factorizing as a function of its marginal and pairwise posteriors $b_i(x_i)$ and $b_{ij}(x_i, x_j)$ as follows:

$$b(\boldsymbol{x}) \approx \prod_{i,j} \frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \prod_i b_i(x_i)$$

This approximation is exact only if the graph contains no cycle (it is a tree).

Fractional BP (Wiegerinck and Heskes, 2002) is based on assuming a more general form for the approximate distribution factorization:

$$b(\boldsymbol{x}) \approx \prod_{i,j} \left(\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right)^{1/\alpha_{ij}} \prod_i \left(b_i(x_i) \right)^{1/\kappa_i}$$

The KL divergence between this distribution and $p(\boldsymbol{x})$ defines a generalized Bethe free energy (see also Yedidia et al., (2005) and Wainwright et al., (2005). This KL divergence is minimized by the Fractional BP algorithm. The Circular BP algorithm is very close to Fractional BP (Bouttier, 2021), which is especially visible in Ising models: they differ in the particular form of the function f_{ij}. More precisely, one algorithm derives from the other by using the following (tight) approximation: $\frac{1}{\alpha_{ij}} \phi^{-1}\big(\phi(\alpha_{ij} J_{ij}) \phi(x)\big) \approx \phi^{-1}\big(\phi(J_{ij}) \phi(x)\big)$. Therefore, Circular BP has very similar properties to Fractional BP, and thus minimizes (approximately) the KL divergence between $b(\boldsymbol{x})$ (assuming it takes the form of Equation $(S13)$) and $p(\boldsymbol{x})$.

Chapter V

General discussion

Context of the thesis

The number of published research on adherence to conspiracy theories (CTs) and its underlying mechanisms has literally exploded over the last decade. The resurgence of these unshakeable beliefs in contexts of distress and uncertainty has led some authors to propose the emergence of CTs as a possible coping mechanism. Indeed, socio-political crises can trigger feelings of anxiety, uncertainty and subjective lack of control over one's own life. In some circumstances, like during the COVID-19 lockdowns, these feelings were also amplified by a sense of isolation and loneliness (Palgi et al., 2020). Conspiracy beliefs have even been proposed as an adaptative strategy designed to cope with stress by addressing the needs to restore certainty, a sense of control and social ties.

However, while CTs might serve adaptative functions on the short run, they also have several detrimental effects in the long term. CTs have notably been show to influence decisions related to important issues, such as healthcare (Bertin et al., 2020; Bird and Bogart, 2005; Jolley and Douglas, 2014; Marinthe et al., 2020) or socio-political engagement (Butler et al., 1995; Imhoff et al., 2021). As the pace of socio-political crises is gaining momentum, a better understanding of CTs constitutes a crucial challenge of our time. If socio-demographic features of CTs have been effectively highlighted in the literature in recent years (see Douglas et al., 2019), several gaps remain in terms of underlying mechanisms and behavioral impacts of these beliefs.

First, some probabilistic biases associated with CTs have been recently investigated such as motivated reasoning, jumping-to-conclusions and conjunction fallacy (see **Table 1**, for a review see Gagliardi, 2022), but little is known about the formal mechanisms at the roots of these associations. Furthermore, data on certain important processes previously linked with rigid beliefs are lacking. This is for instance the case for attentional resources allocation, which can be easily explored using salience attribution tasks or with questionnaires.

Proved effective in accounting for other inflexible beliefs, the Bayesian framework appears well-suited to address this gap. Surprisingly, despite some noticeable exceptions, very few papers have tried to approach CTs through a

computational lens. These attempts have notably fitted computational models to high-level cognitive functions, using tasks such as the reversal-learning paradigm (Suthaharan et al., 2021; Zhang et al., 2022). While the principles governing the Bayesian approach of cognition also rule perceptual processing, the potential association between CTs and lower-level biases is still prone to debate and emerged as an important question to resolve.

Some research groups proposed the use of lower-level stimuli in pure behavioral experiments (such as *illusory pattern detection*, (Hartmann and Müller, 2023; Müller and Hartmann, 2023; van Prooijen et al., 2018), which unfortunately resulted in conflicting findings (Dieguez et al., 2015). The association between CTs and illusory pattern detection was also proposed rooted in a common compensatory mechanism designed to address control threats (Hogg et al., 2010; Kay et al., 2008; Walker et al., 2019; Whitson and Galinsky, 2008). Yet, papers that investigated the potential association between CTs and personal sense of control also yielded conflicting results (Hart and Graether, 2018; Kofta et al., 2020; Nyhan and Zeitzoff, 2018; Pantazi et al., 2022; Stojanov et al., 2020, 2022; Sullivan et al., 2010; van Elk and Lodder, 2018; van Prooijen and Acker, 2015; Whitson and Galinsky, 2008). To explain these discrepancies, some authors stressed the role uncertainty could exert on the association (Stojanov et al., 2020) and called for more naturalistic designs that could integrate this covariate (Stojanov et al., 2022).

Finally, approaching extreme beliefs with Bayesian models designed to reproduce the dynamics of belief propagation in social networks could complement research conducted at the individual-level. This particular line of research is built on the idea that an extreme world-view could result from optimal Bayesian inference performed in intrinsically sub-optimal network structures (Madsen et al., 2017). However, no clear solution taking these structures into account was proposed until then, notably to better understand how we could counter polarization and radicalization phenomena.

Summary of the main findings

In this thesis, my main objective was to address these different gaps through four complementary experimental approaches. First, I investigated the associations between CTs, aberrant salience attribution and health-related decisions in the context of the COVID-19 pandemic (Chapter II.1.). Testing 691 participants online, we demonstrated that incorrect assignment of salience to innocuous stimuli was associated with both CTs and vaccine hesitancy. We found that this association was domain-specific by highlighting a stronger relationship between hypersalience and conspiracies revolving around health-issues and personal well-being. While the cognitive factors associated with CTs have been extensively explored, this thesis provides, to our knowledge, a first evidence for the role of biased salience attribution.

In the second experiment (Chapter II.2.), I questioned the debated link between CTs and perceived *lack of control* (LoC). To address this matter, I used self-reported assessments combined with a more objective measure of the actual control over a visual stimulus. We demonstrated that this association could be uncovered by conditions of heightened real-world uncertainty and was mediated by individual stress vulnerability. Again, I evidenced the domain-specific nature of this association in 411 participants tested online, notably showing that different types of control were associated with different themes of conspiracy ideations. We can note that for the first time the LoC/CTs association was validated using an experimental measure of control based on bistable perception, paving the way for the use of more objective measures of the sense of control in future work.

In the third experiment (Chapter III), I investigated further the mechanistic processes underpinning CTs and how they relate to stress. We did so by coupling a perceptual task with a computational model able to capture hidden mechanistic variables during periods of heightened real-world uncertainty. Using a repeated-measures design testing 623 individuals before and after a major stressful political event, I was able to identify different stages of CTs rigidification, together with their respective inference mechanisms. Specifically, using the *Circular Inference* (CI) model, we evidenced that when uncertainty peaks, there is a significant association between CTs and sensory overweighting, later followed by an increased reliance on priors in participants who decreased their stress level.

Finally, in the last chapter (Chapter IV), I demonstrated how sub-optimal network structures (i.e., echo-chambers) can aberrantly increase confidence leading to polarization and radicalization of beliefs. We proposed the *Circular Belief Propagation* (CBP) as a novel algorithmic solution designed to bring agents closer to informed rationality. We finally demonstrated its validity in unleashing the true information sharing capabilities of online communities of increasing size using realistic network structures.

Taken together, these results form the basis on which I will discuss the adaptative value of CTs. First, I will discuss how findings from my PhD thesis support the idea that CTs can serve as a compensatory strategy designed to address uncertainty and control threats. Second, I will discuss why the CI model provides a comprehensive framework to understand the inferential biases underlying the dynamical process of belief rigidification under real-world uncertainty. Finally, I will illustrate how these results account for the multifaceted aspect of CTs and highlight subtle nuances in the mechanisms underlying different types of CTs.

Conspiracy ideations in the face of uncertainty

Endorsing conspiracy theories has been largely proposed as a coping mechanism that responds to the need of making sense of the world when facing uncertainty (Douglas et al., 2017). We saw that several cognitive biases might be adopted to reduce uncertainty, which could in turn play a crucial part in the emergence of CTs. Reinforcing adherence to conspiracy ideations was thus proposed to help making sense of socio-political crises at a low cognitive cost. Congruent with this idea, some authors proposed that CTs might not result from a cognitive dysfunction per se, but rather from a range of cognitive biases attempting to fulfil adaptative functions (Bortolotti, 2023).

However, these coping strategies stay maladaptive since the endorsement of rigid beliefs can have a drastic impact over an individual's life and decisions. Furthermore, CTs were shown associated with some inefficient coping strategies such as avoidant and immature defence strategies (Gioia et al., 2023; Marchlewska et al.,

2021). While it can be argued that CTs serve an adaptative purpose, they still constitute an "*imperfect response to psychological and epistemic needs*" (Bortolotti, 2023). According to Coltheart (2010), adhering to an erroneous hypothesis is not pathological as soon as the individual is able to integrate new evidence that can discredit the adopted belief. In the case of CTs, integrating this new evidence could prove very costly, as these beliefs are often embedded in a broader conspiracist system where each theory is consistent with a fundamental worldview that other people might be malevolent and hold harmful intentions. This would explain why several motivational and reasoning biases can prevent the abandonment of CT beliefs to preserve the credibility of a wider belief system. Congruent with this idea the role of *confirmation bias*, *bias against disconfirmatory evidence* and belief updating differences have also been evidenced in CTs (Georgiou et al., 2021; Kabengele et al., 2023; Kuhn et al., 2022; McHoskey, 1995; Pytlik et al., 2020; Suthaharan et al., 2021; Woodward et al., 2007; Zhang et al., 2022). Future work could challenge this assumption that differences in information processing have a protective function. For example, Zhang and colleague (Zhang et al., 2022) found that CTs were associated with differences in probabilistic processing using a decision task that involved deceitful agents (i.e., thieving leprechauns). It would be interesting to see if these results can be replicated using a more neutral version of the task, like choosing between symbols instead of anthropomorphized characters.

Findings from my PhD bolster the idea that uncertainty peaks might alter subjective sense of being safe and in control. We notably show that around a high-stakes political event, uncertainty can trigger LoC, both subjectively and objectively measured. Some individuals would be particularly sensitive to these control threats and might exhibit higher levels of distress. In an attempt to reduce this stress and make sense of the world, they would adopt an exploration strategy rooted in a more intuitive thinking mode. This strategy could lead to the attribution of enhanced importance to random stimuli from the environment (as shown in experiment 1), accompanied by an overconfidence in this new information (as show in experiment 3). The use of the CI model notably allowed us to evidence a significant association between CTs and sensory overweighting. While computational approaches of CTs have already stressed the influence of biased belief-updating processes using high-level cognitive tasks, this is the first successful attempt at fitting perceptual data to account for CTs.

One of the strengths of this PhD lies in the proposal of a dynamic view of belief rigidification. Indeed, we know that CTs can be stable over a period of several years (Stieger et al., 2013). How can we explain such persistence? Coltheart (Coltheart, 2010) argued that a good model of delusion should account for: (i) the endorsement of implausible hypothesis at an early stage, referred to as "*belief adoption*", and (ii) belief maintenance despite the existence of discordant evidence at the "*belief maintenance*" stage. Drawing on this idea, while conspiracy adherence can efficiently reduce stress in the short term, a complete investigation aimed at better understanding its later effects, when the threat is no longer present, is called for. Some psychosocial accounts of CTs based on this two-factors model have been proposed, stressing the contribution of epistemic distrust (Pierre, 2020). My PhD work complements this initiative by providing a dynamic Bayesian inference account of the different stages of information processing underpinning the emergence and rigidification of CTs.

By combining the strengths of computational modeling with a naturalistic design able to capture real-world uncertainty, the results brought here complement the existing literature in proposing a new mechanism for belief rigidification dynamics. Using the CI model, we showed that after uncertainty resolution, the decrease in stress was associated with a concomitant increase in CTs and the progressive reliance on prior information. This could be interpreted as if individuals who mainly subscribed to conspiracy theories seek to reduce the potential cost of revising their beliefs by self-reinforcing their established worldview. By shifting from an exploration strategy to an exploitation strategy integrated into highly cognitive processes, these individuals overweighted their prior beliefs, while discarding counterevidence.

These findings stress how crucial it is to account for the socio-political context nurturing CTs. We notably showed that real-world uncertainty could uncover CTs mechanisms. That means that the time at which we assess probabilistic processes matters. A same mechanism might be highlighted during uncertain times but remains hidden six months later in the same population. Congruent with Stojanov and colleagues (Stojanov et al., 2020), we preconise to further explore and account for the impact of perceived uncertainty and uncertainty tolerance in future works. Such measures could be performed using simple visual analog scales, such as "*How uncertain do you feel about the future?*" to more sophisticated tools, such as the *Intolerance to Uncertainty Scale* (Freeston et al., 1994).

A domain-specific construct

The Bayesian account of CTs I proposed in this thesis is embedded in a dimensional approach which posits that beliefs can be situated on a continuum. Extreme conspiracy views ("*The Earth is flat*") and more reasonable assumptions ("*Imperfect vaccines have been put on the market by greedy labs to the detriment of public health*") could then be explained by common mechanisms of different intensities rooted in the tendency of inferential bias to be amplified by the individual vulnerability to stress. Several authors have nevertheless stressed the importance of taking CTs diversity into account (Franks et al., 2013, 2017; Klein et al., 2018). In line with this idea, I took a closer look at the specific conspiracy themes that were most closely related to the different mechanisms investigated, allowing us to highlight the domain-specific nature of CTs.

Findings presented across the different chapters provide new insights indicating that different mechanisms could potentially account for the polymorphic nature of CTs. We notably show that hypersalience and perceptual control, two cognitive-perceptual factors that have been proved to play a part in delusions, were most strongly associated with CTs clustered as "*extraterrestrial cover-up*" and "*personal well-being*". These two dimensions revolve around the supernatural, health issues and mind-controlling tech themes that are also commonly found in delusional ideations.

On the other hand, the subjective sense of global control was associated with conspiracies related to "*control of information*" and "*government malevolence*", an association mediated by political distress. These two dimensions that were overall the most represented in our samples reflected concerns at play in the period of our testing marked by heightened socio-political uncertainty. These findings suggest that the need to cope with the stress induced by the feeling of not being in control of the political situation triggered the adherence to CTs directly linked to the perceived threat.

The domain-specificity evidenced here could also account for some literature discrepancies. For example, in the third chapter, we revealed a negative association between age and CTs, mirroring the findings of some previous papers (Wagner-Egger et al., 2022), but contrasting with others which reported the reverse association

(Marinthe et al., 2020). In reality, these differences could reflect the fact that CTs under investigation could relate to threats perceived differently across different age groups. For example, Marinthe and colleagues (Marinthe et al., 2020) found a positive association between age and COVID-19-related conspiracies, which might simply reflect the fact that older adults are at higher risk of severe disease after COVID-19 infections (Federico and Malka, 2018).

To summarize, we can see that while cognitive-perceptual factors provide a global account for conspiracy adherence, there might be subtle differences in the mechanisms at play in the emergence of different types of CTs. Borlototti (Bortolotti, 2023) advanced that belief updating differences are more likely to be a product of environmental contingencies at play than an information processing dysfunction. In an effort to bridge these opposing viewpoints, we propose CTs as multifactorial constructs modulated by different factors of vulnerability and tested these assumptions experimentally. On one hand, transient societal distress may preferentially impact individuals who are more susceptible to stress and anxiety, triggering a more important need to cope with feelings of loss of control and uncertainty. These individuals might be more inclined to endorse conspiracies directly related to societal issues. Conversely, delusion-like beliefs might rely more on perceptual aspects of the inference system triggered by more subtle vulnerability factors closer to those previously found in schizotypy. While the existence of these distinct constructs should be explored, we can already note that the different dimensions of CTs are strongly correlated together, suggesting potential interactions between these different mechanisms at play in the emergence of CTs.

Social components of conspiracy beliefs

I think we could also challenge whether the belief that “an international coalition of governments commanded the intentional spread of a virus in order to conceal the mass implantation of 5G chips through harmful vaccines” is really a “*simple, easy-to-understand*” explanation of the Covid-19 pandemic. While our model captures why and how an individual might endorse a first conspiracy explanation before sinking into more

complex conspiratorial narratives in a self-reinforcing system, a complementary social approach is needed to fully understand adherence to such complex constructs.

As for any framework, the Bayesian account of beliefs presented throughout my thesis was not fully able to capture the multifaceted and complex nature of CTs. First, to keep the model simple, a wide range of socio-demographic factors and personality traits previously proposed associated with conspiracy endorsement were not considered. Among them we can notably cite the desire for uniqueness (Imhoff and Lamberty, 2018; Lantian et al., 2017), self-esteem and narcissism (Cichocka et al., 2016), masculinity (Adam-Troian et al., 2021), low agreeability (Swami et al., 2010), Machiavellianism (March and Springer, 2019) and openness to experience (Swami et al., 2010).

Some authors proposed *distrust* as a key social component of CTs (Goertzel, 1994; Wagner-Egger and Bangerter, 2007). CTs have indeed been found associated with higher level of distrust towards authorities (Imhoff and Bruder, 2014; Imhoff and Lamberty, 2018) and scientific knowledge (Lobato et al., 2014). According to Pierre's two-factors model (Pierre, 2020), the rejection of official narratives could constitute a socio-cultural response to breaches of trust, inequities of power, and existing racial prejudices. As a consequence, some individuals would start searching for alternative explanations that would be concealed from the public. This component would be heightened in populations subject to systemic violence who might have internalized stronger priors for distrust in authorities. Such a phenomenon would explain why CTs can be particularly revived during social crises that bring to light power inequities, such as the riots that followed George Floyd's assassination.

This conceptualization of CTs sheds light on some specificities of the present thesis I also wanted to discuss. First, we only investigated *Western Educated Industrialized Rich and Democratic* (W.E.I.R.D.) countries. Future research endeavors might be extended to more diversified populations. Furthermore, some heterogeneity could also hide within the W.E.I.R.D. countries we assessed. Indeed, while we attempted to frame similar political events across the samples we recruited, we can question whether these events are really comparable. If the uncertainty inherent to the French Presidential elections had a clear climax and resolution, worrying political upheavals were still underway at the time of our second test of the US sample (such as the United States Capitol attack or the *Black Lives Matter* movement-related riots).

Moreover, the event we framed in the UK (the Brexit Implementation) might have elicited less uncertainty than in the other countries we tested. While the uncertainty surrounding the political event was total in the French and American populations, the societal issue had already reached a partial resolution for the UK sample (participants already knew whether the Brexit was going to happen, and the uncertainty may more in how it would change everyday life after January 31, 2020).

Moreover, in each country, political affiliations were able to substantially influence the emotional and cognitive impact of the conflicting political events we studied. Congruent with this idea, a recent study showed that CTs are associated with extreme political views and that this link is modulated by deprivation of political control (Imhoff et al., 2022). It would have been really interesting to have access to this information for our participants and take it into account in the analyses. Such a factor might partially explain the amount of variance in our data.

In line with previous work (Reed et al., 2020), my PhD work aimed at addressing general domain decision-making processes involved in CTs. While it partially explains the underpinning mechanisms of the emergence and maintenance of these rigid beliefs, adding a social lens might surely help to better understand CTs phenomenology. We could for example use probabilistic tasks framed in a more social context, such as the ones already used in paranoia (Barnby et al., 2020, 2022). Simonsen and colleagues (Simonsen et al., 2021) for instance used a social variant of the beads task, designed to assess how participants integrated other agent's beliefs when reaching decisions. They found that patients with schizophrenia or schizo-affective disorders tended to over-weight and overcount their own experience, while under-weighing information coming from external agents. Crucially, they also found that the *Circular Inference* best described patient's behavior compared to other computational models, suggesting that CI is still a good option to account for social inference processing.

Limitations and perspectives

The experimental paradigms used in the thesis are not exempt from limitations. We will review the limitations and constraints related to online testing, the perceptual task we administered, the psychometric assessments we performed and finally the computational framework we used.

First, the web-based nature of the experimental work might rise general questions. While online settings have gained increasing popularity over recent years due to substantial advantages, such as reduced demand characteristics, automation, and generalization of results to wider populations (Birnbaum, 2004; Reips, 2000, 2002a, 2002b), online paradigms can also be subject to several biases.

Online testing gives easy access to a wide panel of participants that might be more diversified than the populations usually recruited for lab experiments. This heterogeneity might partially explain the small amplitude of our effects. In addition, it is questionable to what extent participants who decide to devote their time and cognitive resources to online studies are representative of the general population. We can note that a substantial amount of these participants had to be excluded on the basis of failed attentional checks, questioning their general engagement with the protocol, and requiring strict definitions for outliers detection.

Second, we might wonder whether administrating online questionnaires might elicit different responses from those we would have obtained through traditional experimentation. While some authors showed that the psychometric characteristics of online versus traditional in-lab administered questionnaires stay comparable (Riva et al., 2003; Simmons et al., 2023), it has also been suggested that people tend to lie more online, either to look more attractive or because “everybody does it” (Drouin et al., 2016). However, while participants may lie about their demographic characteristics, another study showed that online tests do not facilitate falsified questionnaire answers in order to project a certain self-image (Grieve and de Groot, 2011). We would like to argue that the sense of anonymity elicited by online methods appear to foster more honest answers.

Moreover, a perceptual task such as the *Necker Cube* paradigm might prevent participants from “faking” answers and be less subject to social desirability biases.

Nevertheless, the online implementation of such sensitive psychophysical protocol raises other questions. We know that the use of a problem-solving experiment displays a good concordance between online and in-lab testing (Dandurand et al., 2008), but comparisons between the two methods on neuropsychological performance tests yielded more modest results (Simmons et al., 2023). This is why in this thesis, I initially checked the good online-inlab concordance in the Necker Cube task performances using an independent sub-sample of participants.

Our procedure could definitely be improved in future protocols, but I would like to stress the fact that the data presented in this thesis represents the first attempt at implementing the NC task online; an approach that has yielded positive results, but also reflects the ability of researchers to quickly adapt to unprecedented constraints, such as those imposed by the 2020 pandemic.

Again, even if the NC task versions used in this thesis were already validated in previous research papers from the team (Leptourgos, Notredame, et al., 2020), it can always be improved. The main point of discussion that were brought to my attention while presenting my data along these past four years is the need to control for the confounding effects of eye-movements. While eye-position and eye-movements can influence neutral and consciously biased perception of the *Necker Cube* and thus constitutes an important factor to control for (Einhäuser et al., 2004; Polgári et al., 2023), eye-tracking devices could even represent a good alternative to self-reports in assessing perceptual dynamics (Polgári et al., 2020). Since the beginning of my PhD thesis, online assessment has gained increasing attention and webcam-based eye-tracking solutions have been refined, providing exciting possibilities for future research.

Also, as mentioned earlier in this discussion, more precise measures of uncertainty but also of trait-anxiety could constitute an interesting addition to future protocols. The same could be done for salience, which could be assessed at different levels of processing. Chun and colleagues (Chun et al., 2019) found that while schizotypy was associated with self-reported aberrant salience experiences similar to the ones we measured in the first chapter, it was not associated with visual salience. They argue that perceptual and cognitive salience might be rooted in different mechanisms. While this thesis provides first evidence for an implication of aberrant salience in CTs and health-related attitude, further research will have to decipher the implication of these different levels of processing, notably by adding a perceptual task.

These points provide good directions for future research endeavours. Similarly, the Circular Inference framework have been proven equally effective in accounting for cognitive and perceptual inference. However, how these two levels relate to each other has never been properly investigated yet. That is why a follow-up study is scheduled to assess for the concordance degree between the CI model parameters that can be extracted from the *Necker Cube* paradigm and the *Fisher task* in the same participants.

Finally, I started this thesis reminding that inference processes could be studied at three complementary levels (Marr, 1982). A pilot study is currently being conducted in the lab that aims at investigating the hardware level, i.e., the neural implementation of the inference processes underlying rigid beliefs. Relying on high-density electroencephalographic recordings, this project, initially scheduled at the beginning of my PhD, will notably rely on a novel version of the NC task and will search for physiological signatures of *Circular Inference*… and the “loop should be complete” !

Conclusion: towards interventional practices

Despite few limitations, my PhD thesis addresses a key dimension of belief rigidification through the study of belief propagation. Similarly to what has been hypothesized at the neuronal level (Jardri and Denève, 2013), sub-optimality of information regulation in online social-networks can lead to extreme confidence levels. Furthermore, in addition to the mechanistic account it provides, this last contribution proposes a valid algorithmic solution to "*rescue rationality*", i.e., to bring agents closer to informed rationality. The strength of this proposal lies in its no-censorship policy, that is, removing redundant and irrelevant traffic information. In a previous attempt, other authors found that interventional practices such as psychological inoculation (i.e., the preparation of individuals to resist persuasion and the influence of fallacious arguments) only yielded temporary results defeated by echo chambers persistence (Pilditch et al., 2022). We argue that addressing the problem at its source, by targeting the algorithmic flaws causing echo-chambers in the first place might constitute a more efficient strategy on the long run.

To conclude, from what we learned across the thesis, and more broadly since the pandemic, addressing belief rigidification constitutes a major challenge that will only gain importance with the increase in socio-political crises and the acceleration of information. While chapter IV offers the beginnings of a solution to this issue, real-life interventions are needed as well. In a recent review, O'Mahony and colleagues (O'Mahony et al., 2023) found that interventions designed at encouraging critical and analytical thinking were the most effective in countering overconfidence in conspiracy theories. In addition, the present findings also suggest that targeting uncertainty might be a good starting point for developing new complementary strategies. In the context of the COVID-19 pandemic, Farias and Pilati (Farias and Pilati, 2023) proposed to directly address the causes of CTs by promoting basic scientific knowledge that might 'fill the informational gaps' and assisting individuals in coping with uncertainty. In these uncertain times, I believe that the provision of timely, transparent and accurate information, together with professional psychological support, could help prevent CT beliefs and promote informed rationality.

Figures

References

Abelson RP (1988) Conviction. *American Psychologist* 43(4). US: American Psychological Association: 267–275.

Acerbi L, Vijayakumar S and Wolpert DM (2014) On the Origins of Suboptimality in Human Probabilistic Inference. *PLOS Computational Biology* 10(6). Public Library of Science: e1003661.

Adam-Troian J, Wagner-Egger P, Motyl M, et al. (2021) Investigating the Links Between Cultural Values and Belief in Conspiracy Theories: The Key Roles of Collectivism and Masculinity. *Political Psychology* 42(4): 597–618.

American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders (5th Ed.)*. Available at: https://www.zotero.org/briancroxall/items/P7AD22C8 (accessed 9 October 2023).

Appelbaum PS, Robbins PC and Vesselinov R (2004) Persistence and stability of delusions over time. *Comprehensive Psychiatry* 45(5): 317–324.

Baker C (2007) Theory-based Social Goal Inference. Epub ahead of print 2007.

Banisch S and Olbrich E (2019) Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology* 43(2). Routledge: 76–103.

Barnby JM, Bell V, Mehta MA, et al. (2020) Reduction in social learning and increased policy uncertainty about harmful intent is associated with pre-existing paranoid beliefs: Evidence from modelling a modified serial dictator game. *PLOS Computational Biology* 16(10). Public Library of Science: e1008372.

Barnby JM, Mehta MA and Moutoussis M (2022) The computational relationship between reinforcement learning, social inference, and paranoia. *PL o S Computational Biology* 18(7).

Baron RM and Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51(6). US: American Psychological Association: 1173–1182.

Barrat A and Weigt M (2000) On the properties of small-world network models. *The European Physical Journal B - Condensed Matter and Complex Systems* 13(3): 547–560.

Barron D, Morgan K, Towell T, et al. (2014) Associations between schizotypy and belief in conspiracist ideation. *Personality and Individual Differences* 70: 156–159.

Baumann F, Lorenz-Spreen P, Sokolov IM, et al. (2020) Modeling Echo Chambers and Polarization Dynamics in Social Networks. *Physical Review Letters* 124(4): 048301.

Baumann F, Lorenz-Spreen P, Sokolov IM, et al. (2021) Emergence of Polarized Ideological Opinions in Multidimensional Topic Spaces. *Physical Review X* 11(1). American Physical Society: 011012.

Bechara A, Damasio AR, Damasio H, et al. (1994) Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50(1–3): 7–15.

Beck J and Forstmeier W (2007) Superstition and belief as inevitable by-products of an adaptive learning strategy. *Human Nature* 18(1): 35–46.

Beck JM, Ma WJ, Pitkow X, et al. (2012) Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron* 74(1). Elsevier: 30–39.

Bertin P, Nera K and Delouvée S (2020) Conspiracy Beliefs, Rejection of Vaccination, and Support for hydroxychloroquine: A Conceptual Replication-Extension in the COVID-19 Pandemic Context. *Frontiers in Psychology* 11.

Bialystok E and Shapero D (2005) Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures. *Developmental Science* 8(6). United Kingdom: Blackwell Publishing: 595–604.

Bierwiaczonek K, Gundersen AB and Kunst JR (2022) The role of conspiracy beliefs for COVID-19 health responses: A meta-analysis. *Current Opinion in Psychology* 46: 101346.

Bilewicz M (2022) Conspiracy beliefs as an adaptation to historical trauma. *Current Opinion in Psychology* 47: 101359.

Binnendyk J and Pennycook G (2022) Intuition, reason, and conspiracy beliefs. *Current Opinion in Psychology* 47: 101387.

Bird ST and Bogart LM (2005) Conspiracy Beliefs About HIV/AIDS and Birth Control Among African Americans: Implications for the Prevention of HIV, Other STIs, and Unintended Pregnancy. *Journal of Social Issues* 61(1): 109–126.

Birnbaum MH (2004) Human research and data collection via the Internet. *Annual Review of Psychology* 55. US: Annual Reviews: 803–832.

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer New York, NY. Epub ahead of print 2006.

Bishop S, Duncan J, Brett M, et al. (2004) Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. *Nature Neuroscience* 7(2). 2. Nature Publishing Group: 184–188.

Blackmore S and Moore R (1994) Seeing things: Visual recognition and belief in the paranormal. *European Journal of Parapsychology* 10. United Kingdom: University of Edinburg: 91–103.

Boddington A (1963) Sejanus. Whose Conspiracy? *The American Journal of Philology* 84(1). Johns Hopkins University Press: 1–16.

Bortolotti L (2023) Is it pathological to believe conspiracy theories? *Transcultural Psychiatry*. SAGE Publications Ltd: 13634615231187243.

Bortolotti L, Ichino A and Mameli M (2021) Conspiracy theories and delusions. *Reti, saperi, linguaggi* (2/2021). Epub ahead of print 2021. DOI: 10.12832/102760.

Bouttier V (2021) *La propagation circulaire de croyances comme modèle d'inférences optimales et sous-optimales dans le cerveau: extension de l'algorithme et proposition*

d'implémentation neurale. These en préparation. Université Paris Cité. Available at: https://www.theses.fr/s190689 (accessed 7 October 2023).

Bouttier V, Duttagupta S, Denève S, et al. (2022) Circular inference predicts nonuniform overactivation and dysconnectivity in brain-wide connectomes. *Schizophrenia Research* 245: 59–67.

Brakoulias V and Starcevic V (2011) The Characterization of Beliefs in Obsessive–Compulsive Disorder. *Psychiatric Quarterly* 82(2): 151–161.

Breen R (1999) BELIEFS, RATIONAL CHOICE AND BAYESIAN LEARNING. *Rationality and Society* 11(4). SAGE Publications Ltd: 463–479.

Brennan JH and Hemsley DR (1984) Illusory correlations in paranoid and non-paranoid schizophrenia. *The British Journal of Clinical Psychology* 23 (Pt 3): 225–226.

Brett-Jones J, Garety P and Hemsley D (1987) Measuring delusional experiences: A method and its application. *British Journal of Clinical Psychology* 26(4): 257–265.

Bronstein MV, Kummerfeld E, MacDonald A, et al. (2022) Willingness to vaccinate against SARS-CoV-2: The role of reasoning biases and conspiracist ideation. *Vaccine* 40(2): 213–222.

Brotherton R and Eser S (2015) Bored to fears: Boredom proneness, paranoia, and conspiracy theories. *Personality and Individual Differences* 80. Netherlands: Elsevier Science: 1–5.

Brotherton R and French CC (2014) Belief in Conspiracy Theories and Susceptibility to the Conjunction Fallacy. *Applied Cognitive Psychology* 28(2): 238–248.

Brotherton R and French CC (2015) Intention seekers: conspiracist ideation and biased attributions of intentionality. *PloS One* 10(5): e0124125.

Brotherton R, French CC and Pickering AD (2013) Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale. *Frontiers in Psychology* 4. Frontiers.

Brotherton R, French C and Pickering A (2013) Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale. *Frontiers in Psychology* 4.

Brouwer GJ and van Ee R (2006) Endogenous influences on perceptual bistability depend on exogenous stimulus characteristics. *Vision Research* 46(20): 3393–3402.

Bruder M, Haffke P, Neave N, et al. (2013) Measuring Individual Differences in Generic Beliefs in Conspiracy Theories Across Cultures: Conspiracy Mentality Questionnaire. *Frontiers in Psychology* 4.

Brugger P, Regard M, Landis T, et al. (1993) 'Meaningful' patterns in visual noise: Effects of lateral stimulation and the observer's belief in ESP. *Psychopathology* 26(5–6). Switzerland: Karger: 261–265.

Bukowski M, de Lemus S, Rodriguez-Bailón R, et al. (2017) Who's to blame? Causal attributions of the economic crisis and personal control. *Group Processes & Intergroup Relations* 20(6). SAGE Publications Ltd: 909–923.

Butler LD, Koopman C and Zimbardo PG (1995) The Psychological Impact of Viewing the Film 'JFK': Emotions, Beliefs, and Political Behavioral Intentions. *Political Psychology* 16(2). [International Society of Political Psychology, Wiley]: 237–257.

Byford J (2011) *Conspiracy Theories*. London: Palgrave Macmillan UK. Available at: http://link.springer.com/10.1057/9780230349216 (accessed 10 October 2023).

Carcea I and Froemke RC (2013) Cortical plasticity, excitatory-inhibitory balance, and sensory perception. *Progress in Brain Research* 207: 65–90.

Castellano C, Marsili M and Vespignani A (2000) Nonequilibrium Phase Transition in a Model for Social Influence. *Physical Review Letters* 85(16). American Physical Society: 3536–3539.

Chadwick PD and Lowe CF (1994) A cognitive approach to measuring and modifying delusions. *Behaviour Research and Therapy* 32(3): 355–367.

Chater N, Oaksford M, Hahn U, et al. (2010) Bayesian models of cognition. *WIREs Cognitive Science* 1(6): 811–823.

Chun CA, Brugger P and Kwapil TR (2019) Aberrant Salience Across Levels of Processing in Positive and Negative Schizotypy. *Frontiers in Psychology* 10. Frontiers.

Cicero DC, Kerns JG and McCarthy DM (2010) The Aberrant Salience Inventory: a new measure of psychosis proneness. *Psychological Assessment* 22(3): 688–701.

Cichocka A, Marchlewska M and Golec de Zavala A (2016) Does self-love or self-hate predict conspiracy beliefs? Narcissism, self-esteem, and the endorsement of conspiracy theories. *Social Psychological and Personality Science* 7(2). 2. Sage: 157–166.

Cinelli M, De Francisci Morales G, Galeazzi A, et al. (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118(9): e2023301118.

Claridge G (1997) *Schizotypy: Implications for Illness and Health*. Schizotypy: Implications for illness and health. New York, NY, US: Oxford University Press.

Coltheart M (2010) The neuropsychology of delusions. *Annals of the New York Academy of Sciences* 1191: 16–26.

Constantinidis C and Steinmetz MA (2001) Neuronal Responses in Area 7a to Multiple-stimulus Displays: I. Neurons Encode the Location of the Salient Stimulus. *Cerebral Cortex* 11(7): 581–591.

Constantinidis C and Steinmetz MA (2005) Posterior Parietal Cortex Automatically Encodes the Location of Salient Stimuli. *Journal of Neuroscience* 25(1). Society for Neuroscience: 233–238.

Cook J and Lewandowsky S (2016) Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science* 8(1): 160–179.

Corlett PR, Murray GK, Honey GD, et al. (2007) Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. *Brain: A Journal of Neurology* 130(Pt 9): 2387–2400.

Corlett PR, Frith CD and Fletcher PC (2009) From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology* 206(4): 515–530.

Coyle JT (2006) Glutamate and schizophrenia: beyond the dopamine hypothesis. *Cellular and Molecular Neurobiology* 26(4–6): 365–384.

Dagnall N, Drinkwater K, Parker A, et al. (2015) Conspiracy theory and cognitive style: a worldview. *Frontiers in Psychology* 6.

Dagnall N, Denovan A, Drinkwater K, et al. (2017) Statistical Bias and Endorsement of Conspiracy Theories. *Applied Cognitive Psychology* 31(4): 368–378.

Dandurand F, Shultz TR and Onishi KH (2008) Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40(2): 428–434.

Darwin H, Neave N and Holmes J (2011) Belief in conspiracy theories. The role of paranormal belief, paranoid ideation and schizotypy. *Personality and Individual Differences* 50(8): 1289–1293.

Daws RE, Timmermann C, Giribaldi B, et al. (2022) Increased global integration in the brain after psilocybin therapy for depression. *Nature Medicine* 28(4): 844–851.

Deffuant G, Neau D, Amblard F, et al. (2000) Mixing beliefs among interacting agents. *Advances in Complex Systems* 03(01n04). World Scientific Publishing Co.: 87–98.

Denève S and Jardri R (2016) Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences* 11. Computational modeling: 40–48.

Deneve S and Jardri R (2016) Circular inference: Mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences* 11: 40–48.

Derome M, Kozuharova P, Diaconescu AO, et al. (2023) Functional connectivity and glutamate levels of the medial prefrontal cortex in schizotypy are related to sensory amplification in a probabilistic reasoning task. *NeuroImage* 278: 120280.

Devereux G (1970) *Essais d'ethnopsychiatrie Générale*. Paris: Gallimard.

Dieguez S, Wagner-Egger P and Gauvrit N (2015) Nothing Happens by Accident, or Does It? A Low Prior for Randomness Does Not Explain Belief in Conspiracy Theories. *Psychological Science* 26(11). SAGE Publications Inc: 1762–1770.

Dobbins AC and Grossmann JK (2010) Asymmetries in Perception of 3D Orientation. *PLOS ONE* 5(3). Public Library of Science: e9553.

Douglas KM, Sutton RM and Cichocka A (2017) The Psychology of Conspiracy Theories. *Current Directions in Psychological Science* 26(6). SAGE Publications Inc: 538–542.

Douglas KM, Uscinski JE, Sutton RM, et al. (2019) Understanding Conspiracy Theories. *Political Psychology* 40(S1): 3–35.

Dow BJ, Menon T, Wang CS, et al. (2022) Sense of control and conspiracy perceptions: Generative directions on a well-worn path. *Current Opinion in Psychology* 47: 101389.

Drinkwater K, Dagnall N and Parker A (2012) Reality testing, conspiracy theories and paranormal beliefs. *Journal of Parapsychology* 76(1). US: Rhine Research Ctr: 57–77.

Drouin M, Miller D, Wehle SMJ, et al. (2016) Why do people lie online? "Because everyone lies on the internet". *Computers in Human Behavior* 64: 134–142.

Dubé E and MacDonald NE (2020) How can a global pandemic affect vaccine hesitancy? *Expert Review of Vaccines* 19(10): 899–901.

Dubé E, Gagnon D and MacDonald NE (2015) Strategies intended to address vaccine hesitancy: Review of published reviews. *Vaccine* 33(34). WHO Recommendations Regarding Vaccine Hesitancy: 4191–4203.

Dudley R, Taylor P, Wickham S, et al. (2016) Psychosis, Delusions and the "Jumping to Conclusions" Reasoning Bias: A Systematic Review and Meta-analysis. *Schizophrenia Bulletin* 42(3): 652–665.

Dunbar RIM (2016) Do online social media cut through the constraints that limit the size of offline social networks? *Royal Society Open Science* 3(1). Royal Society: 150292.

Einhäuser W, Martin KAC and König P (2004) Are switches in perception of the Necker cube related to eye position? *European Journal of Neuroscience* 20(10): 2811–2818.

Eysenck HJ (1960) Personality and Behaviour Therapy. *Proceedings of the Royal Society of Medicine* 53(7). SAGE Publications: 504–508.

Farias J and Pilati R (2023) COVID-19 as an undesirable political issue: Conspiracy beliefs and intolerance of uncertainty predict adhesion to prevention measures. *Current Psychology (New Brunswick, N.J.)* 42(1): 209–219.

Federico CM and Malka A (2018) The contingent, contextual nature of the relationship between needs for security and certainty and political preferences: Evidence and implications. *Political Psychology* 39(Suppl 1). United Kingdom: Wiley-Blackwell Publishing Ltd.: 3–48.

Ferrara E (2020) Bots, Elections, and Social Media: A Brief Overview. In: Shu K, Wang S, Lee D, et al. (eds) *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Cham: Springer International Publishing, pp. 95–114. Available at: https://doi.org/10.1007/978-3-030-42699-6_6 (accessed 5 April 2022).

Finocchiaro MA (2005) *Retrying Galileo, 1633–1992*. University of California Press.

Fletcher PC and Frith CD (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews. Neuroscience* 10(1): 48–58.

Fletcher R, Cornia A and Nielsen RK (2020) How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries. *The International Journal of Press/Politics* 25(2). SAGE Publications Inc: 169–195.

Foster KR and Kokko H (2009) The evolution of superstitious and superstition-like behaviour. *Proceedings of the Royal Society B: Biological Sciences* 276(1654): 31–37.

Franceschi J and Pareschi L (2022) Spreading of fake news, competence and learning: kinetic modelling and numerical approximation. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 380(2224): 20210159.

Franks B, Bangerter A and Bauer MW (2013) Conspiracy theories as quasi-religious mentality: an integrated account from cognitive science, social representations theory, and frame theory. *Frontiers in Psychology* 4: 424.

Franks B, Bangerter A, Bauer MW, et al. (2017) Beyond 'Monologicality'? Exploring Conspiracist Worldviews. *Frontiers in Psychology* 8: 861.

Freeman D (2007) Suspicious minds: the psychology of persecutory delusions. *Clinical Psychology Review* 27(4): 425–457.

Freeman D and Bentall RP (2017) The concomitants of conspiracy concerns. *Social Psychiatry and Psychiatric Epidemiology* 52(5): 595–604.

Freeman D, Garety PA, Bebbington PE, et al. (2005) Psychological investigation of the structure of paranoia in a non-clinical population. *The British Journal of Psychiatry: The Journal of Mental Science* 186: 427–435.

Freeston MH, Rhéaume J, Letarte H, et al. (1994) Why do people worry? *Personality and Individual Differences* 17(6): 791–802.

Friston K (2005) A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360(1456): 815–836.

Fromm S, Katthagen T, Deserno L, et al. (2023) Belief Updating in Subclinical and Clinical Delusions. *Schizophrenia Bulletin Open* 4(1): sgac074.

Gagliardi L (2022) The role of cognitive biases in conspiracy beliefs: A literature review. 4261122, SSRN Scholarly Paper. Rochester, NY. Available at: https://papers.ssrn.com/abstract=4261122 (accessed 22 September 2023).

Galam S, Gefen (Feigenblat) Y and Shapir Y (1982) Sociophysics: A new approach of sociological collective behaviour. I. mean-behaviour description of a strike. *The Journal of Mathematical Sociology* 9(1): 1–13.

Garety PA, Hemsley DR and Wessely S (1991) Reasoning in deluded schizophrenic and paranoid patients. Biases in performance on a probabilistic inference task. *The Journal of Nervous and Mental Disease* 179(4): 194–201.

Garety PA, Freeman D, Jolley S, et al. (2005) Reasoning, emotions, and delusional conviction in psychosis. *Journal of Abnormal Psychology* 114(3): 373–384.

Geisler WS and Kersten D (2002) Illusions, perception and Bayes. *Nature Neuroscience* 5(6): 508–510.

Georgiou N, Delfabbro P and Balzan R (2020) COVID-19-related conspiracy beliefs and their relationship with perceived stress and pre-existing conspiracy beliefs. *Personality and Individual Differences* 166: 110201.

Georgiou N, Delfabbro P and Balzan R (2021) Conspiracy theory beliefs, scientific reasoning and the analytical thinking paradox. *Applied Cognitive Psychology* 35(6): 1523–1534.

Giachanou A, Zhang X, Barrón-Cedeño A, et al. (2022) Online information disorder: fake news, bots and trolls. *International Journal of Data Science and Analytics* 13(4): 265–269.

Gianotti LR, Mohr C, Pizzagalli D, et al. (2001) Associative processing and paranormal belief. *Psychiatry and Clinical Neurosciences* 55(6): 595–603.

Gigerenzer G (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press.

Gilbert P, Boxall M, Cheung M, et al. (2005) The relation of paranoid ideation and social anxiety in a mixed clinical population. *Clinical Psychology & Psychotherapy* 12(2): 124–133.

Gioia F, Imperato C, Boursier V, et al. (2023) The role of defense styles and psychopathological symptoms on adherence to conspiracy theories during the COVID-19 pandemic. *Scientific Reports* 13(1). 1. Nature Publishing Group: 3482.

Girvan M and Newman MEJ (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12): 7821–7826.

Gkinopoulos T and Mari S (2023) How exposure to real conspiracy theories motivates collective action and political engagement? The moderating role of primed victimhood and underlying emotional mechanisms in the case of 2018 bushfire in Attica. *Journal of Applied Social Psychology* 53(1): 21–38.

Goertzel T (1994) Belief in Conspiracy Theories. *Political Psychology* 15(4). [International Society of Political Psychology, Wiley]: 731–742.

Gosselin F and Schyns PG (2003) Superstitious perceptions reveal properties of internal representations. *Psychological Science* 14(5): 505–509.

Grant DA and Berg EA (1948) A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology* 38(4): 404–411.

Gray R, Franci A, Srivastava V, et al. (2018) Multi-agent decision-making dynamics inspired by honeybees. arXiv:1711.11578. arXiv. Available at: http://arxiv.org/abs/1711.11578 (accessed 7 October 2023).

Greenburgh A and Raihani NJ (2022) Paranoia and conspiracy thinking. *Current Opinion in Psychology* 47: 101362.

Grieve R and de Groot HT (2011) Does online psychological test administration facilitate faking? *Computers in Human Behavior* 27(6): 2386–2391.

Grzesiak-Feldman M (2013) The Effect of High-Anxiety Situations on Conspiracy Thinking. *Current Psychology* 32(1): 100–118.

Guloksuz S and van Os J (2018) The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. *Psychological Medicine* 48(2): 229–244.

Hacking I (1975) *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press.

Hart J and Graether M (2018) Something's going on here: Psychological predictors of belief in conspiracy theories. *Journal of Individual Differences* 39(4). Germany: Hogrefe Publishing: 229–237.

Hartmann M and Müller P (2023) Illusory perception of visual patterns in pure noise is associated with COVID-19 conspiracy beliefs. *i-Perception* 14(1): 204166952211447.

Haselgrove M, Le Pelley ME, Singh NK, et al. (2016) Disrupted attentional learning in high schizotypy: Evidence of aberrant salience. *British Journal of Psychology* 107(4): 601–624.

Helmholtz H von (1866) *Treatise on Physiological Optics, Volume III*. Courier Corporation.

Helmholtz H von (1948) Concerning the perceptions in general, 1867. In: *Readings in the History of Psychology*. Century psychology series. East Norwalk, CT, US: Appleton-Century-Crofts, pp. 214–230.

Heskes T (2002) Stable Fixed Points of Loopy Belief Propagation Are Local Minima of the Bethe Free Energy. In: *Advances in Neural Information Processing Systems*, 2002. MIT Press. Available at: https://proceedings.neurips.cc/paper_files/paper/2002/hash/d2a27e83d429f0dcae6b937cf440aeb1-Abstract.html (accessed 7 October 2023).

Heyes C (2012) New thinking: the evolution of human cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367(1599). Royal Society: 2091–2096.

Hogg MA, Adelman JR and Blagg RD (2010) Religion in the face of uncertainty: an uncertainty-identity theory account of religiousness. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* 14(1): 72–83.

Hollander E (1997) Obsessive-compulsive disorder: the hidden epidemic. *The Journal of Clinical Psychiatry* 58 Suppl 12: 3–6.

Holt DJ, Titone D, Long LS, et al. (2006) The misattribution of salience in delusional patients with schizophrenia. *Schizophrenia Research* 83(2): 247–256.

Horst S (2005) The Computational Theory of Mind. In: *Stanford Encyclopedia of Philosophy*.

Hugdahl K, Craven AR, Nygård M, et al. (2015) Glutamate as a mediating transmitter for auditory hallucinations in schizophrenia: a (1)H MRS study. *Schizophrenia Research* 161(2–3): 252–260.

Huq SF, Garety PA and Hemsley DR (1988) Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* 40(4-A). United Kingdom: Taylor & Francis: 801–812.

Ihler AT, Iii JWF and Willsky AS (2005) Loopy Belief Propagation: Convergence and Effects of Message Errors. *Journal of Machine Learning Research* 6(31): 905–936.

Imhoff R and Bruder M (2014) Speaking (Un–)Truth to Power: Conspiracy Mentality as A Generalised Political Attitude. *European Journal of Personality* 28(1). SAGE Publications Ltd: 25–43.

Imhoff R and Lamberty P (2018) How paranoid are conspiracy believers? Toward a more fine-grained understanding of the connect and disconnect between paranoia and belief in conspiracy theories. *European Journal of Social Psychology* 48(7): 909–926.

Imhoff R and Lamberty PK (2017) Too special to be duped: Need for uniqueness motivates conspiracy beliefs. *European Journal of Social Psychology* 47(6). US: John Wiley & Sons: 724–734.

Imhoff R, Dieterle L and Lamberty P (2021) Resolving the Puzzle of Conspiracy Worldview and Political Activism: Belief in Secret Plots Decreases Normative but Increases Nonnormative Political Engagement. *Social Psychological and Personality Science* 12(1). SAGE Publications Inc: 71–79.

Imhoff R, Zimmer F, Klein O, et al. (2022) Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour* 6(3): 392–403.

Intaitė M, Koivisto M, Rukšėnas O, et al. (2010) Reversal negativity and bistable stimuli: Attention, awareness, or something else? *Brain and Cognition* 74(1): 24–34.

Irwin HJ and Young JM (2002) Intuitive versus Reflective Processes in the Formation of Paranormal Beliefs. Koestler Chair of Parapsychology. Epub ahead of print 2002.

Irwin HJ, Schofield MB and Baker IS (2014) Dissociative tendencies, sensory-processing sensitivity and aberrant salience as predictors of anomalous experiences and paranormal attributions. *Journal of the Society for Psychical Research* 78. United Kingdom: Society for Psychical Research: 193–206.

Jardri R and Denève S (2013) Circular inferences in schizophrenia. *Brain: A Journal of Neurology* 136(Pt 11): 3227–3241.

Jardri R, Duverne S, Litvinova AS, et al. (2017) Experimental evidence for circular inference in schizophrenia. *Nature Communications* 8(1). 1. Nature Publishing Group: 14218.

Jolley D and Douglas KM (2014) The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS One* 9(2): e89177.

Kabengele M-C, Gollwitzer PM and Keller L (2023) Conspiracy Beliefs and Jumping to Conclusions. PsyArXiv. Available at: https://psyarxiv.com/63apz/ (accessed 25 July 2023).

Kalivas PW and Paulus MP (2021) *Intrusive Thinking: From Molecules to Free Will*. The MIT Press. Available at: https://direct.mit.edu/books/edited-volume/5267/Intrusive-ThinkingFrom-Molecules-to-Free-Will (accessed 11 October 2023).

Kapur S (2003) Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *The American Journal of Psychiatry* 160(1): 13–23.

Kapur S, Mizrahi R and Li M (2005) From dopamine to salience to psychosis--linking biology, pharmacology and phenomenology of psychosis. *Schizophrenia Research* 79(1): 59–68.

Karamched B, Stolarczyk S, Kilpatrick ZP, et al. (2020) Bayesian Evidence Accumulation on Social Networks. *SIAM Journal on Applied Dynamical Systems* 19(3). Society for Industrial and Applied Mathematics: 1884–1919.

Kasper J, Fiedler K, Kutzner F, et al. (2023) On the role of exploitation and exploration strategies in the maintenance of cognitive biases: Beyond the pursuit of instrumental rewards. *Memory & Cognition*. Epub ahead of print 24 January 2023. DOI: 10.3758/s13421-023-01393-8.

Kay AC, Gaucher D, Napier JL, et al. (2008) God and the government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology* 95(1). US: American Psychological Association: 18–35.

Kay AC, Gaucher D, McGregor I, et al. (2010) Religious belief as compensatory control. *Personality and Social Psychology Review* 14(1). US: Sage Publications: 37–48.

Kendler KS, Glazer WM and Morgenstern H (1983) Dimensions of delusional experience. *The American Journal of Psychiatry* 140(4): 466–469.

Kersten D, Mamassian P and Yuille A (2004) Object perception as Bayesian inference. *Annual Review of Psychology* 55: 271–304.

Klein C, Clutton P and Polito V (2018) Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology* 9. Switzerland: Frontiers Media S.A.

Kofta M, Soral W and Bilewicz M (2020) What breeds conspiracy antisemitism? The role of political uncontrollability and uncertainty in the belief in Jewish conspiracy. *Journal of Personality and Social Psychology* 118(5): 900–918.

Koller D and Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. Cambridge, MA: MIT Press.

Krummenacher P, Mohr C, Haker H, et al. (2010) Dopamine, paranormal belief, and the detection of meaningful stimuli. *Journal of Cognitive Neuroscience* 22(8): 1670–1681.

Kuhn SAK, Lieb R, Freeman D, et al. (2022) Coronavirus conspiracy beliefs in the German-speaking general population: endorsement rates and links to reasoning biases and paranoia. *Psychological Medicine* 52(16). Cambridge University Press: 4162–4176.

Lachman ME and Weaver SL (1998a) Sociodemographic variations in the sense of control by domain: findings from the MacArthur studies of midlife. *Psychology and Aging* 13(4): 553–562.

Lachman ME and Weaver SL (1998b) The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology* 74(3). US: American Psychological Association: 763–773.

Lantian A, Muller D, Nurra C, et al. (2016) Measuring Belief in Conspiracy Theories: Validation of a French and English Single-Item Scale. *International Review of Social Psychology* 29(1): 1.

Lantian A, Muller D, Nurra C, et al. (2017) "I Know Things They Don't Know!" *Social Psychology* 48(3). Hogrefe Publishing: 160–173.

Lantian A, Bagneux V, Delouvée S, et al. (2021) Maybe a free thinker but not a critical one: High conspiracy belief is associated with low critical thinking ability. *Applied Cognitive Psychology* 35(3): 674–684.

Laurin K, Kay AC and Moscovitch DA (2008) On the belief in God: Towards an understanding of the emotional substrates of compensatory control. *Journal of Experimental Social Psychology* 44(6). Netherlands: Elsevier Science: 1559–1562.

Lee FLF (2016) Impact of social media on opinion polarization in varying times. *Communication and the Public* 1(1): 56–71.

Leeser J and O'Donohue W (1999) What is a delusion? Epistemological dimensions. *Journal of Abnormal Psychology* 108(4). US: American Psychological Association: 687–694.

Leptourgos P, Denève S and Jardri R (2017) Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Current Opinion in Neurobiology* 46. Netherlands: Elsevier Science: 154–161.

Leptourgos P, Bouttier V, Jardri R, et al. (2020) A functional theory of bistable perception based on dynamical circular inference. *PLOS Computational Biology* 16(12). Public Library of Science: e1008480.

Leptourgos P, Notredame C-E, Eck M, et al. (2020) Circular inference in bistable perception. *Journal of Vision* 20(4). The Association for Research in Vision and Ophthalmology: 12–12.

Lewandowsky S, Gignac GE and Oberauer K (2013) The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLOS ONE* 8(10). Public Library of Science: e75637.

Liekefett L, Christ O and Becker JC (2021) Can Conspiracy Beliefs Be Beneficial? Longitudinal Linkages Between Conspiracy Beliefs, Anxiety, Uncertainty Aversion, and Existential Threat. *Personality and Social Psychology Bulletin*. SAGE Publications Inc: 01461672211060965.

Liu S, Zhang L and Yan Z (2018) Predict Pairwise Trust Based on Machine Learning in Online Social Networks: A Survey. *IEEE Access* 6: 51297–51318.

Lobato E, Mendoza J, Sims V, et al. (2014) Examining the Relationship Between Conspiracy Theories, Paranormal Beliefs, and Pseudoscience Acceptance Among a University Population. *Applied Cognitive Psychology* 28(5): 617–625.

Lockwood PL, Apps MAJ and Chang SWC (2020) Is There a 'Social' Brain? Implementations and Algorithms. *Trends in Cognitive Sciences* 24(10): 802–813.

Long GM and Toppino TC (1981) Multiple representations of the same reversible figure: Implications for cognitive decisional interpretations. *Perception* 10(2). United Kingdom: Pion: 231–234.

Love S (2022) Inside the Dispute Over a High-Profile Psychedelic Study. In: *Vice*. Available at: https://www.vice.com/en/article/4awj3n/inside-the-dispute-over-a-high-profile-psychedelic-study (accessed 7 October 2023).

Madsen JK, Bailey R and Pilditch TD (2017) Growing a Bayesian Conspiracy Theorist: An Agent-Based Model. London, UK: Cognitive Science Society. Available at: https://mindmodeling.org/cogsci2017/papers/0503/index.html (accessed 25 July 2023).

Madsen JK, Bailey RM and Pilditch TD (2018) Large networks of rational agents form persistent echo chambers. *Scientific Reports* 8(1). 1. Nature Publishing Group: 12391.

Mamassian P and Goutcher R (2005) Temporal dynamics in bistable perception. *Journal of Vision* 5(4): 7.

Mamassian P, Landy M and Maloney LT (2002) Bayesian modelling of visual perception. In: *Probabilistic Models of the Brain: Perception and Neural Function*. Neural information processing series. Cambridge, MA, US: The MIT Press, pp. 13–36.

March E and Springer J (2019) Belief in conspiracy theories: The predictive role of schizotypy, Machiavellianism, and primary psychopathy. *PLOS ONE* 14(12). Public Library of Science: e0225964.

Marchlewska M, Green R, Cichocka A, et al. (2021) From bad to worse: Avoidance coping with stress increases conspiracy beliefs. *British Journal of Social Psychology*. Epub ahead of print 31 August 2021. DOI: 10.1111/bjso.12494.

Marinthe G, Brown G, Delouvée S, et al. (2020) Looking out for myself: Exploring the relationship between conspiracy mentality, perceived personal risk, and COVID-19 prevention measures. *British Journal of Health Psychology* 25(4): 957–980.

Marr D (1982) Vision : a computational investigation into the human representation and processing of visual information | BibSonomy. Available at: https://www.bibsonomy.org/bibtex/31060780234ce2036de55d261cc63a61 (accessed 5 October 2023).

Masumi A and Sato T (2021) Model-based analysis of learning latent structures in probabilistic reversal learning task. *Artificial Life and Robotics* 26(3): 275–282.

Mattson MP (2014) Superior pattern processing is the essence of the evolved human brain. *Frontiers in Neuroscience* 8.

McAuley J and Leskovec J (2012) Learning to discover social circles in ego networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, Red Hook, NY, USA, 2012, pp. 539–547. NIPS'12. Curran Associates Inc.

McBain R, Norton DJ, Kim J, et al. (2011) Reduced Cognitive Control of a Visually Bistable Image in Schizophrenia. *Journal of the International Neuropsychological Society* 17(3). Cambridge University Press: 551–556.

McHoskey JW (1995) Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology* 17(3). US: Lawrence Erlbaum: 395–409.

McLean BF, Mattiske JK and Balzan RP (2017) Association of the Jumping to Conclusions and Evidence Integration Biases With Delusions in Psychosis: A Detailed Meta-analysis. *Schizophrenia Bulletin* 43(2): 344–354.

McNeish DM and Stapleton LM (2016) The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review* 28(2). Germany: Springer: 295–314.

Meehl PE (1962) Schizotaxia, schizotypy, schizophrenia. *American Psychologist* 17(12). US: American Psychological Association: 827–838.

Meller T (2002) "Agency panic and the culture of conspiracy," in Conspiracy Nation: The Politics of Paranoia in Post-War America. *New York University Press*. Knight P. New York.

Mengin A, Allé MC, Rolling J, et al. (2020) [Psychopathological consequences of confinement]. *L'Encephale* 46(3S): S43–S52.

Mittal C and Griskevicius V (2014) Sense of control under uncertainty depends on people's childhood environment: A life history theory approach. *Journal of Personality and Social Psychology* 107(4). US: American Psychological Association: 621–637.

Molenda Z, Green R, Marchlewska M, et al. (2023) Emotion dysregulation and belief in conspiracy theories. *Personality and Individual Differences* 204: 112042.

Mooij JM and Kappen HJ (2007) Sufficient Conditions for Convergence of the Sum–Product Algorithm. *IEEE Transactions on Information Theory* 53(12): 4422–4437.

Müller P and Hartmann M (2023) Linking paranormal and conspiracy beliefs to illusory pattern perception through signal detection theory. *Scientific Reports* 13(1): 9739.

Nera K, Jetten J, Biddlestone M, et al. (2022) 'Who wants to silence us'? Perceived discrimination of conspiracy theory believers increases 'conspiracy theorist' identification when it comes from powerholders – But not from the general public. *British Journal of Social Psychology* 61(4): 1263–1285.

Newman MEJ and Park J (2003) Why social networks are different from other types of networks. *Physical Review E* 68(3). American Physical Society: 036122.

Noordewier MK and Rutjens BT (2021) Personal need for structure shapes the perceived impact of reduced personal control. *Personality and Individual Differences* 170: 110478.

Notredame C-E, Pins D, Deneve S, et al. (2014) What visual illusions teach us about schizophrenia. *Frontiers in Integrative Neuroscience* 8.

Nyhan B and Zeitzoff T (2018) Conspiracy and Misperception Belief in the Middle East and North Africa. *The Journal of Politics* 80(4). The University of Chicago Press: 1400–1404.

Oaksford M and Chater N (2001) The probabilistic approach to human reasoning. *Trends in Cognitive Sciences* 5(8). Netherlands: Elsevier Science: 349–357.

Oleksy T, Wnuk A, Maison D, et al. (2021) Content matters. Different predictors and social consequences of general and government-related conspiracy theories on COVID-19. *Personality and Individual Differences* 168: 110289.

O'Mahony C, Brassil M, Murphy G, et al. (2023) The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE* 18(4). Public Library of Science: e0280902.

Pahuja E, Manjunatha N, Kumar CN, et al. (2020) Repetitive superficial self harm as an acting out on delusion of persecution: A case report and mini review. *Asian Journal of Psychiatry* 48: 101904.

Palgi Y, Shrira A, Ring L, et al. (2020) The loneliness pandemic: Loneliness and other concomitants of depression, anxiety and their comorbidity during the COVID-19 outbreak. *Journal of Affective Disorders* 275: 109–111.

Pantazi M, Papaioannou K and van Prooijen J-W (2022) Power to the People: The Hidden Link Between Support for Direct Democracy and Belief in Conspiracy Theories. *Political Psychology* 43(3): 529–548.

Pariser E (2011) *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press. New York.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Peluso AM and Pichierri M (2021) Effects of socio-demographics, sense of control, and uncertainty avoidability on post-COVID-19 vacation intention. *Current Issues in Tourism* 24(19). Routledge: 2755–2767.

Peralta AF, Neri M, Kertész J, et al. (2021) Effect of algorithmic bias and network structure on coexistence, consensus, and polarization of opinions. *Physical Review E* 104(4). American Physical Society: 044312.

Pertwee E, Simas C and Larson HJ (2022) An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nature Medicine* 28(3). 3. Nature Publishing Group: 456–459.

Peruzzi A, Zollo F, Schmidt AL, et al. (2019) From confirmation bias to echo-chambers: a data-driven approach. *SOCIOLOGIA E POLITICHE SOCIALI* (2018/3). FrancoAngeli Editore. Epub ahead of print 2019. DOI: 10.3280/SP2018-003004.

Peters E, Ward T, Jackson M, et al. (2016) Clinical, socio-demographic and psychological characteristics in individuals with persistent psychotic experiences with and without a 'need for care'. *World psychiatry: official journal of the World Psychiatric Association (WPA)* 15(1): 41–52.

Phillips LD and Edwards W (1966) Conservatism in a simple probability inference task. *Journal of Experimental Psychology* 72(3). US: American Psychological Association: 346–354.

Pierre JM (2020) Mistrust and Misinformation: A Two-Component, Socio-Epistemic Model of Belief in Conspiracy Theories. *Journal of Social and Political Psychology* 8(2). 2: 617–641.

Pilditch TD, Roozenbeek J, Madsen JK, et al. (2022) Psychological inoculation can reduce susceptibility to misinformation in large rational agent networks. *Royal Society Open Science* 9(8). Royal Society: 211953.

Pilowsky LS, Bressan RA, Stone JM, et al. (2006) First in vivo evidence of an NMDA receptor deficit in medication-free schizophrenic patients. *Molecular Psychiatry* 11(2): 118–119.

Polgári P, Causin J-B, Weiner L, et al. (2020) Novel method to measure temporal windows based on eye movements during viewing of the Necker cube. *PLOS ONE* 15(1). Public Library of Science: e0227506.

Polgári P, Weiner L, Causin J-B, et al. (2023) Investigating racing thoughts via ocular temporal windows: deficits in the control of automatic perceptual processes. *Psychological Medicine* 53(4). Cambridge University Press: 1176–1184.

Poth N and Dolega K (2023) Bayesian belief protection: A study of belief in conspiracy theories. *Philosophical Psychology* 36(6). Routledge: 1182–1207.

Pretti M (2005) A message-passing algorithm with damping. *Journal of Statistical Mechanics: Theory and Experiment* 2005(11): P11008.

Prooijen J-W van and Lange PAM van (2014) *Power, Politics, and Paranoia: Why People Are Suspicious of Their Leaders*. Cambridge University Press.

Proskurnikov AV, Matveev AS and Cao M (2016) Opinion Dynamics in Social Networks With Hostile Camps: Consensus vs. Polarization. *IEEE Transactions on Automatic Control* 61(6): 1524–1536.

Pytlik N, Soll D and Mehl S (2020) Thinking Preferences and Conspiracy Belief: Intuitive Thinking and the Jumping to Conclusions-Bias as a Basis for the Belief in Conspiracy Theories. *Frontiers in Psychiatry* 11: 568942.

Rado S (1953) Dynamics and classification of disordered behavior. *The American Journal of Psychiatry* 110(6): 406–416.

Raihani NJ and Bell V (2019) An evolutionary perspective on paranoia. *Nature Human Behaviour* 3(2): 114–121.

Reed EJ, Uddenberg S, Suthaharan P, et al. (2020) Paranoia as a deficit in non-social belief updating. *eLife* Schoenbaum G, de Lange FP, and Schoenbaum G (eds) 9. eLife Sciences Publications, Ltd: e56345.

Reips U-D (2000) The Web experiment method: Advantages, disadvantages, and solutions. In: *Psychological Experiments on the Internet*. San Diego, CA, US: Academic Press, pp. 89–117.

Reips U-D (2002a) Internet-Based Psychological Experimenting: Five Dos and Five Don'ts. *Social Science Computer Review* 20(3). SAGE Publications Inc: 241–249.

Reips U-D (2002b) Standards for Internet-based experimenting. *Experimental Psychology* 49(4). Germany: Hogrefe & Huber Publishers: 243–256.

Reynolds JH and Desimone R (2003) Interacting Roles of Attention and Visual Salience in V4. *Neuron* 37(5). Elsevier: 853–863.

Riekki T, Lindeman M, Aleneff M, et al. (2013) Paranormal and religious believers are more prone to illusory face perception than skeptics and non-believers. *Applied Cognitive Psychology* 27(2). US: John Wiley & Sons: 150–155.

Rieth CA, Lee K, Lui J, et al. (2011) Faces in the mist: Illusory face and letter detection. *i-Perception* 2(5). United Kingdom: Pion.

Rigoli F (2022) Deconstructing the Conspiratorial Mind: the Computational Logic Behind Conspiracy Theories. *Review of Philosophy and Psychology*: 1–18.

Riva G, Teruzzi T and Anolli L (2003) The Use of the Internet in Psychological Research: Comparison of Online and Offline Questionnaires. *CyberPsychology & Behavior* 6(1). Mary Ann Liebert, Inc., publishers: 73–80.

Rodriguez MG, Gummadi K and Schoelkopf B (2014) Quantifying Information Overload in Social Media and its Impact on Social Contagions. arXiv:1403.6838. arXiv. Available at: http://arxiv.org/abs/1403.6838 (accessed 7 October 2023).

Rodríguez-Martínez G (2023) Perceptual reversals and creativity: is it possible to develop divergent thinking by modulating bistable perception? *Revista de Investigación Desarrollo e Innovación* 13: 129–144.

Rodríguez-Martínez GA and Castillo-Parra H (2018) Bistable perception: neural bases and usefulness in psychological research. *International Journal of Psychological Research* 11(2): 63–76.

Romer D and Jamieson KH (2020) Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Social Science & Medicine (1982)* 263: 113356.

Rossell SL, Labuschagne I, Castle DJ, et al. (2020) Delusional themes in Body Dysmorphic Disorder (BDD): Comparisons with psychotic disorders and non-clinical Controls. *Psychiatry Research* 284: 112694.

Rowling JK (1997) *Harry Potter. The Philosopher's Stone; The Chamber of Secrets; The Prisoner of Azkaban; The Goblet of Fire; The Order of the Phoenix; The Half-Blood Prince; The Deathly Hallows*. First edition. Rowling, J. K. Harry Potter series ; Arthur A. Levine Books,.

Santos FP, Lelkes Y and Levin SA (2021) Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences of the United States of America* 118(50): e2102141118.

Sasahara K, Chen W, Peng H, et al. (2021) Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science* 4(1): 381–402.

Sato R (2023) The rabbit-hole of conspiracy theories: An analysis from the perspective of the free energy principle. *Philosophical Psychology* 36(6). Routledge: 1160–1181.

Sauvé G, Lavigne KM, Pochiet G, et al. (2020) Efficacy of psychological interventions targeting cognitive biases in schizophrenia: A systematic review and meta-analysis. *Clinical Psychology Review* 78: 101854.

Shapley R and Hawken M (2002) Neural mechanisms for color perception in the primary visual cortex. *Current Opinion in Neurobiology* 12(4): 426–432.

Sheffield JM, Suthaharan P, Leptourgos P, et al. (2022) Belief Updating and Paranoia in Individuals With Schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 7(11): 1149–1157.

Shermer M (2011) *The Believing Brain: From Ghosts and Gods to Politics and Conspiracies---How We Construct Beliefs and Reinforce Them as Truths*. Macmillan.

Shrout PE and Bolger N (2002) Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* 7(4). US: American Psychological Association: 422–445.

Simmons A, McGatlin K and Lustig C (2023) How well do online, self-administered measures correspond to in-lab assessments? A preliminary examination of three measures in healthy older adults. *Neuropsychology* 37(4). US: American Psychological Association: 424–435.

Simonsen A, Fusaroli R, Petersen ML, et al. (2021) Taking others into account: combining directly experienced and indirect information in schizophrenia. *Brain: A Journal of Neurology* 144(5): 1603–1614.

Šrol J, Ballová Mikušková E and Čavojová V (2021) When we are worried, what are we thinking? Anxiety, lack of control, and conspiracy beliefs amidst the COVID-19 pandemic. *Applied Cognitive Psychology* 35(3): 720–729.

Stein A (2021) *Terror, Love and Brainwashing: Attachment in Cults and Totalitarian Systems*. Routledge.

Sterzer P, Kleinschmidt A and Rees G (2009) The neural bases of multistable perception. *Trends in Cognitive Sciences* 13(7): 310–318.

Stieger S, Gumhalter N, Tran U, et al. (2013) Girl in the cellar: a repeated cross-sectional investigation of belief in conspiracy theories about the kidnapping of Natascha Kampusch. *Frontiers in Psychology* 4.

Stojanov A, Bering JM and Halberstadt J (2020) Does Perceived Lack of Control Lead to Conspiracy Theory Beliefs? Findings from an online MTurk sample. *PLoS ONE* 15(8): e0237771.

Stojanov A, Bering JM and Halberstadt J (2022) Perceived lack of control and conspiracy theory beliefs in the wake of political strife and natural disaster. *Psihologija* 55(2): 149–168.

Stojanov A, Halberstadt J, Bering JM, et al. (2023) Examining a domain-specific link between perceived control and conspiracy beliefs: a brief report in the context of COVID-19. *Current Psychology* 42(8): 6347–6356.

Sullivan D, Landau MJ and Rothschild ZK (2010) An existential function of enemyship: Evidence that people attribute influence to personal and political enemies to compensate for threats to control. *Journal of Personality and Social Psychology* 98(3). US: American Psychological Association: 434–449.

Suthaharan P and Corlett PR (2023) Assumed shared belief about conspiracy theories in social networks protects paranoid individuals against distress. *Scientific Reports* 13(1): 6084.

Suthaharan P, Reed EJ, Leptourgos P, et al. (2021) Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*: 1–13.

Swami V, Chamorro-Premuzic T and Furnham A (2010) Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology* 24(6): 749–761.

Swami V, Coles R, Stieger S, et al. (2011) Conspiracist ideation in Britain and Austria: evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology (London, England: 1953)* 102(3): 443–463.

Swami V, Pietschnig J, Tran US, et al. (2013) Lunar Lies: The Impact of Informational Framing and Individual Differences in Shaping Conspiracist Beliefs About the Moon Landings. *Applied Cognitive Psychology* 27(1): 71–80.

Swami V, Voracek M, Stieger S, et al. (2014) Analytic thinking reduces belief in conspiracy theories. *Cognition* 133(3): 572–585.

Swami V, Furnham A, Smyth N, et al. (2016) Putting the stress on conspiracy theories: Examining associations between psychological stress, anxiety, and belief in conspiracy theories. *Personality and Individual Differences* 99: 72–76.

Tenenbaum JB, Griffiths TL and Kemp C (2006) Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences* 10(7): 309–318.

Thompson SC, Sobolew-Shubin A, Galbraith ME, et al. (1993) Maintaining perceptions of control: Finding perceived control in low-control circumstances. *Journal of Personality and Social Psychology* 64(2). US: American Psychological Association: 293–304.

Tokita CK, Guess AM and Tarnita CE (2021) Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences* 118(50). Proceedings of the National Academy of Sciences: e2102147118.

Tolkien JRR (1954) *The Lord of the Rings. The Fellowship of the Ring; The Two Towers; The Return of the King*. London: George Allen and Unwin.

Tollefson J (2021) Tracking QAnon: how Trump turned conspiracy-theory research upside down. *Nature* 590(7845): 192–193.

UNESCO Institute for Statistics and Statistics UI for (2020) International Standard Classification of Education (ISCED). Available at: http://uis.unesco.org/en/topic/international-standard-classification-education-isced (accessed 25 July 2023).

Uscinski JE and Olivella S (2017) The conditional effect of conspiracy thinking on attitudes toward climate change. *Research & Politics* 4(4). SAGE Publications Ltd: 2053168017743105.

van Elk M and Lodder P (2018) Experimental Manipulations of Personal Control do Not Increase Illusory Pattern Perception. *Collabra: Psychology* Vazire S and Tullett A (eds) 4(1): 19.

van Harreveld F, Rutjens BT, Schneider IK, et al. (2014) In doubt and disorderly: Ambivalence promotes compensatory perceptions of order. *Journal of Experimental Psychology. General* 143(4): 1666–1676.

van Mulukom V, Pummerer LJ, Alper S, et al. (2022) Antecedents and consequences of COVID-19 conspiracy beliefs: A systematic review. *Social Science & Medicine (1982)* 301: 114912.

van Prooijen J, Douglas KM and De Inocencio C (2018) Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural. *European Journal of Social Psychology* 48(3): 320–335.

van Prooijen J-W (2017) Why Education Predicts Decreased Belief in Conspiracy Theories. *Applied Cognitive Psychology* 31(1): 50–58.

van Prooijen J-W and Acker M (2015) The Influence of Control on Belief in Conspiracy Theories: Conceptual and Applied Extensions. *Applied Cognitive Psychology* 29(5): 753–761.

van Prooijen J-W and Douglas KM (2017) Conspiracy theories as part of history: The role of societal crisis situations. *Memory Studies* 10(3). SAGE Publications: 323–333.

van Prooijen J-W and Jostmann NB (2013) Belief in conspiracy theories: The influence of uncertainty and perceived morality. *European Journal of Social Psychology* 43(1): 109–115.

van Prooijen J-W and van Vugt M (2018) Conspiracy Theories: Evolved Functions and Psychological Mechanisms. *Perspectives on Psychological Science* 13(6). SAGE Publications Inc: 770–788.

Vespignani A (2018) Twenty years of network science. *Nature* 558(7711): 528–529.

Volk DW and Lewis DA (2002) Impaired prefrontal inhibition in schizophrenia: relevance for cognitive dysfunction. *Physiology & Behavior* 77(4): 501–505.

Volkan K, Department of Psychology CSUCI and Graduate Medical Education Program CMHS (2021) Schizophrenia, Culture, and Culture-Bound Syndromes. *Psychology Research and Applications* 3(1).

Voon V, Chang-Webb YC, Morris LS, et al. (2016) Waiting Impulsivity: The Influence of Acute Methylphenidate and Feedback. *International Journal of Neuropsychopharmacology* 19(1): pyv074.

Wagner-Egger P and Bangerter A (2007) The Truth Lies Elsewhere: Correlates of Belief in Conspiracy Theories. *Revue internationale de psychologie sociale* 20(4): 31–61.

Wagner-Egger P, Bangerter A, Delouvée S, et al. (2022) Awake together: Sociopsychological processes of engagement in conspiracist communities. *Current Opinion in Psychology* 47: 101417.

Wainwright MJ and Jordan MI (2008) Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning* 1(1–2). Now Publishers, Inc.: 1–305.

Wainwright MJ, Jaakkola TS and Willsky AS (2005) A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory* 51(7): 2313–2335.

Walker AC, Turpin MH, Stolz JA, et al. (2019) Finding meaning in the clouds: Illusory pattern perception predicts receptivity to pseudo-profound bullshit. *Judgment and Decision Making* 14(2).

Wheeler EA (2021) How Belief in Conspiracy Theories Addresses Some Basic Human Needs. In: Sinnott JD and Rabin JS (eds) *The Psychology of Political Behavior in a Time of Change*. Identity in a Changing World. Cham: Springer International Publishing, pp. 263–276. Available at: https://doi.org/10.1007/978-3-030-38270-4_11 (accessed 10 October 2023).

Whitson JA and Galinsky AD (2008) Lacking Control Increases Illusory Pattern Perception. *Science* 322(5898). American Association for the Advancement of Science: 115–117.

Wiegerinck W and Heskes T (2002) Fractional Belief Propagation. In: *Advances in Neural Information Processing Systems*, 2002. MIT Press. Available at: https://proceedings.neurips.cc/paper_files/paper/2002/hash/35936504a37d53e03abdfbc7318d9ec7-Abstract.html (accessed 7 October 2023).

Winn J and Bishop CM (2005) Variational Message Passing. *Journal of Machine Learning Research* 6(23): 661–694.

Wiseman R and Watt C (2006) Belief in psychic ability and the misattribution hypothesis: a qualitative review. *British Journal of Psychology (London, England: 1953)* 97(Pt 3): 323–338.

Wood MJ, Douglas KM and Sutton RM (2012) Dead and Alive: Beliefs in Contradictory Conspiracy Theories. *Social Psychological and Personality Science* 3(6). SAGE Publications Inc: 767–773.

Woodward TS, Buchy L, Moritz S, et al. (2007) A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophrenia Bulletin* 33(4). United Kingdom: Oxford University Press: 1023–1028.

Wycha N (2015) It's a Conspiracy: Motivated Reasoning and Conspiracy Ideation in the Rejection of Climate Change. *Electronic Theses and Dissertations*. Epub ahead of print 1 January 2015.

Yedidia JS, Freeman WT and Weiss Y (2005) Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* 51(7): 2282–2312.

Yuille A and Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences* 10(7). Special issue: Probabilistic models of cognition: 301–308.

Zhang lili, Ward T, Vashisht H, et al. (2022) Belief in Conspiracy Theories is associated with decreased adaptive learning to contingency volatilit. Available at: https://doi.org/10.31234/osf.io/nrqb8 (accessed 1 October 2023).

Zhu N, O J, Lu HJ, et al. (2020) Debate: Facing uncertainty with(out) a sense of control – cultural influence on adolescents' response to the COVID-19 pandemic. *Child and Adolescent Mental Health* 25(3): 173–174.

Zmigrod L and Tsakiris M (2021) Computational and neurocognitive approaches to the political brain: key insights and future avenues for political neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376(1822). Royal Society: 20200130.

UNRAVELLING THE FABRIC OF CONSPIRACY THEORIES FROM THE INDIVIDUAL TO THE COMMUNITY
Cognitive and Computational Dimensions

Abstract : The pace of socio-political crisis has dramatically accelerated over the past decade and was accompanied by a surge in conspiratorial and pseudo-scientific beliefs. These inflexible ideas, widespread to varying degrees in the non-clinical population, can drastically influence a wide range of attitudes from health-related behaviors to political engagement. Furthermore, while it can be observed at an individual level, this global phenomenon of belief rigidification is mirrored at the community level by the major polarization and radicalization of online opinions. Adherence to conspiracy theories (CTs) has been proposed by some authors as a coping strategy aimed at restoring predictability in highly uncertain situations. This compensatory mechanism would be rooted in cognitive and perceptual inference biases that can be captured by Bayesian belief models. This PhD thesis aimed at deciphering the mechanisms underpinning the emergence and maintenance of unshakeable conspiracy beliefs through three intertwined levels of comprehension: the cognitive, perceptual and computational approaches. In the first axis, I explored the cognitive mechanisms associated with conspiracy ideations. I notably showed that hypersalience, a cognitive bias that consists in attributing great significance to irrelevant stimuli, was associated with adherence to CTs and vaccine hesitancy in the context of the COVID-19 pandemic. We also demonstrated that the debated link between the perceived lack of control over one's life could be experimentally captured with a behavioral task. I further showed that this association was stress sensitive and could be uncovered by real-world uncertainty. Drawing on that idea, the second axis aimed at deciphering the dynamics of CTs around distressing and uncertain political events, combining an online bistable perception task with computational model fitting. Using the *Circular Inference* framework, we notably showed that when uncertainty peaks, CTs were associated with an overweighting of sensory information. In an attempt to cope with uncertainty, some participants particularly sensitive to stress adopted an exploration strategy that consisted in searching for simple and intuitive answers to complex issues. Progressively, this exploration strategy could shift to an exploitation strategy in which increased adherence to CTs is associated with the amplification of prior information leading to a self-reinforcement of the belief system. Finally, in the third axis, I addressed the question of belief rigidification at the community scale by modeling belief propagation in large social networks, as a form of probabilistic inference. We notably approached the phenomenon of polarization and radicalization observed in online communities as aberrant overconfidence rooted in some form of circularity in messages-passing, considered inherent to the network's structure. Going further, we demonstrated the validity of a novel algorithm, *Circular Belief Propagation*, in countering this aberrant overconfidence using data from Facebook© and Twitter©.

Keywords : Conspiracy; belief; bistability; circular inference; Bayesian inference

COMPRENDRE LES IDEATIONS COMPLOTISTES DE L'INDIVIDU AU GROUPE SOCIAL
Dimensions cognitives et computationnelles

Résumé : Le rythme des crises sociopolitiques s'est considérablement accéléré au cours des dernières années, et s'est accompagné d'une montée en flèche des croyances conspirationnistes et pseudo-scientifiques. Ces idées inflexibles, largement répandues au sein de la population générale, ont eu un impact non négligeable sur les comportements individuels, qu'il s'agisse de leur choix de santé ou de leur choix d'engagement politique. Ce phénomène global de rigidification des croyances, s'il peut être observé au niveau individuel, trouve également son reflet à l'échelle des groupes sociaux, via la polarisation et la radicalisation des opinions en ligne. Certains auteurs ont formulé l'hypothèse du « complotisme » en tant que stratégie d'adaptation, qui viserait à rétablir la prévisibilité du monde face à des évènements très incertains. Ce mécanisme compensatoire s'appuierait sur des biais d'inférence, cognitifs et perceptifs, que les modèles Bayésiens sont censés pourvoir capturer. Cette thèse de doctorat visait donc à décrypter les mécanismes qui sous-tendent l'émergence et le maintien des croyances complotistes à travers trois niveaux de compréhension : les approches cognitives, perceptives et computationnelles. Dans le premier axe, j'ai exploré les mécanismes cognitifs associés aux idéations complotistes (IC). J'y montre notamment que l'attribution aberrante de saillance, un biais cognitif qui consiste à attribuer une trop grande importance à des stimuli non-pertinents, est associée aux IC et à l'hésitation vaccinale lors de la pandémie de COVID-19. Nous démontrons également que le lien, encore débattu, entre IC et perte du sentiment de contrôle sur le monde, peut être capturé expérimentalement par une tâche comportementale. Je montre enfin que cette association est sensible au stress et peut être révélée par l'incertitude du monde réel. En m'appuyant sur ces résultats, je me suis appliqué dans le deuxième axe du travail à déchiffrer la dynamique des IC autour d'événements politiques incertains, en combinant l'utilisation d'une tâche de perception bistable en ligne et d'un modèle computationnel. En utilisant le modèle de l'*Inférence Circulaire*, nous montrons notamment que lorsque l'incertitude atteint son paroxysme, les IC sont associées à une prise en compte plus importante des informations sensorielles. Pour faire face à l'incertitude, certains participants, particulièrement sensibles au stress, adopteraient une stratégie d'exploration consistant à rechercher des réponses simples et intuitives à des questions complexes. Progressivement, cette stratégie d'exploration évoluerait vers une stratégie d'exploitation dans laquelle l'adhésion accrue aux théories du complot est associée à l'amplification des connaissances a priori conduisant à un auto-renforcement du système de croyance. Enfin, dans le troisième axe de la thèse, j'aborde la question de la rigidification des croyances à l'échelle du groupe, en modélisant la propagation des croyances dans les réseaux sociaux comme une forme d'inférence probabiliste. Nous avons notamment abordé les phénomènes de polarisation et de radicalisation communément observés dans les communautés en ligne comme un excès de confiance qui trouverait ses origines dans une forme de circularité inhérente à la structure du réseau. En outre, nous avons pu démontrer la validité d'un nouvel algorithme, le *Circular Belief Propagation* (CBP), capable de contrer cet excès de confiance en utilisant des données issues de réseaux réels, tels que Facebook© et Twitter©.

Mots-clef : Complotisme; croyance; bistabilité; inférence circulaire; inférence Bayésienne.