



**HAL**  
open science

# Neurocognitive mechanisms underlying transmission of uncertain information in humans

Valentin Guigon

► **To cite this version:**

Valentin Guigon. Neurocognitive mechanisms underlying transmission of uncertain information in humans. Neuroscience. Université Claude Bernard - Lyon I, 2022. English. NNT : 2022LYO10183 . tel-04817845

**HAL Id: tel-04817845**

**<https://theses.hal.science/tel-04817845v1>**

Submitted on 4 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**THESE de DOCTORAT DE  
L'UNIVERSITE CLAUDE BERNARD LYON 1**

**Ecole Doctorale N° 476 - NSCo  
Neurosciences et Cognition**

**Discipline :**  
Neurosciences

Soutenue publiquement le 02/12/2022, par :  
**Valentin Guigon**

---

**Neurocognitive mechanisms underlying  
transmission of uncertain information  
in humans**

/

**Mécanismes neurocognitifs sous-  
jacents à la transmission d'informations  
incertaines chez l'humain**

---

Devant le jury composé de :

Derrington, Edmund, P.U., ISCMJ - UCBL1  
Borst, Grégoire, P.U. LaPsyDé - Université Paris-Cité  
Schwieren, Christiane, Professeure, AWI - Université d'Heidelberg  
Charpentier, Caroline, Chercheure, HSS - Caltech  
Palminteri, Stefano, D.R., LNC2 - INSERM  
Dreher, Jean-Claude, D.R., ISCMJ - CNRS  
Villeval, Marie Claire, D.R., GATE - CNRS

Président  
Rapporteur  
Rapporteuse  
Examinatrice  
Examineur  
Directeur de thèse  
Co-directrice de thèse

# Table of contents

Remerciements / Acknowledgments .....	5
Abstract .....	6
Résumé .....	8
General introduction.....	10
I. Modern news consumption.....	10
II. Beliefs about the state of the world .....	13
III. Motivated beliefs shape information selection .....	16
IV. The value of information .....	18
V. Metacognitive abilities .....	20
VI. Evaluating information .....	22
VII. Sharing news .....	26
VIII. Neural correlates of rewards processing .....	28
IX. Neural correlates of information valuation.....	32
X. Neural correlates of others-oriented processing .....	34
Synthesis.....	38
Synthèse .....	41
Chapter I The demand for information when the truthfulness of news is uncertain ...	45
Abstract .....	45
I. Introduction .....	46
II. Methods .....	51
III. Results .....	57
IV. Discussion.....	69
Chapter II Neurocomputational processes of inferring others' preferences for information and fake news .....	74
Abstract .....	74

---

I. Introduction .....	75
II. Methods .....	77
III. Results .....	93
IV. Discussion.....	106
Chapter III Testosterone Causes Decoupling of Orbitofrontal Cortex-Amygdala Relationship While Anticipating Primary and Secondary Rewards .....	112
Abstract .....	112
I. Introduction .....	113
II. Methods .....	116
III. Results .....	124
IV. Discussion.....	132
General conclusion .....	137
Conclusion générale .....	140
Chapter I. Supplementary Materials.....	143
Chapter II. Supplementary Materials .....	162
Chapter III. Supplementary Materials .....	195
References .....	199





# Remerciements / Acknowledgments

La thèse est un exercice d'une grande intensité. Elle pousse son rédacteur à aller puiser loin, jusque dans ses fondations, et en fait surgir autant l'impérieuse nécessité de la naissance à soi-même que les nombreux doutes qui viennent s'y fracasser. En cela, la thèse est vecteur de refondation. Ma famille m'aura vue grandir, la mer Méditerranée qui m'est si chère m'aura vue grandir et la thèse à sa façon en aura fait de même. Je voudrais remercier mes parents Bernard et Evelyne, mon frère Noam, ceux qui sont partis et ceux qui ont rejoint la famille pour leur soutien indéfectible. Je voudrais remercier l'environnement et la culture dans lesquels j'ai baigné, puisque je n'ai pas grandi seul. Une créature constituée de chaleur, d'eau salée et de multiples figures me transporte. Je voudrais également remercier Jean-Claude et Marie Claire qui m'ont fait confiance tout au long de l'exercice. C'est cette confiance qui m'a donné la possibilité d'essayer, d'échouer et de faire face à la thèse dans sa totalité. Cette thèse aura traversé jusqu'à ma vie la plus intime durant quatre ans. Sans cette confiance, je me tiendrais différemment devant vous. Je voudrais aussi exprimer ma gratitude aux membres du jury, qui permettent de nouer le travail effectué. Merci à vous, Grégoire Borst et Christiane Schwieren, d'avoir accepté le rôle majeur de rapporteur. Merci à vous, Edmund Derrington, Caroline Charpentier et Stefano Palminteri d'avoir accepté de participer à ma défense.

Mes amis les plus chers m'auront également été des piliers durant cet exercice. A leur façon, leur présence est infusée dans le manuscrit. La thèse a de beau qu'elle ne s'effectue pas seul et qu'elle invite à se mêler à ses semblables. Chacun a une place indisputable dans mon cœur, comme autant de figures dans un Panthéon personnel. Je souhaite remercier Nicolas, Adrien, Robin, Eva, Joffrey, Amaury, Loïck, Thibaut, Romain, Pierre, Mathilde, Quentin, Alexis, Florian, Jessica, Jérémy, Jean-Christophe, Adrien, Albane, Auriane, Thibault, William, Sonja, Hugo, Alexandre, Tanguy, Thomas. Vous m'avez apporté votre amour et m'avez appris à vous donner le mien. Je voudrais particulièrement remercier Rémi, Rémi et Julien sans qui cet exercice n'aurait pu être mené à son terme. Vous m'avez non seulement apporté une grande amitié mais également votre aide la plus directe et impliquée. Celles-ci m'ont été immensément précieuses. D'autres grandes amitiés ont bourgeonné durant la thèse, qui se sont rajoutées à ma trame personnelle. Je les espère fermement cimentées. Je remercie Mathilda, Maëva, Sébastien, Felipe, Alice, Etienne, Rossella, Maxime, Nicolas, Jimmy, Holly, Inès, Hadiyah, Sorravich, Axel, Flore, Toan, Marie, Alejandra, Jona, Siwar, Yao, Maxime, Mojtaba, Sanaz, Hager, Giulia, Jacopo, Maciek. J'en oublie et je vous prie de m'en excuser. Je donne aussi une place toute particulière à Ninine, à Dédé, à mes vieux amis et à ceux qui me sont liés mais que je n'ai pas ou peu connus. Enfin, j'inscris ici Buck en gardien de ce Panthéon. Merci.

# Abstract

The decisions to acquire and share information are of key importance in modern societies. Yet, the mechanisms underlying these decisions remain unclear. Acquiring new information helps update one's beliefs. Sharing information requires inferring whether this information is sufficiently valuable for others in terms of belief updating. These essays study why people seek information when it has no instrumental value, and how individuals infer others' information-seeking preferences. Combining behavioral economics and neuroimaging (model-based fMRI), we investigated the mechanisms and brain computations engaged in the decisions to *Receive News*, on one hand, and *Sending news*, on the other.

A first study explored a) how individuals evaluate the veracity of true and false information ("fake news") and b) the role of one's confidence in the decision to acquire extra information. We hypothesized that the confidence individuals hold about their estimation of information truthfulness is unreliable. Despite such a low metacognition, we expected that confidence drives information-seeking behavior. In a controlled environment, participants were asked their confidence levels about the truthfulness of various news before reporting their willingness to receive or not extra information. The main findings showed that participants' metacognition was not calibrated to assess information accurately, but it determined the desire to acquire extra information, demonstrating a key role of metacognitive monitoring.

A second study used fMRI to explore how individuals infer others' preferences for information and the underlying neurocomputational bases. We tested a Bayesian model in which individuals weighed beliefs about information truthfulness, the simulated population preference and the target agent's social distance. Participants were rewarded for matching the reception choice of other participants by choosing whether to provide them or not with extra information. The results showed that senders initially integrated beliefs about truthfulness probability and priors on receivers' beliefs, overestimating the receivers' preferences for information. When informed on the receivers' social distance to political organizations that were related to the information content, senders shifted their weight in favor of their beliefs about the latter. We showed that beliefs associated with truthfulness probability correlated positively with the Ventral Medial Prefrontal Cortex, bilateral Striatum and Dorsolateral Prefrontal Cortex. Beliefs associated with the distance elicited brain activity within the bilateral

Temporo-Parietal Junction. These results reveal the brain systems engaged in updating one's beliefs about others' preferences for acquiring extra information about news.

# Résumé

Les décisions d'acquérir et de partager des informations sont d'une importance capitale dans les sociétés modernes. Pourtant, les mécanismes qui sous-tendent ces décisions restent flous. L'acquisition de nouvelles informations permet de mettre à jour ses croyances. Le partage de l'information nécessite de déduire si cette information a suffisamment de valeur pour les autres en termes de mise à jour des croyances. Ces essais étudient pourquoi les gens recherchent des informations lorsqu'elles n'ont aucune valeur instrumentale, et comment les individus infèrent les préférences des autres en matière de recherche d'informations. En combinant l'économie comportementale et la neuro-imagerie basée sur des modèles, nous avons étudié les mécanismes et les calculs cérébraux engagés dans les décisions de recevoir des informations, d'une part, et d'en envoyer, d'autre part.

Une première étude a exploré a) comment les individus évaluent la véracité des informations vraies et fausses ("fake news") et b) le rôle de la confiance d'une personne dans la décision d'acquérir des informations supplémentaires. Nous avons émis l'hypothèse que la confiance que les individus accordent à leur estimation de la véracité des informations n'est pas fiable. Malgré une métacognition aussi faible, nous nous attendions à ce que la confiance détermine le comportement de recherche d'informations. Dans un environnement contrôlé, les participants ont été interrogés sur leur niveau de confiance quant à la véracité de diverses informations avant de déclarer leur volonté de recevoir ou non des informations supplémentaires. Les principaux résultats ont montré que la métacognition des participants n'était pas calibrée pour évaluer l'information avec précision, mais qu'elle déterminait le désir d'acquérir des informations supplémentaires, démontrant ainsi un rôle clé de la surveillance métacognitive.

Dans une deuxième étude, nous l'imagerie par résonance magnétique fonctionnelle pour explorer la manière dont les individus infèrent les préférences des autres en matière d'information ainsi que les bases neurocomputationnelles sous-jacentes. Nous avons testé un modèle bayésien dans lequel les individus pondéraient leurs croyances sur la véracité des informations, la préférence simulée de la population et la distance sociale de l'agent cible. Les participants jouaient le rôle d'émetteurs. Ils étaient récompensés pour s'aligner sur le choix de réception de récepteurs en choisissant de leur fournir ou non des informations supplémentaires. Les résultats ont montré que les émetteurs intégraient initialement des croyances sur la

probabilité de véracité et des priors sur les croyances des récepteurs, surestimant les préférences des récepteurs pour l'information. Lorsqu'ils recevaient un indice sur les croyances des récepteurs concernant le contenu de l'information, les émetteurs modifiaient leur comportement pour privilégier leurs croyances sur les croyances des récepteurs. Nous avons montré que les croyances associées à la probabilité de la véracité d'une information étaient corrélées positivement avec le cortex préfrontal médian ventral, le striatum bilatéral et le cortex préfrontal dorsolatéral. Les croyances associées aux croyances des récepteurs ont déclenché une augmentation du signal dans la jonction temporo-pariétale bilatérale. Ces résultats révèlent les systèmes cérébraux engagés dans la mise à jour des croyances sur les préférences des autres pour acquérir des informations supplémentaires sur les nouvelles.

# General introduction

## I. Modern-day news consumption

The information environment nowadays abounds in news and is characterized by an immense variety of supply and demand. The internet space progressively allowed for a wider connection between individuals among networks and for a faster and wider diffusion of information. Specifically, search engines, social medias, messaging applications and news applications have allowed for more frequent behaviours of information seeking, information aggregation and information sharing. With such multiplicity, both promises of better informed individuals and concerns about same information-pits they could fall into arose. Reports show that online news consumption is actually rather restricted. In 2019, a panel of 1711 UK internet users spent three to six percent of their online time with news media (Fletcher, Newman, & Schulz, 2020). In 2021, half of UK internet users from a different survey reported they consumed information from brands online – news websites or apps – in the preceding week. The other half’s diet was based on offline services, online search engines or social medias (F. Kelly et al., 2022). To assess how informed individuals might be in modern days, one has to ask how generalizable these observations about news consumption are.

Data shows there are disparities amidst consumers of information. In the UK, 22% of users engage in different news sources on a regular basis, 55% engage in a few different sources daily and 23% do not access news daily (Arguedas, Robertson, Fletcher, & Nielsen, 2021). In the Global North (Europe and Northern America), younger people consume news on their smartphone and through social medias while older generations use TV, radio and print. Generational divides extend in social medias as well: Twitter, Instagram and Snapchat are targeted by 18-to-24-year-olds while Facebook is more popular among those aged 25 to 34 (Röttger & Vedres, 2020). Internet use in 2018 in Europe was 80.1%, but geographical demographics further divide the information environment. This number drops from 74.6% in the Americas to 49.5% in Arab states, 46.2% in Asia & Pacific regions and 26.3% in Africa. Usage differentiates even more when generations, sub-regions and socioeconomic status intersect (Blank, 2017; Blank & Lutz, 2017; Hargittai, 2015).

It remains that mobile and messaging services are becoming the primary sources of news in many communities. Free online resources and advertisement in social medias have

affected the market. More competition erupted among new providers and professional journalism has led towards new survival strategies, such as paywalls, donations, digital subscriptions. This has limited local journalism, affected the survival of news outlets and the daily exposition to news (Newman et al., 2021). In summary, there is a growing portion of mobile or internet users, decentralization of news, fragmentation of actors in the news production, a greater emphasis on the role of news aggregators and a thinning portion of traditional news outlets.

The 2016 Presidential election in the U.S.A. has raised many concerns regarding the spread and impact of misinformation. What we define as fake news is typically low on facticity – it doesn't rely much on facts – and intentionally misleading – its immediate intention is to deceive readers (Tandoc, Lim, & Ling, 2018). Evidence suggests that, at the time of the 2016 election, fake news on Twitter accounted for about 6% of all news consumptions in months before the election. This consumption was heavily concentrated, with 1% of users exposed to 80% of the fake news (Allcott & Gentzkow, 2017; Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). It has been estimated that the average US citizen was exposed to one or two news articles at most during the election period (Allcott & Gentzkow, 2017). Recent evidences point towards a limited effect of misinformation campaigns. Among US citizens on Twitter, roughly 0.1% of users were responsible for sharing 80% of fake news (Grinberg et al., 2019). Users who engaged the most with misinformation content seemed to be already highly polarized. Average individuals are usually sceptical of information from the internet, perform cross-validation and diversify their news diet for such key events (Allcott & Gentzkow, 2017). Among all respondents from a survey of 46 media markets, 74% of them reported they prefer news that reflect a range of views, rather than seeking partisan news (Newman et al., 2021). In consequence, these campaigns weakly impacted political attitudes or behaviours at the time of the election (Bail et al., 2020). Although it has been a subject largely disputed over the years, this is consistent with a growing literature on the low impact of political campaigns aiming at persuading voters (Kalla & Broockman, 2018; Mercier, 2017).

Alongside diffusion of misinformation, there is a risk that online communities are separated from each other, blocking information from spreading between them and widening the sociocultural or economic gaps. The term *echo-chamber* encapsulates this phenomenon. Echo-chambers are “bounded, enclosed media space that [have] the potential to both magnify the messages delivered within it and insulate them from rebuttal” (Jamieson & Cappella, 2008). Consuming information from few sources can lock individuals in spaces of attitude-consistent



information and insulate them from cross-cutting exposure. The same effect can also arise from the over-personalisation of news feeds in search engines and social medias. A universe of information tailored or narrowed to its consumer is said to be a *filter bubble*. These two factors are candidates for explaining the concentrated fake news consumption that the 2016 election case illustrated.

Besides algorithmic selection, that showcases like-minded content, echo-chambers might construct on a behavior called selective exposure. Selective exposure is the willingness to expose oneself to information that align with prior beliefs. For instance, people belonging to groups might choose to favour news sources that emit content about their groups or that prime group stereotypes. Recent research minimizes the prevalence of selective exposure. People in general maintain diversity in their information diet (Flaxman, Goel, & Rao, 2016), engage preferentially in cross-cutting exposure (i.e., exposure to disagreement) and use mostly public service broadcasting, paper news and news aggregating apps as principal news sources (Fletcher, Robertson, & Nielsen, 2021; Newman et al., 2021). However, people do also engage in selective avoidance and selective exposure (L. Bos, Kruikemeier, & De Vreese, 2016; Charpentier, Bromberg-Martin, & Sharot, 2018). Findings have highlighted the importance of political motivations in selective exposure, with great importance of partisan beliefs (Arguedas et al., 2021; Garrett, 2009; Tucker et al., 2018). When partisanship is at play, people are more prone to select information that reinforces their opinion and develop a slight aversion to what challenges it. With a tendency to favour reading opinion-reinforcing information, politically motivated individuals end up being more exposed to partisan news despite eagerness to engage with different perspectives (Garrett, 2009). Online survey data in seven countries estimated that between 2% and 5% of people consume only news sources with partisan content leaning on one side of the political spectrum (Fletcher et al., 2021). The portion of surveyed individuals inhabiting politically partisan echo-chambers is limited presumably because most people do not hold strong political beliefs. Only a small fraction may engage in information consuming behaviours that reduce cross-cutting exposure (L. Bos et al., 2016).

Observations about access to information call for caution when interpreting results about small communities. There is a fear that communities get locked inside echo chambers, fuelling political fragmentation and polarization and diminishing abilities to form accurate beliefs. Exposure to like-minded political content has the potential to polarise people or to

strengthen pre-existing beliefs. The risk of strengthening inaccurate beliefs amongst populations is however limited. Only a small and politically polarized subset of the population is the most exposed to and engaged with misinformation. Furthermore, few evidences point towards a prevalent existence of echo chambers or filter bubbles. Rather, it seems that people are generally quite eager to seek diverse information. A question that remains is, are people able to integrate this diverse information to their beliefs?

## II. Beliefs about the state of the world

How can one know something?

When one feels the weather at winter gets colder and colder sooner every year, one may be tempted to think that the world is going increasingly faster towards an ice age. One can then predict that crops will tend to freeze earlier each year. And one may very likely see this prediction verified. This would give more weight to the former proposition, even though a global warming predicts the same results.

How can one then, from a few singular facts such as “it gets colder at winter sooner every year”, form a universal proposition? This directional process from facts to proposition is called an inductive inference. It considers that what has been reliable in the past becomes knowledge about the future (Hume, 2003). The proposition “It gets colder at winter sooner every year” becomes “there is a global freezing”. However, repeated observations of a behavior, such as the sound coming from the struck of a chord, doesn’t necessarily imply that what was not observed will behave in a similar way. Repetition doesn’t necessarily imply truth. Repetition doesn’t necessarily increase knowledge.

To form knowledge, one may then choose to turn towards the use of a universal proposition to predict new singular facts. This is called a deductive inference, a process that gives which facts come necessarily from the premises of the proposition. They allow to form inference of what is not yet observed. If ‘there is a global freezing’ is a universal proposition, one can deduce that ‘crops will tend to freeze earlier each year’. However, the power of deductive inference is limited. It gives knowledge about what is already implied by the proposition. Furthermore, it doesn’t give knowledge about what is outside the proposition. Even though repetition is not a guarantee, one cannot escape inductive inference if one wants to form new knowledge (Russell, 1912) or if one navigates an uncertain environment.

Many solutions have been brought to this problem. One of them is Russell's *principle of induction* that helps performing inferences from which we can learn something (Russell, 1912): A) the higher the number of cases an object *A* has been associated to an object *B*, provided we didn't find them dissociated, the greater the probability they will be associated in a new case; B) under the same conditions, enough associations between *A* and *B* will make the probability of a new association tend to certainty. Repeated observations of an association increase the probability that these two things are associated.

The issue remains that propositions admitted as universal have been verified for a limited number of singular cases. To account for what we didn't see, a now widely accepted solution comes from Karl Popper (Popper, 1959). Let's consider a proposition about an association between two objects. We can consider this proposition, a theory, as true until enough contradictory evidence falsifies the theory: the more it resists to tests without having been falsified, the more this theory is corroborated. Accordingly, we will look for tests to falsify the theory of the 'global freezing'. To complement this 'falsification' principle, a second 'replication' principle tells us that the more the results predicted by the theory replicate, the more the theory is corroborated.

Although disputed, Popper's two principles is a great way to tackle the uncertainty of the world. One updates the corroboration, or truthlikeness, she gives to a proposition according to the results of the tests it went through. Such a proposition, which can be more or less true, is termed 'belief'. Most beliefs are about hidden states of the world that can be only partially, if not incompletely, observed. Consequently, they are held with certain degrees of uncertainty. Thoughts and intentions of neighbors are very partially accessible; noise in a long distance call blur messages exchanged between recipients; detection of black holes requires inference from indirect observation of its effect on interstellar matter. Beliefs help people form internal models of the world, which is useful in pursuing goals and obtaining outcomes as long as they are accurate. To form knowledge, one should thus be motivated to form accurate beliefs from the information one encounters, be it perceptions or statements. Assuming a new information is accurate, integrating it to one's beliefs should increase the accuracy of representations of the world.

Examining people's beliefs about themselves or events, however, often reveals unrealistic perceptions (Sharot & Garrett, 2016). For instance, people often hold false representations of themselves, such as perceiving themselves as more skilled than they really

are (i.e., meta-ignorance, known as the Dunning-Kruger effect, Kruger & Dunning, 1999). People may attribute great value to impressive but vague assertions, even though they are devoid of meaning (Pennycook, Allan, Nathaniel, Derek, & Fugelsang, 2015). People are prone to ontological confusions, such as attributing mental states to inanimate objects (Lindeman, Svedholm-Häkkinen, & Lipsanen, 2015) or holding beliefs about supernatural events that contradict folk beliefs (Pennycook et al., 2015). People may also simply suffer from an optimism bias. They may overestimate probabilities of a favorable future event, such as good news, and fail in predicting its outcome (Armor & Taylor, 2012; Shepperd, Waters, Weinstein, & Klein, 2015).

These inaccurate beliefs question the way human beings form beliefs and integrate new information to update them. Evidence suggests beliefs are not solely based on evidence but are motivated as well. Not only do beliefs are valuable because they help action selection but they also induce feelings in people who hold them. Maintaining an optimistic view about oneself and events induces positive feelings, acting as a reward: people that care about expected future utility are happier if they are also optimistic, overestimating probabilities of high reward (Brunnermeier et al., 2003). After a stressful social-evaluation, self-beneficial belief updating can act as a coping strategy to recover from stress-induced negative affect (Czekalla et al., 2021). On the other hand, most individuals are averse to ambiguity (i.e, when probabilities about the possible states of the world are unknown, see Camerer & Weber, 1991) in part because it may induce feelings of anxiety (Birrell, Meares, Wilkinson, & Freeston, 2011). People are then motivated to form positive beliefs rather than remaining uncertain, and to discount negative beliefs. This phenomenon affects low-level processes such as learning as well. In a simple task of learning associations between symbols and monetary gains and losses, participants learn more easily when symbols are associated with better-than-expected outcomes, relative to worse-than-expected ones (Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017). This shows the preference for good news is at the heart of unrealistic optimism.

The consequence of motivated beliefs in the face of competing evidence is an asymmetry towards new information. Someone who believes *a priori* that there is no current temperature rise, because crops are freezing, will integrate more easily new information that are in line with their beliefs – a good news in regard to prior beliefs. In contrast, information which goes against are less unlikely to make him change his beliefs. An unexpected bad news

– that temperature rise is likely to be greater – will fail to change beliefs (Sunstein, Bobadilla-Suarez, Lazzaro, & Sharot, 2016). For believers in climate change, evidence that temperature is greater than planned is a bad news for them and for the planet, but a good news justifying their previous concerns. Such information may even be a factor of polarization between the two camps.

### III. Motivated beliefs shape information selection

As illustrated by our previous example, to form or update beliefs, people select the information they integrate. The motivation behind beliefs underlies the motivation to select information that produces beliefs. For instance, humans, monkeys, and other animals seek cues in their environment that may inform them of future outcomes, even when the cues produce no effects on the outcomes (Blanchard, Hayden, & Bromberg-Martin, 2015; Charpentier et al., 2018). Monkeys are willing to sacrifice water to satisfy this curiosity. In situations where beliefs at play are self-relevant, people integrate differently desirable and undesirable information. We have seen that they usually discount bad news and integrate good news. This valence-dependent asymmetry motivates beliefs updating, with under-weighting of undesirable information. For instance, people favour the good news about the likelihood of encountering aversive events, relative to bad news (Sharot & Garrett, 2016). They favour the integration of positive information over negative about objective personal abilities or traits (Eil & Rao, 2011), social feedbacks (Korn, Prehn, Park, Walter, & Heekeren, 2012). Besides helping in pursuing goals, beliefs bear an intrinsic value, as it reduces feelings of uncertainty or elicit positive feelings about oneself. People also show a preference for knowledge over ignorance that is valence-dependent. In a fMRI experiment, participants observed at each trial a pie representing the likelihood of winning or losing money. Then, they were proposed to choose between two offers representing the likelihood of the outcome being revealed. Participants selected the most informative offer more often during gain trials than loss trials. They were more likely to choose the most informative offer on gain trials when they were more likely to win. On loss trials, the most informative offer was more likely to be chosen when they were less likely to lose.

There is evidence that organisms select internal information they integrate to form beliefs. In many perceptual tasks, participants are rewarded for correctly reporting the true state of a noisy sensory stimulus. For instance, when asked to report the most frequent movement in a cluster of arrows pointing in many directions, participants receive a mix of information and noise that is transformed into evidence about the true state. Participants apply a decision rule to

report from the mix what the true state they believe is. This belief about the true state from sensory evidence is represented at the neural level by the activity of neurons (Gold J & Shadlen M, 2001). A study in humans recently reported a distinction between what the brain oculomotor system perceives and what participants report. The oculomotor system was able to detect fluctuations in velocity of a moving target, with corresponding correct tracking of the target by the eyes. Despite this correct tracking, participants made fewer verbal correct responses than the eye movements. Perceptual errors were “independent of the accuracy of pursuit eye movements” (Tavassoli & Ringach, 2010). Ethan S. Bromberg-Martin and Tali Sharot argue this result is evidence that the brain simultaneously contains multiple beliefs about the world. This ambivalence would offer opportunity for motivations to influence beliefs formation (Bénabou, 2015; Bénabou & Tirole, 2016; Bromberg-Martin & Sharot, 2020). In a fMRI task, when participants were shown ambiguous images (i.e., scene versus face), monetary incentives favoured the belief for a particular state of an image. The motivation to see a scene increased the neural activity within the visual cortex associated with the categorization of an image as a scene (Leong, Hughes, Wang, & Zaki, 2019). These observations advocate for motivated information selection behind perceptual beliefs as well. We seem to select information that form or update beliefs in accordance with our own interests.

We have seen that beliefs are motivated by their utility in regard to external and internal outcomes. Both drive information selection, which might result, potentially, in multiple beliefs simultaneously at play. However, one is held over another. Motivated by their beliefs, people might choose to disregard information that may bring negative internal utility. They might pursue this choice even though this information might be valuable in regard to their goals, leading to inaccurate beliefs (Charpentier et al., 2018).

Two main aspects of beliefs at least are valued by individuals: valence and certainty. People are averse to negative beliefs. Uncertainty can induce negative feelings as well, motivating people to avoid it. In anticipation of an impending electric shock, many people prefer receiving a higher shock sooner than waiting for the initial shock. The cost of the waiting – the dread – influences the cost of the outcome (Berns et al., 2006). This suggests that the information of the timing about the future outcome – whether pleasure or pain – has a value in itself.

## IV. The value of information

In contrast to information selection, information-seeking is a behaviour of active pursuit of knowledge. People seek information to gain knowledge about uncertain future outcomes. Once they estimate they acquired enough information, they can stop seeking and exploit information for decision-making (Friston et al., 2015). For instance, animals foraging for food or shelter have to decide which environmental locations to allocate their efforts and when to stop. Many species integrate both personal and social information to decide about effort allocation (Hills et al., 2015). Information-seeking is to be differentiated from curiosity, the feeling of wanting to know. Individuals engage in curiosity-driven behaviors to experience the observation of states of the environment that have predictive power, such as learning a skill. Experiencing a state of the environment (i.e., being curious) is rewarding in itself (Loewenstein, 1994) and drive individuals to come back to it when they feel the state has not been exploited enough (Still & Precup, 2012).

To decide whether seeking information and for how long, people estimate the future impact of that information regarding different motives. Three motives have been distinguished : instrumental, hedonic or cognitive (Sharot & Sunstein, 2020). For each motive, the information value can be positive, negative or null. The instrumental value of information is the classical value in decision-making. It is positive when it is related to the perceived ability of the information to help making better future decisions, such as escaping losses or winning gains. If an information leads to worse outcomes, relative to not knowing the information, it bears a negative value. The hedonic value of information is tied to its expected capacity to impact feelings. It relates to how much an information induces positive or negative feelings. People usually seek what induces positive feelings and avoid what induces negative ones. For instance, people that worry about cancer and perceive a high cancer risk avoid visiting doctors (Persoskie, Ferrer, & Klein, 2014). Finally, the cognitive value of information is tied to its capacity to alter the understanding of the state of the world. People seek information about concepts they often think about (Loewenstein, 1994) in order to bring their internal models closer to the external reality. It is likely they seek information related to concepts that are highly activated and interconnected for them (Sharot & Sunstein, 2020). For instance, the concept of ‘global freezing’ may be frequently activated within one person and interconnected to the concept of ‘extinction’, leading to seeking information related to the ice Age.



After combining its estimated value in regard of the three motives, people predict what the information may reveal and choose whether to seek information or to remain ignorant of its content. This choice is made according to the uncertainties individuals wish to maintain or to clarify (Charpentier et al., 2018; Dana, Weber, & Kuang, 2007; Exley, 2016; Exley & Kessler, 2019, 2021; Foerster & van der Weele, 2018; Shalvi, Soraperra, van der Weele, & Villeval, 2019). Even though knowledge is seen as always valuable, people may choose active deliberate ignorance (i.e., not knowing), as a choice to seek uncertainty rather than reducing it (Hertwig & Engel, 2016). Ignorance might as well be chosen to avoid negative feelings or to have license to behave, for instance, free of ethical concerns (Freddi, 2021; Grossman, 2014; Grossman & van der Weele, 2017). Such a choice is said to be active in the sense that it is deliberate despite the availability and free access of information (Golman, Hagmann, & Loewenstein, 2017).

Evidence shows that individuals are able to accurately predict the impact that information will have on their internal states and external outcomes (Cogliati Dezza, Maher, & Sharot, 2022). In a lottery task with a possibility to receive information, the authors manipulated the uncertainty of rewards, the likelihood to receive good news and the instrumental utility of information. Gains and losses varied in probability and magnitude whereas information varied in usefulness. Participants were first presented with a lottery then were asked to rate their anticipated mood and their estimated probability to win. Next, they were presented with a choice to seek information about the outcome of the lottery. After being presented with the choice, they were asked to rate their current mood and to estimate their probability to win a second time. Authors found that participants' rating of anticipated affect correlated with the expected value of the lottery. Participants' ratings of probability to win correlated with the actual uncertainty of the lottery as measured by the dispersion of the probability to win. Models showed that participants then used those predictions to decide when to seek information. Participants were more eager to seek information when they expected information to positively impact their affect, when they were uncertain about the lottery outcome and when information had instrumental utility. Results showed that after receiving the information, participants were happier, less uncertain and made better decisions in the task.

A series of experiments investigated whether the information-seeking behaviours are stable across domains. By manipulating three domains (i.e., self-traits, finance, health), evidence has been brought that individual differences in the weighing of each motive (i.e., instrumental, hedonic, cognitive) were domain general (C. A. Kelly & Sharot, 2021). 89% of participants showed a dominant motive and 53% of participants assigned twice the weight to



the dominant motive relative to the other two motives. The dominant motives were different across participants. Within participants, motives showed stability over time and across domains. There was however an average tendency to favour the action motive for health and finance domains rather than the self-trait domain. Authors found as well that self-reported mental health predicted the weighing of motives, especially that of cognitive utility. These observations suggest that each individual's information-seeking behavior results from a combination of self-traits, hedonic states and domains.

It remains that beliefs can be inaccurate in regard to reality. People may mistakenly estimate the utility of information. Tali Sharot and Cass Sunstein report some biases that are likely to influence the perceived utility of information during information-seeking (Sharot & Sunstein, 2020). For instance, the illusion that oneself has control over situations and outcomes could lead to overestimate an information's instrumental value. Unrealistic optimism affects predictions of hedonic utility by overestimating the probability of good news and underestimating those of bad news. The illusion of knowledge consists in underestimating the value of an information, impacting most likely predictions of cognitive utility. These observations have important implications regarding individuals' perception of their abilities.

## V. Metacognitive abilities

There is countless data showing that humans have an explicit sense of their decisions' expected accuracy (Boldt & Yeung, 2015; Fleming, Huijgen, & Dolan, 2012; Fleming & Lau, 2014; Yeung & Summerfield, 2012). We monitor and control our own cognitive processes, a process called metacognition. This monitoring provides a representation of our errors and performances. It is influenced by tasks properties such as difficulty or uncertainty (Pouget, Drugowitsch, & Kepecs, 2016) and allows one to assign a value of confidence to one's decisions. This process can occur in the absence of external feedbacks and relies on mechanisms of internal evaluation (Boldt & Yeung, 2015). Evidence shows that humans make use of their metacognition to decide on their information-seeking behavior. For instance, a study investigated participants' ability to learn the value of two lotteries while tracking their beliefs about their value over time. The authors found that uncertainty about the value of lotteries led to an increase in exploration: the more participants were uncertain about their beliefs, the more they were likely to explore the two lotteries. Thus, the confidence in their beliefs was used as arbitration to decide between exploration and exploitation (Boldt, Blundell, & De Martino, 2019). It has also been found that low confidence is predictive of one's tendency to change

one's mind in similar situations and to seek more evidence before taking a new decision (Boldt et al., 2019; Desender, Boldt, & Yeung, 2018; Folke, Jacobsen, Fleming, & De Martino, 2017). Because of how central confidence in the outcome of a decision is to explaining information-seeking behavior, it is primordial to understand how the confidence people have in their decisions relate to the actual accuracy of their decisions.

It has been found that the correlation between one's degree of confidence and one's objective accuracy in a task highly depends on the decision processes at play. For instance, this correlation is typically pretty high for perceptual decision tasks (Boldt & Yeung, 2015) but can be low when involved in memory processes such as witness identification (Brewer & Wells, 2006). Several measures have been developed to assess the link between one's confidence and one's objective accuracy. For example, metacognitive sensitivity is a measure that gives the accuracy of one's confidence in a task. In the case one's confidence is not able to discriminate between correct and incorrect judgments, one has a low metacognitive sensitivity; in the opposite case, one has high metacognitive sensitivity. Another example is the metacognitive bias, a measure of the overall level of confidence independent of correctness of judgments (Fleming & Lau, 2014). When both confidence and accuracy (CA) are highly correlated, the CA relationship is said to be well-calibrated. As an illustration, a confidence of 60% in predicting days of freezing cold is well calibrated if freezing cold is observed on 60 out of 100 days that are subject to the prediction – provided that freezing cold is a sufficiently well-defined term to operate probabilities on.

Confidence on a proposition corresponds to the belief that the proposition is correct, based on evidence available. Defined as a belief, confidence takes the form of a probability about variables that mostly take two values: correct or incorrect. The variables on which confidence operates, however, can take more than two values and are often characterized by uncertainty. This uncertainty has its origin in the external world or in brain processes (Pouget et al., 2016). This makes metacognition vulnerable to uncertainty and prone to make both beliefs and subsequent information-seeking behaviors drift away from optimization of outcomes.

Unfortunately, people have an idea of when their personal uncertainty (i.e., their ignorance) is high but have little skills in estimating how high it is (Ungar et al., 2012). Intrinsically, uncertainty conveys the simple idea that knowledge is not fully established and that many outcomes are possible. Besides uncertainty arising from brain processes and represented at the neural level, people also face explicit representations of uncertainty. They

can take various forms, such as confidence numeric intervals, verbal expressions, graphics or a combination, with as many ways to interpret them. Because uncertainty conveys the idea that many outcomes are possible, because people interpret differently the various representations of uncertainty and because they are motivated to avoid it, uncertainty encourages biased interpretations of the uncertainty itself as well as interpretations of the content that is subject to uncertainty (Budescu, Por, & Broomell, 2012). In consequence, people form interpretations of uncertain objects according to their prior beliefs (Dieckmann, Gregory, Peters, & Hartman, 2017). This suggests for individuals the need to reconsider their ability to monitor errors and to form confidence about information when it involves uncertainty.

In summary, to decide on whether or not to seek information, individuals are driven by information already available to them and by the expected value of future information that they may encounter. Once a new information has been revealed and integrated, the seeking behavior is revised. Individuals have to estimate again the value of what is available and what may come in further search. When information is associated with decision-making, individuals assess first the confidence they have about the decision. This confidence determines whether uncertainty is low enough to engage in exploitation or whether there is a need for more exploration. The ability of individuals to estimate the value of exploring is likely to justify which choice is taken. The role of individuals' ability to monitor their errors, and the corresponding confidence they have in their judgments and decisions, is likely to bring them closer or farther from the formation of accurate beliefs.

## VI. Evaluating information

Humans implement their eagerness to seek diverse and challenging information with diverse media diets. Still, they are subject to their own beliefs when seeking information. In addition, there are victim of their own beliefs when evaluating information. We have seen that beliefs about the true state of sensory evidence is represented by neural activity of many single neurons, which is then aggregated in upper brain areas. During this process, the brain can adjunct additional beliefs, resulting in a discrepancy between a sensory signal and a reported judgment. Any information is either correct or incorrect; news is either true or false. Is there a difference between news, clouds of arrows moving in two directions, and information about the ups and downs of a financial market?

Information is a signal with statistical properties about the content – its uncertainty – that is subject to noise that alters the signal. Information also has semantic properties – linguistic units that contribute to the meaning of the signal. The brain treats different categories of stimuli such as faces, places, bodies, images, sounds, or motor stimuli via distinct neural patterns (Kanwisher, 2010). Every stimulus can embed different levels of semantic information, triggering different brain processes. For instance, patterns of responses when processing images of faces last longer than scrambled versions of the same images (Coggan, Baker, & Andrews, 2016). Reading meaningful items is supported by different brain areas than reading meaningless strings of letters (Simos et al., 2002). Reading news requires processing meaning. News elicit concepts that people more or less often think of. Moreover, interpreting the content of news is subject to social influences. It is reasonable to expect that processing news elicit specific cognitive processes relative to other types of information.

Beliefs about true and false news can be apprehended with two approaches. The first one is the sensitivity, or truth discernment. It is a measure of the accuracy of one's beliefs in information, computed as the average accuracy ratings of news that are true minus the average accuracy ratings of news that are false. For instance, a participant might be asked to report whether the claim of a news headline is true or false, and its response compared with the news truthfulness. The higher the accuracy ratings of true news relative to the accuracy ratings of false news, the more one's discernment of truth is accurate. The other approach is the bias, the extent to which one believes in news. It is computed as the mere sum of beliefs, such as the number of news headlines that were rated as true (Pennycook & Rand, 2021b). Multiple studies have investigated individuals' ability to detect lies, with results hovering around chance-level (Belot & van de Ven, 2017; Bond & DePaulo, 2006, 2008). Recent evidence showed that individuals are not better than chance at distinguishing between true and false videos about news events (Serra-Garcia & Gneezy, 2021) and worse than chance at detecting false income reports (Dwenger & Lohse, 2019; Konrad, Lohse, & Qari, 2014). Most studies investigating evaluation of news have been interested in politically motivated truth discernment – the ability of individuals to discern truth from falsehood when news that elicit political beliefs are either concordant or discordant with their beliefs. There is an indisputable link between partisanship and beliefs in concordant news. We could expect that partisanship biases truth discernment, decreasing capacity to discern news truthfulness. However, people are better at discerning truth from falsehood when news are politically concordant, relative to politically discordant news (Bago, Rand, & Pennycook, 2020; Pennycook, Epstein, et al., 2021; Pennycook & Rand,

2019a). These results may imply that political identity is not the prime explanation in discernment failures.

People use mental shortcuts to make decisions – called heuristics. These are efficient mechanisms that use less information, diminish computation time and often improve judgment accuracy. In counterpart, they can introduce errors or systematic deviations in judgments—called biases. Both heuristics and biases are tools of living beings, bringing adaptive solutions to problems (Gigerenzer & Brighton, 2009). Gigerenzer and Brighton take the example of a baseball player running to catch a ball flying. The player could, virtually, perform a set of differential equations from a collection of variables to predict the trajectory of the ball on the field. Or, he could also run and rely on a simple gaze heuristic with only one variable: maintaining the angle of the gaze constant by adjusting the running speed. The latest solution is as efficient as the former but less computationally demanding. The rationale behind heuristics is finding a good-enough solution rather than the best. Rules are typically used to find such solutions: first, search for cues in order of their validity or accessibility; then, stop as soon as two cues point to the same object; if there is a stop, infer that there is enough evidence for the object. The bias is simply the difference of the judgment from the rational norm, meaning that part of available information is ignored to reach the judgment. Heuristics ignore some information so a rule of thumb applies to many situations (more in: Gigerenzer & Brighton, 2009).

In the case of evaluating news, individuals use shortcuts such as fluency-based heuristics to form truthfulness judgments. Fluency-based heuristics are mental shortcuts based on how easily objects are processed. For instance, the more fluently individuals can process a painting, the more positive their aesthetic response will be (Belke, Leder, Strobach, & Carbon, 2010). In the case of information, cues of reliable and easy processing are likely to signal that an information is true (Marsh, Cantor & Brashier, 2016). As such, familiarity with content correlates with perceived accuracy of news (Pennycook & Rand, 2020). Repetition favours processing fluency and is known to increase belief in statements – even for those that contradict prior knowledge or that are highly implausible (Fazio, Rand, & Pennycook, 2019). Consistence between elements of a message and elements in memory creates a cohesion effect (Ecker et al., 2022). These illusions of truth are responsible for vulnerability to advertisement, repeated propaganda and misinformation (Unkelbach, 2007). Many more heuristics increase perceived credibility of content, such as the number of prior endorsements, social feedbacks, technical

quality, grammatical correctness, perceived expertise of the source or emotional processing (Maier, 2005; Pennycook & Rand, 2020).

Biases intervene in information processing as well. Partisan bias refers to a tendency in people to think or act in ways that favor their own political group or their own ideologically based beliefs. This engagement in acts or thoughts may be deliberate; we usually refer to the notion of bias when people are unaware that their political affinities have affected their behavior (Ditto et al., 2019). Whereas selective exposure is a motivated bias at play in information-seeking behavior, individuals tend to attribute greater weight to what aligns with their beliefs. In consequence, they favor examination of information they expect will align with their beliefs (Hart et al., 2009). Because perceptions are shaped by beliefs and values, when beliefs are at stake, individuals act and think so their identity is protected (Bago et al., 2020; Kahan, 2018). This motivated reasoning serves different motives. Motivated by defence motives, individuals defend their attitudes or behavior by either challenging contrary information or engaging in attitude-consistent information. Motivated by accuracy-motives, individuals engage with information with an open-mindedness to reach informed conclusion. Motivated by impression-motives, individuals engage with information to satisfy social goals. Accuracy considerations tend to reduce confirmation biases (Hart et al., 2009) whereas partisan motivated reasoning reinforces motives of defending partisan identification (Bolsen, Druckman, & Cook, 2014). The effect of political concordance on beliefs, however, seems no stronger than the effect of actual truthfulness of news (Bago et al., 2020; Pennycook & Rand, 2019b). People have a tendency to believe that information is predominantly true – in part because information we come across is often reliable (Marsh, Cantor, & M. Brashier, 2016; Pennycook et al., 2015).

Evidence about evaluation of political news articles suggests that most people mainly fall for false news due to a lack of cognitive reflection (Bago et al., 2020; Pennycook & Rand, 2019b). One of the main hypotheses about cognitive reflection was that the sophistication of reasoning magnifies politically motivated reasoning (Tappin, Pennycook, Rand, & Hanser, 2021). In turn, it would increase the effects of partisan biases. According to recent papers, it seems rather that engaging in cognitive reflection is weakly associated with the selective conformation of evidence to one's political ideology (i.e., identity-protective cognition). Instead cognitive reflection seems to improve performances in detecting false content, regardless of political concordance (Pennycook & Rand, 2021b). The reason might be that individuals engaged in sophisticated reasoning tend to weigh more their prior beliefs in face of new evidence (Tappin et al., 2021). When taking political beliefs into account, reasoning is

much more affected by beliefs that are moderated by cognitive reflection than by those moderated by partisanship. Moreover, deliberation tends to correct intuitive mistakes (Bago et al., 2020). Truth discernment is therefore more accounted for by cognitive reflexion. Beliefs in political news (i.e., believing that a politically concordant news is true), on the other hand, are more accounted for by partisanship. Indeed, relevant prior knowledge is associated with the ability to discern truth. When individuals evaluate news that are politically concordant, political knowledge correlates with truth discernment (Pennycook & Rand, 2021b). An effect of the same kind has been found for media and digital literacy (Sirlin, Epstein, Arechar, & Rand, 2021). Cases when reasoning may not improve accuracy are when prior beliefs are distorted or when people refuse to update their beliefs. Many processes impact the formation and updating of beliefs. How much one believes that evidence should lead to updating her beliefs (i.e., individuals' meta-belief) offers a general account of the role of high-level cognition in evaluating information (Pennycook, Cheyne, Koehler, & Fugelsang, 2020).

In summary, news would be signals not so different from other types of information. The level of semantic properties of information seems to be what adds complexity to the cognitive processing. We could reasonably expect the same complexity when processing movies or music. Because of complexity, forming judgments about news is all the more difficult. Processing news involves many more beliefs than simply judging the direction of a visual stimulus. The more beliefs are engaged, the more individuals might be personally involved in the processing. They become motivated to defend their beliefs. Reaching a decision of forming judgments involves heuristics and biases. In the case of news, these heuristics and biases become intertwined with self-regarding and other-regarding motivations.

## VII. Sharing news

One of the many sources of information is people themselves. To achieve external and internal outcomes, not only do people seek information but they also share information with others. Sharing affects attitudes and behavior by affecting others' beliefs. We share facts, thoughts, narratives, ideas, perceptions and content, may it be music, videos or statements. We share to different types and sizes of audience – either local communities or broad audiences – and we share information about ourselves, others or about the world at large. Sharing information helps achieve self-oriented goals, such as deriving meaning from events, triggering



social interactions or providing a cooperative agent with instrumental information. In the case of sharing negative information, it can also create opportunity for coping with events or facilitating bonding via gossip. Sharing also helps achieve other-oriented information, such as creating intimacy, building trust via reciprocation or promote a sense of community (Barasch, 2020).

Sharing has consequences. Individuals can incur monetary costs, social punishment in retribution for violated norms, reputation losses or be disadvantaged after disclosing private strategic information. To decide whether to share information or not, individuals have to estimate potential gains and losses in different dimensions. In other words, they have to estimate the value of the content and integrate social information such as social norms, attitudes or preferences (Scholz, Jovanova, Baek, & Falk, 2020). This suggests that they consider the extent to which information they might share is accurate. People might be motivated to share accurate information if they feel their reputation might be hurt otherwise. They might also be motivated to share inaccurate information if they want to disadvantage others. However, recent studies pointed a discrepancy between accuracy judgment and news sharing. When people consider sharing content that might be misinformation, they often fail to attend to accuracy of news. As a consequence, news truthfulness might have little impact on sharing intentions, resulting in unknowingly disseminating misinformation (Epstein et al., 2021; Pennycook, Epstein, et al., 2021, Serra Garcia and Gneezy, 2021). Using cues or strategies that make individuals shift attention to accuracy does increase the quality of news that they share (Pennycook, Epstein, et al., 2021). It could be that people are focused on factors other than accuracy.

Self-related processing is a meaningful input to the sharing decision (Scholz & Falk, 2020). People consider a wide array of features about information, such as its framing, valence, emotional arousing, salience or instrumental value of content (Cosme et al., 2022). They are also biased towards granting more weight to objects and attributes perceived as related to the self and attribute value to information in accordance with motivated beliefs. It has been hypothesized that information related to the self, either self-oriented or others-oriented, has higher subjective value when considering sharing information (Cosme et al., 2022). For instance, individuals consider the impact of an information on their relationship with others prior to sharing (Berger, 2014). They also engage in different sharing behaviors according to the audience that will or might be targeted by the information. Broadcasting an information is sharing with a large and ill-defined audience, whereas narrowcasting is sharing with a small and controlled audience (Scholz, Baek, Brook O'Donnell, & Falk, 2020). For instance, when



engaged in narrowcasting, people tend to tailor what they share to the specificities of groups or individuals they target. In the case of broadcasting, people face greater uncertainty regarding opinions and beliefs of their audience. To avoid risks, they monitor the appropriateness of the content they share. Accordingly, people consider the self-relevance and social-relevance of stimuli to compute the overall value of what is likely to be shared. The lower the overall relevance, the lower the propensity to share. Evidence shows that broadcasting is more strongly related to self-relevance than social-relevance whereas narrowcasting is more linked to social-relevance (Cosme et al., 2022). Additionally, the decision to share information with others is affected by the perception of others beliefs and preferences. Identification and involvement with a group as well as expected outcomes predict motivations to contribute in sharing information. For instance, individuals spread moralized content with identity-based motivations, such as maintaining a distinctiveness between the ingroup and outgroups (Brady, Crockett, & Van Bavel, 2020). The effect that receivers' motivated beliefs have on senders was investigated in a study with a sender-receiver game. In these games there is a state of the world. A sender observes the true state of the world and a receiver isn't aware of the true state of the world but wants to learn it. The sender has a preference for telling the truth and may receive a benefit for having the receiver to think the sender is telling the truth. The receiver receives a benefit for correctly assessing whether the sender was telling the truth. The first main finding of the study was a greater likelihood for senders to send false messages when the political party of receivers were aligned with the false message. The second main finding was a willingness of senders to pay to learn the political party of receivers on political topics. The senders then used the receivers' party information to choose the false messages (Thaler, 2021).

## VIII. Neural correlates of rewards processing

Choosing to seek information requires computing the value of what might be revealed, for better or for worse. Value computation is a central process not only to anticipate rewards and losses and but also to update value upon reception. When facing a choice, one has expectations about the outcome of each option. The option associated with the best expected outcomes has the highest subjective value and is therefore the most likely to be chosen. These expected outcomes stem from cues about the value of options and may evolve with time as similar choices repeat. Once the outcome is experimented, the true value of the option is revealed. One can then update its expectations for future choices. We might incur money or reputation losses as much as we could gain hedonic rewards when engaging in pursuit of

information. Correctly balancing the expected outcomes is key to reach a satisfying decision. Updating them once an information has been revealed is key to form accurate beliefs.

Many choices we make in the real world are based upon options whose outcomes are complexly commensurable. For a task as simple as selecting vegetables in the grocery store, the brain operates computations on options of different types, of varying quantities and defined by a combination of different attributes (e.g., color, size, taste, price, current metabolic state). This particular choice operates on a class of reinforcers we call primary rewards. Food, drinks, erotic rewards or shelter are essential for survival and reproduction, hence they have an innate value. The term reward is chosen for objects that induce positive reinforcement of behavior, by opposition to the term punishment for objects that induce negative reinforcement such as avoidance. Secondary rewards, such as money, power or reputation, are indirectly linked to survival and reproduction. Their value is abstract and is learned through associations to lower-level rewards (Knutson & Bossaerts, 2007).

A major neural circuit, the mesolimbic pathway, is dedicated to processing rewards. Its central component, the ventral tegmental area (VTA), projects its dopaminergic neurons to the circuit's other members, the nucleus accumbens (NAcc), hippocampus, amygdala and prefrontal cortex (PFC) (Nieh et al., 2013). Dopaminergic neurons in this circuit represent the value of rewards or reward-predicting stimuli by the means of corresponding neural activity. First the activity of individual neurons responds to the detection of a stimulus by increasing in intensity, then gradually increase or decrease to represent the subjective value of the stimulus. Besides the subjective value, some dopaminergic neurons process other aspects of rewards, such as magnitude or probability. The value is said to be subjective in the sense that is modulated by goals and motivations. Neural activity of these neurons induces movements towards the rewarding objects, induce emotions, such as pleasure, and induce learning. This third function is especially important for one to modulate in time its expectations of rewards. Once a reward has been experienced, the activity of most of the dopaminergic neurons represent the reward-prediction errors, the difference between the predicted reward and the experienced reward. Rewards with higher-than-predicted value induce dopamine activations, rewards with lower-than-predicted value induce decreases in dopamine activity and rewards whose value was accurately predicted do not affect the dopamine activity. This reward-prediction error activity represents the update of the value of rewards of reward-predicting stimuli (more in Schultz, 2016).

Activity in these neurons is not sufficient to explain how the organism selects the most appropriate option from a set of possibilities. The organism is required to evaluate and compare peers with peers. Humans are willing to sacrifice money to catch sight of attractive faces (Hayden, Parikh, Deaner, & Platt, 2007). Nonhuman male primates are willing to sacrifice drinks to glimpse female perinea (Deaner, Khera, & Platt, 2005; Gomes & Boesch, 2009). To output a decision, the values of rewards are all put on the same neural scale. Strong evidence has showed the ventral striatum (VS) and medial PFC (mPFC) are responsible for encoding the subjective value of rewards in a common neural currency, a neural signal that represents diverse information on a single scale. The VS Blood-oxygen-level dependent (BOLD) activity represents magnitude and probability of rewards at the neural level, both when anticipating or receiving primary and monetary rewards (Bartra, McGuire, & Kable, 2013; Kable & Glimcher, 2007; Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007a). The ventromedial PFC (vmPFC) is also systematically activated upon receipt of rewards (Sescousse, Caldú, Segura, & Dreher, 2013) and responds to cues of both primary and secondary reward cues (Haber & Knutson, 2010; Sescousse, Li, & Dreher, 2013). It is sensitive to their subjective value, varying with rewards magnitude, probability and delays of reception, satiety upon reception and personal preferences (Haber & Knutson, 2010; Kable & Glimcher, 2007; McClure, Laibson, Loewenstein, & Cohen, 2004; J. O'Doherty et al., 2000; Plassmann, O'Doherty, Shiv, & Rangel, 2008; Sescousse, Li, et al., 2013; Small, Zatorre, Dagher, Evans, & Jones-Gotman, 2001). However, its role in computing a common reward currency is more prominent (Levy & Glimcher, 2012). For instance, following exposition to images of money and female faces, an fMRI study required participants to choose their willing to pay for the opportunity to view female faces varying in attractiveness. An anterior part of the vmPFC tracked the subject-specific values for each reward type whereas a posterior part of the vmPFC predicted the exchange rate between money and faces (D. V. Smith et al., 2010). Another study reported neural activity in the same region within subjects associated to rewards subjective values, regardless of their type (Sescousse et al., 2010, 2014).

Reward processing is separated into two temporally distinct processes: reward anticipation and reward reception (Knutson et al., 2007a). Processing rewards upon reception recruits a large network (Sescousse, Caldú, et al., 2013; Sescousse, Redouté, & Dreher, 2010). The vmPFC and bilateral VS, along with bilateral amygdala, bilateral anterior insula (AI), and mediodorsal thalamus all respond to the hedonic value of monetary, erotic or food rewards once the outcome has been unveiled. Accordingly, it has been deemed to be the 'common reward

network'. The vmPFC, extending to the pregenual anterior cingulate cortex (pgACC), and VS reciprocally project to each other and all areas within the network maintain structural or functional connections (Kringelbach & Radcliffe, 2005; Sescousse, Caldú, et al., 2013). Reward anticipation elicits a rather similar network of areas. Ventral striatal activity and vmPFC has been regularly reported during reward anticipation; vmPFC, amygdala, AI and thalamus activity are found with more inconsistency in the literature (Bartra et al., 2013; Diekhof, Kaps, Falkai, & Gruber, 2012; Knutson, Adams, Fong, & Hommer, 2001; Oldham et al., 2018; Wilson et al., 2018). Oldham and colleagues shown that monetary incentive delay tasks prompt VS, insula, amygdala and thalamus activity. Whether the network processes anticipation of all rewards similarly is yet to be confirmed. Primary and secondary rewards elicit different processes and overlaps: values associated to secondary rewards are abstract and both reward types processing show a distinct abstract-to-concrete organization in the OFC (Li et al., 2015).

All areas demonstrate overlapped specialization within the network. The amygdala activity is primarily sensitive to stimulus salience (Metereau & Dreher, 2013), such as the hedonic value of sexual stimuli, and receives numerous projections from cortical areas, including the VS and OFC (Carmichael & Price, 1995; Haber & Knutson, 2010; E. A. Murray, 2007). The AI integrates information about stimulus salience (Menon & Uddin, 2010; Rutledge, Dean, Caplin, & Glimcher, 2010), risk and uncertainty (Knutson & Bossaerts, 2007; Preuschoff, Quartz, & Bossaerts, 2008; Singer, Critchley, & Preuschoff, 2009). It is associated with the anterior cingulate cortex (ACC) within both the salience detection network and the frontoparietal control network (Cauda et al., 2011; Cauda, Geminiani, & Vercelli, 2014; Seeley et al., 2007; Vincent, Kahn, Snyder, Raichle, & Buckner, 2008). The AI is innervated by dopaminergic neurons and has anatomical and functional connections to all regions of the common network, as well as to the OFC (Cauda et al., 2011; Cloutman, Binney, Drakesmith, Parker, & Lambon Ralph, 2012; Ghaziri et al., 2017; Naqvi & Bechara, 2009; A. R. Smith, Steinberg, & Chein, 2014; Van Den Heuvel, Mandl, Kahn, & Hulshoff Pol, 2009). The mediodorsal thalamus is a relay structure between the basal ganglia and the PFC. It receives afferent connections from both the VS and the insula and projects to the PFC (Haber & Knutson, 2010; Öngür & Price, 2000). Its activity covary with the probability of reception (Galvan et al., 2005; Roiser, Stephan, den Ouden, Friston, & Joyce, 2010) and the intensity of rewards (Blood & Zatorre, 2001; Martin-Soelch, Missimer, Leenders, & Schultz, 2003; Redouté et al., 2000).

The VMPFC activity has been found to extend to a more anterior portion of the brain, the orbitofrontal cortex (OFC) (Sescousse, Caldú, et al., 2013). The OFC has neurons that

covariate with variations in the subjective value of stimuli and its neural activity represents the value of rewards across many dimensions (Grabenhorst & Rolls, 2011). This region is particularly interesting due to its anteroposterior segregation. Primary rewards engage the most posterior part; while secondary rewards engage the most anterior (Kringelbach & Rolls, 2004; Sescousse, Caldú, et al., 2013; Sescousse et al., 2010). This functional segregation follows an anatomical segregation, from agranular in its posterior region to granular in the anterior (Haber & Knutson, 2010). The anterior part is the most recent, ontologically speaking, and is more developed in humans relative to non-human primates (Öngür & Price, 2000). Interestingly, this segregation may correspond to a general organization of the region, where the progression from posterior to anterior is a progression towards abstraction and complexity of stimuli (Badre & D'Esposito, 2009; Dreher, Koechlin, Tierney, & Grafman, 2008; Koechlin & Summerfield, 2007).

## IX. Neural correlates of information valuation

People seek information to fill gaps in their knowledge (Golman & Loewenstein, 2018; Golman, Loewenstein, Molnar, & Saccardo, 2021). They pursue information when it is likely to reveal good news, when it has the ability to guide future actions towards high rewards or when uncertainty is high. Besides decreasing uncertainty, information is rewarding in itself. Not only is it endowed with a subjective value but evidence in primates reveals that individuals even respond to cues that signal upcoming rewards. In a neural recording procedure, during trials where receiving information did not affect the reception of an upcoming reward, monkeys still showed higher preference for receiving information rather than not. The midbrain dopamine neurons that coded for expectation of water rewards also coded for the expectation of information. The cue that indicated a large upcoming reward elicited neural excitation whereas the cue that indicated a small upcoming reward elicited neural inhibition (Bromberg-Martin & Hikosaka, 2009). Visual cues associated with future information gain about uncertain outcomes have also been reported to elicit activity in prefrontal cortical neurons, in both primates and humans (Blanchard et al., 2015; Charpentier et al., 2018). These results highlight the involvement of the dopaminergic reward system in processing information as rewards. They also indicate the predictive nature of information-processing signals in dopaminergic neurons.

Neurons typically encoding reward prediction have sustained activity from the first moment rewards can be predicted and scale with the expected magnitude of the reward until reward is delivered. In another study in primates, it has been demonstrated that inter-connected

subregions of ACC and basal ganglia, a structure at the top of the midbrain, responded in the same manner. The neural activity in the network encoded the prediction of information that would resolve outcome uncertainty and motivated behavior to obtain the information. When outcomes were uncertain, the presentation of a stimulus announcing an upcoming information elicited neural activity in populations of neurons in the two structures. The activity was lower when outcomes were certain or when information was not expected. Variations of activity in the ACC predicted future gaze shifts towards the information (White et al., 2019). Furthermore, ACC neurons have been reported to predict opportunities to gain information about upcoming punishments for primates that have preferences for receiving such information. Neurons were found in the structure that selectively responded to either cues associated with punishment uncertainty or cues associated with reward uncertainty. Authors reported similar neural behavior in ventrolateral prefrontal cortex (vlPFC) neurons (Jezzini, Bromberg-Martin, Trambaiolli, Haber, & Monosov, 2021).

Evidence for the involvement of the reward system in processing information has been reported in humans as well. Both states of high curiosity and opportunity to gain knowledge about favorable outcomes have been associated with signals in the ventral tegmental area (VTA) and the nucleus accumbens (NAcc), two regions from the dopaminergic system (Charpentier et al., 2018; Gruber, Gelman, & Ranganath, 2014). The brain activity in the VTA, and the substantia nigra, coding for the opportunity to receive information about gains as good news resembled that of information prediction error (IPE) reported in primates (Charpentier et al., 2018). These signals tracked a valence-dependent difference between the predicted event and the actual event. Activity in the midbrain scaled up for desirable information about uncertain rewards and scaled down for less desirable information about uncertain losses. Additionally, signals in the ventral striatum predicted information preference, indicating that dopamine projections might transform prediction errors into motivation to seek information. More interestingly, instrumental and non-instrumental benefits of information have been shown to be integrated in a single subjective value. This value correlated with activity in the striatum and the ventromedial prefrontal cortex (vmPFC) (Kobayashi & Hsu, 2019). These two regions are involved in processing the subjective values of rewards (Bartra et al., 2013).

When information has instrumental utility, it also has reward-related properties in that it is rewarding in itself. On the other hand, reward signals may embed informative attributes, such as information about gain probabilities in lottery tasks. It has been investigated in a model-based fMRI study whether the prefrontal cortex (PFC) independently encode reward and



information (Cogliati Dezza, Cleeremans, & Alexander, 2022). Authors found that activity for rewards and information overlap in the dorsal ACC and the vmPFC. However, when accounting for their shared variance, the activity in dACC correlated with non-instrumental and instrumental information value but not immediate reward value; the vmPFC showed the inverse pattern. More, authors found that signals of reward and information combined in the striatum. Similar results have also been reported in an electroencephalographic (EEG) study: monetary and informational prediction errors elicited signals with similar spatial and temporal profiles. Given source localization analyses, it is likely that these IPEs originated in the ACC (Brydevall, Bennett, Murawski, & Bode, 2018).

Altogether, these observations show that processing information recruits brain areas involved in reward processing. The dedicated dopaminergic neurons have been found encoding an information prediction error similarly as dopaminergic neurons encode reward prediction error. Evidence also suggests that, whereas the vmPFC and the dACC are both elicited by rewards and information, there might be a specialization in the dACC for processing information.

## X. Neural correlates of others-oriented processing

Studies investigating the neural correlates of information valuation reveal that there is a common neural coding scheme for rewards and information. This might result from properties of information such as an intrinsic rewarding value. In some situations, however, computing the value of information is computing the value in regard to other people. Information one seeks can be intended to be relayed to other individuals, either in raw form or after being processed. This is the case in journalism, research, activism, when reporting for a company or in social circles. Multiple studies have tested predictions that sharing involves self-related and others-related processing.

In a fMRI task, participants read headlines and abstracts of news articles prior to a sharing treatment. Intentions of sharing articles, relative to intentions of reading, elicited higher brain activity in self-referential processing areas, social cognition areas and correlated with preference ratings. Self-referential processing areas were defined as the medial prefrontal cortex (mPFC) and posterior cingulate cortex (PCC) whereas recruited social cognition areas were the dorsomedial prefrontal cortex (dmPFC) and temporo-parietal junction (TPJ) (Baek, Scholz, O'Donnell, & Falk, 2017). Activations in the mPFC and the PCC are consistent with other

studies involving self-relevant processing, such as thinking about personality traits (R. J. Murray, Schaer, & Debbané, 2012). Furthermore, intercultural studies investigating Western and Eastern representations of the self (Y. Zhu, Zhang, Fan, & Han, 2007) or culture-driven general versus contextual representations of the self (Chiao et al., 2009, 2010) found consistent activations in these two areas across cultures. Even though direct evidence is lacking, the elicitation of the mPFC in this situation suggests that in cases of sharing news with others, individuals compute the value of the news from their own perspective. The dmPFC and TPJ are reputed for their implication in understanding others' mental states, such as feelings, attitudes, goals and beliefs (i.e., theory of mind) (Rusch, Steixner-Kumar, Doshi, Spezio, & Gläscher, 2020). In studies investigating the effect of audience on sharing, not only did individuals simulate others' mental states when considering sharing but they also tuned their sharing behavior to their audiences (Barasch & Berger, 2014; Cosme et al., 2022; Scholz, Baek, et al., 2020). Sharing information with a broad audience (i.e., broadcasting) principally involved self-oriented motivations whereas sharing with a narrow audience (i.e., narrowcasting) mostly involved others-oriented motivations. Higher brain activity within the dmPFC and the TPJ have been reported when participants considered narrowcasting news articles, relative to when they considered broadcasting (Baek et al., 2017; Scholz, Baek, et al., 2020).

A model has been proposed to account for the above-mentioned observations about processes underlying decisions to share information with others. The value-based virality framework (Baek et al., 2017) assumes that the likelihood of sharing information is indexed to the perceived value of the act of sharing that information. Individuals maximize the subjective value of sharing by weighing the self-related and the social-related expected outcomes. Costs and benefits associated with each type of outcome are integrated in a single subjective value. In the framework, the more information is self and/or socially relevant, the higher is its subjective value.

Links between individuals' social preferences and social rewards, such as approval, and their valuation network have been long established (Fehr & Camerer, 2007; Rademacher et al., 2010). The neurocognitive processes for comparing values of options are similar for social and non-social decision making. For instance, corresponding neural mechanisms overlap between strategic social decision making and individual decision making (Lee & Seo, 2016). To understand how people put into action representations of the self and others is asking how people learn and integrate in brain structures self-related and others-related signals. Learning about others is the key to model their internal states and beliefs of other entities.



In social strategic behavior, humans are influenced by the presence of other people. They might be influenced by other-regarding preferences, estimate the impact of their own actions on others, punish social norms violations or think about the outcomes of others. In games, players might form and update beliefs about the choices of others, or their internal models, and chose the appropriate responses to maximize their outcomes. For instance, after an opponent chose an action in the rock-paper-scissors game, the value of each response action is increased or decreased according to the hypothetical outcomes they each lead to. These processes might involve specific models of others mind or simple temporal difference learning. The set of structures implicated in social decision making is relatively well identified in humans (Behrens, Hunt, & Rushworth, 2009; Hill et al., 2017; Lee & Seo, 2016) as well as in monkeys (Seo, Cai, Donahue, & Lee, 2014). In games involving strategy against opponents, the TPJ and interconnected parts of the vmPFC are implicated in estimating the impact of their own actions on their opponent's strategy (Hill et al., 2017) The TPJ is thought to encode social-specific outcomes and associated beliefs, such as reactivity from opponents, whereas the vmPFC is associated to a general implementation of behavioral strategies (Konovalov, Hu, & Ruff, 2018). Evidence indicate that the anterior insula (AI) and inferior frontal gyrus (IFG) might be responsible for updating beliefs with private information (Huber, Klucharev, & Rieskamp, 2013; S. A. Park, Goñame, O'Connor, & Dreher, 2017; Shamay-Tsoory, Saporta, Marton-Alper, & Gvirts, 2019; Toelch & Dolan, 2015; Wu, Luo, & Feng, 2016).

When people have to choose on behalf of others, to predict others' preferences or to predict others' decisions, they engage in more specific processes. In such situations, evidence shows that individuals both compute their own subjective preference in addition to simulating that of other agents (Kang, Lee, Sul, & Kim, 2013; Nicolle et al., 2012; Suzuki et al., 2012). They are able to do so by generating error signals when predictions about others' behavior are violated, similarly to reward prediction errors. For instance, individuals compare predicted others' actions with actual actions. 'Simulated-other's action prediction errors' (SAPE) have been found associated to blood-oxygen-level dependent (BOLD) signals in the dorsolateral PFC (dlPFC). However, the social environment is complex and embeds beliefs, preferences and intentions of others. Signals associated with the SAPE have been found in other structures such as the TPJ and the posterior superior temporal sulcus (PSTS) (Suzuki et al., 2012). Predictions about others' behavior can also be based on the simulation of others' valuation of items. 'Simulated reward prediction errors' (SRPE) were found associated with activity within the ventromedial prefrontal cortex (vmPFC) (Suzuki et al., 2012). Reports of activity within the

vmPFC might be puzzling as the area is also engaged in individuals' own valuation processes. Evidence shows that, rather than reflecting the preferences of self versus others, activity within the vmPPFC and dmPFC reflect the value signals for executed and simulated choices (Nicolle et al., 2012).

Inferring others' preferences is forming beliefs about others. This is typically done by sampling others' past actions. Others' past actions reflect their beliefs about the state of the world. Individuals can also use cues to form accurate judgments about others (Ambady, Bernieri, & Richeson, 2000; Ambady & Rosenthal, 1992; J. Park, Kim, Sohn, Choi, & Kim, 2018). Accordingly, they weigh and update both their prior personal beliefs and their beliefs about others' beliefs to estimate others' preferences (Joiner et al., 2017; Park et al., 2017). In tasks where individuals are shown other persons' responses, the ACC reportedly integrates and weighs that social information with the individual information (S. A. Park et al., 2017; Toelch & Dolan, 2015). Additionally, signals related to prediction errors in belief learning have been reported in the rostral ACC (L. Zhu, Mathewson, & Hsu, 2012). The ACC might then be a critical structure in representing and differentiating the self and the others by integrating multiple information. As such, a functional dissociation has been reported within the structure. Prediction errors related to others' choices were found in the ACC gyrus whereas neurons in the ACC sulcus seems to encode more self-referenced reward outcomes, in human and non-human primates (Apps, Lockwood, & Balsters, 2013; Apps, Rushworth, & Chang, 2016; Chang, Gariépy, & Platt, 2013). Overall, prediction errors signals occur for a wide variety of events that can be learned, such as reward values or action values. In the case of information, the valuation system is recruited to form a subjective value of information. The social processing system is recruited as well if individuals consider other-oriented behavior. The ACC seems to be a critical structure in integrating such self-oriented and others-oriented information to output motivations in more frontal structures. This motivation is then employed for actions or beliefs that are self- and/or other-oriented, such as in behavior of seeking and sharing information.

# Synthesis

Modern societies revolve around mass information exchanges. As the amount of information available increases, it is key to find and exchange the right information with the right people so beliefs can be adapted to the state of the world. People pursue information for instrumental motives, helping action selection, and non-instrumental motives. They engage in information-seeking when they estimate that what an information might reveal would be valuable enough (Bromberg-Martin & Monosov, 2020; C. A. Kelly & Sharot, 2021). This is also the case for news from media outlets or content from social medias. Indeed, information may be pursued because individuals are emotionally invested in the content of information (i.e., hedonic utility), because information might turn out useful in the long term (i.e., instrumental utility) or simply because it is related to concepts meaningful to them (i.e., cognitive utility) (Sharot & Sunstein, 2020). The intrinsic value of information is therefore associated to its capacity to reduce uncertainties about environmental contingencies (Cohen, McClure, & Yu, 2007; Golman et al., 2021; Jezzini et al., 2021; Kobayashi & Hsu, 2019). Humans choose to engage in information-seeking depending on the uncertainties they wish to clarify (Charpentier et al., 2018; Freddi, 2021; Shalvi et al., 2019). To decide about sharing an information with others requires inferring whether that information would be valuable enough for others.

Information value being tied to its capacity to reduce uncertainty, the information value one perceives is likely to stem from the perceived truthfulness of the statement, its uncertainty and one's estimated prior knowledge about the content of the statement. In the case of an insufficient estimated value for a specific information, it is to be expected that individuals will disregard information. In the context of information of cognitive utility, individuals favor information that are related to concepts meaningful to them and aim at minimizing the difference between their internal models and external reality. Information seeking minimizes this difference by improving internal models so they fit better the reality (Kappes, Harvey, Lohrenz, Montague, & Sharot, 2020; Sharot & Sunstein, 2020). Such a minimization requires accurate truthfulness discernment to estimate the value of information. However, discerning true from fake news has been shown to be difficult for individuals. Furthermore, the uncertainty about the level of facticity of content or the intention to deceive of the source makes the categorization between true news and false news all the more difficult. Critically, metacognition indexes beliefs about uncertainty and one's estimated abilities to choose the "good" or the "best" decision (Pleskac & Busemeyer, 2010).

In a first chapter, we investigate how individuals evaluate uncertain news with non-instrumental utility and how they subsequently choose whether to receive or not extra information that might resolve the uncertainty about the news. We look at the impact of uncertainty on truthfulness estimation, participants' metacognitive abilities and how both affect information-seeking behavior. To do so, we designed a novel task in which receivers evaluated the truthfulness of various news before declaring their willingness-to-pay to receive or not to receive extra information about the news.

When items have an instrumental value, individuals estimate others' preferences by computing the value they attribute to the item and simulate the other agent's value. This process is the same when choosing on behalf of others (Nicolle et al., 2012), when predicting others' likeness (Kang et al., 2013), or when predicting others' decisions (Suzuki et al., 2012). A similar process has been suggested for non-instrumental information. When deciding on sharing news with others, people consider the value of sharing for themselves and for others (Baek et al., 2017). However, the combination of motives behind information-seeking proves that information valuation is subjective and motivated by beliefs (Golman et al., 2021). Hence, how individuals use their beliefs to infer others' preferences for non-instrumental information remains unclear. And so do neurocomputational bases that underlie such a process. Information-seeking is known to implicate the mesolimbic pathway involved in reward processing. For instance, in primates, preference for information is signaled at the brain-level by activations of dopamine neurons (Bromberg-Martin & Hikosaka, 2009). In humans, preference for information and states of high curiosity are associated with signals in the nucleus accumbens and tegmental ventral area (Charpentier et al., 2018; Gruber et al., 2014). Finally, the integration of both information instrumental and non-instrumental benefits correlates with activity in striatum and ventromedial prefrontal cortex (Kobayashi & Hsu, 2019), two regions known to process subjective value (Bartra et al., 2013). On the other hand, social cognitive processing is known to elicit activity in dorsomedial prefrontal cortex and social areas like temporo-parietal junction and superior temporal sulcus (Joiner et al., 2017; Kang et al., 2013; Lee & Seo, 2016).

In a second chapter, we examine the mechanisms and brain computations engaged in the decision to share with others extra information that can reduce uncertainty about news. To do so, we designed a task in which participants evaluated the same news as the receivers in Chapter 1. Following the evaluation, participants had to infer whether receivers chose to receive extra information about the news, or not. In a control condition, participants were provided no

information about the receivers. In a cue condition, they were provided a cue about receivers' beliefs.

Rewards are generally distinguished into two types. Primary rewards, such as drinks, food or sex, are essential for survival and reproduction. They have an innate value. Secondary rewards, such as money, power or reputation, are indirectly linked to survival and reproduction. Their value is learned through associations to lower-level rewards. However, regardless of their type, rewards are processed within the ventral striatum and the ventromedial prefrontal cortex. When primary or secondary rewards are consumed, a whole brain network, called the common reward network, is dedicated to processing them. This network is composed of the ventromedial prefrontal cortex, ventral striatum, amygdala, anterior insula and medio-dorsal thalamus (Sescousse, Li, et al., 2013; Sescousse et al., 2010). Inconsistencies remain across studies about the implication of the ventromedial prefrontal cortex, amygdala, anterior insula and thalamus activity during reward anticipation (Bartra et al., 2013; Diekhof et al., 2012; Knutson et al., 2001; J. P. O'Doherty, Deichmann, Critchley, & Dolan, 2002; Oldham et al., 2018; Wilson et al., 2018). Nonetheless, evidence points towards a similar network.

In chapter three, we focus on brain activity during the anticipation of rewards of two types. We manipulate erotic and monetary rewards in an incentive delay task. Specifically, we investigate the effect of testosterone on neural activity within the reward network. Testosterone plays a strong role in modulating behavior and is implicated with all areas within the common reward network. It is known to affect the connectivity between the prefrontal cortex and the amygdala, resulting in increased sensitivity to rewards. We study how the effect of testosterone on this connectivity affects the processing of monetary and erotic rewards.

# Synthèse

Les sociétés modernes se sont concentrées autour des échanges d'informations à très grande échelle. Comme la quantité d'informations disponibles augmente continuellement, il est essentiel d'identifier les informations les plus correctes au regard de l'état du monde et de privilégier leur échange afin de garder ses croyances adaptées à l'état du monde. Les individus recherchent des informations pour des motifs instrumentaux, c'est-à-dire qui aident à la sélection d'actions, et pour des motifs non instrumentaux. Ils s'engagent dans la recherche d'informations lorsqu'ils estiment que ce qu'une information pourrait révéler aurait suffisamment de valeur (Bromberg-Martin & Monosov, 2020 ; C. A. Kelly & Sharot, 2021). C'est également le cas pour les brèves provenant d'organes de presse ou de contenus issus des médias sociaux. Les informations peuvent être recherchées parce que les individus sont émotionnellement investis dans le contenu de l'information (motivation hédonique), parce que l'information pourrait s'avérer utile à long terme (motivation instrumentale) ou simplement parce qu'elle est liée à des concepts significatifs pour eux (motivation cognitive) (Sharot & Sunstein, 2020). La valeur intrinsèque de l'information est donc associée à sa capacité à réduire les incertitudes sur les contingences environnementales (Cohen, McClure, & Yu, 2007 ; Golman et al., 2021 ; Jezzini et al., 2021 ; Kobayashi & Hsu, 2019). Les humains choisissent de s'engager dans la recherche d'informations en fonction des incertitudes qu'ils souhaitent clarifier (Charpentier et al., 2018 ; Freddi, 2021 ; Shalvi, Soraperra, & Villeval, 2019). Décider de partager une information avec d'autres nécessite d'inférer si cette information aurait suffisamment de valeur pour les autres.

La valeur de l'information étant liée à sa capacité à réduire les incertitudes, la valeur de l'information que l'on perçoit est susceptible de découler de la véracité perçue de l'énoncé, de son incertitude et de la connaissance préalable estimée du contenu de l'énoncé. Dans le cas d'une valeur estimée insuffisante pour une information spécifique, il faut s'attendre à ce que les individus négligent l'information. Dans le contexte d'informations à valeur cognitive, les individus privilégient les informations qui sont liées à des concepts significatifs pour eux et visent à minimiser la différence entre leurs modèles internes et la réalité externe. La recherche d'informations minimise cette différence en améliorant les modèles internes afin qu'ils correspondent mieux à la réalité (Kappes, Harvey, Lohrenz, Montague, & Sharot, 2020 ; Sharot & Sunstein, 2020). Une telle minimisation nécessite de savoir discerner la véracité d'une information pour estimer sa valeur. Cependant, il s'avère difficile pour les individus de discerner

les vraies des fausses informations (« news »). En outre, l'incertitude quant au niveau de facticité de leur contenu ou à l'intention de tromper de leur source rend la catégorisation entre vraies et fausses informations d'autant plus difficile. De manière critique, la métacognition indexe les croyances sur l'incertitude et les capacités estimées d'une personne à choisir la "bonne" ou la "meilleure" décision (Pleskac & Busemeyer, 2010).

Dans un premier chapitre, nous étudions comment les individus évaluent de brèves informations (« news ») incertaines avec une valeur non instrumentale et comment ils choisissent ensuite de recevoir ou non des informations supplémentaires qui pourraient lever l'incertitude sur ces informations. Nous examinons l'impact de l'incertitude sur l'estimation de la véracité, les capacités métacognitives des participants et la manière dont ces deux éléments affectent le comportement de recherche d'informations. Pour ce faire, nous avons conçu une nouvelle tâche dans laquelle les "récepteurs" évaluent la véracité de diverses informations avant de déclarer qu'ils sont prêts à payer pour recevoir ou non des informations supplémentaires sur ces nouvelles.

Lorsque des objets ont une valeur instrumentale, les individus estiment les préférences des autres à l'égard de ces objets en calculant la valeur qu'eux-mêmes leur attribuent et en simulant la valeur des autres agents. Ce processus est le même que l'inférence concerne le choix au nom d'autres personnes (Nicolle et al., 2012), la prédiction du goût d'autres personnes (Kang et al., 2013) ou la prédiction de décisions d'autres personnes (Suzuki et al., 2012). Un processus similaire a été suggéré pour des objets à valeur non-instrumentale comme les informations. Lorsqu'ils décident de partager des informations avec d'autres, les individus prennent en considération la valeur qu'à l'action du partage pour eux-mêmes et pour les autres. (Baek et al., 2017). Cependant, la combinaison des motivations qui sous-tendent la recherche d'informations prouve que l'évaluation des informations est subjective et motivée par des croyances (Golman et al., 2021). Par conséquent, la façon dont les individus utilisent leurs croyances pour déduire les préférences des autres pour les informations non instrumentales reste floue. Il en va de même pour les bases neurocomputationnelles qui sous-tendent un tel processus. La recherche d'informations est connue pour impliquer la voie mésolimbique impliquée dans le traitement des récompenses. Par exemple, chez les primates, la préférence pour l'information est signalée au niveau du cerveau par l'activation des neurones dopaminergiques (Bromberg-Martin et Hikosaka, 2009). Chez l'homme, la préférence pour l'information et les états de curiosité accrue sont associés à des signaux dans le noyau accumbens et l'aire tegmentale ventrale (Charpentier



et al., 2018 ; Gruber et al., 2014). Enfin, l'intégration des avantages instrumentaux et non instrumentaux de l'information s'est avérée corrélée à l'activité dans le striatum et le cortex préfrontal ventromédial (Kobayashi & Hsu, 2019), deux régions connues pour traiter la valeur subjective (Bartra et al., 2013). D'autre part, les processus socio-cognitifs sont connus pour susciter une activation du cortex préfrontal dorsomédial et des zones sociales comme la jonction temporo-pariétale et le sillon temporal supérieur (Joiner, Piva, Turrin, & Chang, 2017 ; Kang et al., 2013 ; Lee & Seo, 2016).

Dans un deuxième chapitre, nous examinons les mécanismes et les calculs cérébraux engagés dans les décisions de partager avec d'autres personnes des informations qui peuvent réduire les incertitudes sur d'autres informations préalablement traitées. Pour ce faire, nous avons conçu une tâche dans laquelle les participants évaluent les mêmes brèves informations que les "récepteurs" du premier chapitre. Après l'évaluation, les participants devaient déduire si les "récepteurs" avaient choisi ou non de recevoir des informations supplémentaires sur les brèves. Dans une condition contrôle, les participants ne recevaient aucune information sur les "récepteurs". Dans une condition d'indice, ils recevaient un indice sur les croyances des "récepteurs".

On distingue généralement deux types de récompenses. Les récompenses primaires, telles que les boissons, la nourriture ou le sexe, sont essentielles à la survie et à la reproduction. Elles ont une valeur innée. Les récompenses secondaires, comme l'argent, le pouvoir ou la réputation, sont indirectement liées à la survie et à la reproduction. Leur valeur est apprise par le biais d'associations avec des récompenses du premier type. Cependant, quel que soit leur type, les récompenses sont traitées par certaines aires cérébrales similaires, en particulier le striatum ventral et le cortex préfrontal ventromédial. Lorsque des récompenses primaires ou secondaires sont consommées, un réseau cérébral entier, appelé réseau commun de la récompense, est dédié à leur traitement. Ce réseau est composé du cortex préfrontal ventromédial, du striatum ventral, de l'amygdale, de l'insula antérieure et du thalamus médio-dorsal (Sescousse, Li, & Dreher, 2013 ; Sescousse, Redouté, & Dreher, 2010). Des incohérences subsistent entre les études concernant l'implication de l'activité du cortex préfrontal ventromédial, de l'amygdale, de l'insula antérieure et du thalamus pendant l'anticipation de la récompense (Bartra, McGuire, & Kable, 2013 ; Diekhof, Kaps, Falkai, & Gruber, 2012 ; Knutson, Adams, Fong, & Hommer, 2001 ; O'Doherty, Deichmann, Critchley, & Dolan, 2002 ; Oldham et al., 2018 ; Wilson et al., 2018). Néanmoins, certaines études vont dans le sens d'un réseau similaire.



Dans le chapitre trois, nous nous intéressons à l'activité cérébrale lors de l'anticipation de récompenses des deux types, qui ne sont pas des informations. Nous manipulons des récompenses érotiques et monétaires dans une tâche avec incitations retardées. Plus précisément, nous examinons l'effet de la testostérone sur l'activité neuronale du réseau de la récompense. La testostérone joue un rôle important dans la modulation du comportement et est impliquée dans toutes les zones du réseau commun de la récompense. On sait qu'elle affecte la connectivité entre le cortex préfrontal et l'amygdale, ce qui entraîne une sensibilité accrue aux récompenses. Nous étudions comment l'effet de la testostérone sur cette connectivité affecte le traitement des récompenses monétaires et érotiques.

# Chapter I

## The demand for information when the truthfulness of news is uncertain

V. Guigon (ISCMJ/GATE), and M. C. Villeval (GATE) and J-C. Dreher (ISCMJ)  
CNRS, Neuroeconomics lab, ISCMJ and CNRS, Groupe d'Analyse et de Théorie  
Economique (GATE) and Université Claude Bernard Lyon 1, Lyon, France

### Abstract

The growth of social networks platforms and the associated light-speed dissemination of fake news have generated interest in understanding how individuals assess the truthfulness of the news received and how metacognition regulates the decision to acquire extra information that might resolve uncertainty. Understanding how individuals process ambiguous information and integrate them with their own beliefs is of key importance to characterize how the brain evaluates information. We tested experimentally the relationship between the evaluation of the veracity of news with various degrees of content precision and the degree to which confidence modulates the willingness to pay to acquire extra information susceptible to resolve uncertainty. We confronted different Bayesian models to identify what determines one's judgment accuracy about the news truthfulness and confidence in one's judgment. We find that news assessment and confidence are both best explained by a Bayesian model integrating the news content precision and its veracity. A low level of confidence drives the demand for more information when the estimated judgment accuracy is low, revealing a key role of metacognitive monitoring in the assessment of uncertain news.

## I. Introduction

Social networks platforms have developed dramatically, with a number of users of social medias estimated to exceed 4.6 billion in 2022.<sup>1</sup> This growth has been accompanied by a light-speed widespread dissemination of false news, generating fears about the development of a society of misinformation. Indeed, false news and narratives have been shown to spread faster than factual news, notably because they look more novel and thus, attract more attention, and because humans have a limited capacity of discrimination between false and true information. For example, Vosoughi et al. (2018) found that the top 1% of false news cascades on Twitter reached between 1000 and 100000 persons whereas true news cascades rarely reached more than 1000 persons. Discerning true from false news has been proven difficult for individuals and relationships have been established between false news evaluation and major concepts such as cognitive thinking styles (Bago et al., 2020; Pennycook et al., 2015; Pennycook & Rand, 2019b, 2021b; Tappin et al., 2021), motivated reasoning (Kahan, 2018; Redlawsk, Civettini, & Emmerson, 2010; Strickland, Taber, & Lodge, 2011; Washburn & Skitka, 2018), or heuristics and biases (Baron & Jost, 2019; Ditto *et al.*, 2019; Knobloch-Westerwick, Mothes, & Polavin, 2020; Serra-Garcia, & Gneezy, 2021; Taber & Lodge, 2006).<sup>2</sup> The virality of false news has induced in reaction the development of organizations in charge of debunking the misinformation shared on the social medias. Therefore, people are now more aware of the existence of fake news on social medias and thus, on the uncertainty of the information received, but they also have the possibility to acquire information to reduce such uncertainty. If the question of the perception by humans of the truthfulness of information is not novel in the framework of strategic interactions, this major societal evolution due to expansion of social networks has generated a renewed interest in understanding how individuals assess the truthfulness of the news they receive and how they can reduce or resolve such uncertainty.

Correctly assessing whether a statement describes accurately a state or an event about the world and sampling more evidence to revise one's beliefs are primary to informed decision-making. The intrinsic value of information is related to its capacity to reduce uncertainty about

---

<sup>1</sup> Source: <https://datareportal.com/reports/digital-2022-global-overview-report> Retrieved on September 29, 2022.

<sup>2</sup> Note that the formation of inaccurate beliefs is a concern that exceeds the online environment. Phenomena such as pseudo-profound bullshit - seemingly impressive assertions, presented as true and meaningful, without discernible meaning (Pennycook et al., 2015), ontological confusions (Lindeman et al., 2015) or epistemically suspect beliefs (Pennycook et al., 2015) occur in laypersons and pave the way for inaccurate beliefs about the state of the world.

environmental contingencies (Friston et al., 2016) and people actively pursue knowledge by engaging in information-seeking – exploration – in order to later exploit the acquired knowledge. To decide what they want to know, that is, whether to seek information or avoid it, individuals have to estimate the value of such information. This estimated value resides in what the information may reveal and its expected impact regarding subsequent actions (Sharot & Sunstein, 2020). Based on this estimated value, individuals decide to seek information, to avoid it, or remain indifferent. This process has been shown to be asymmetric, as people are more prone to pay for information they seek when they expect this information to be positive rather than when it is expected to bring bad news (Charpentier et al., 2018).

In this study, we contribute to the reflection on individuals' ability to assess the truthfulness in information – whether it is true or false – with the aim of understanding the consequential desirability of additional information susceptible to reduce uncertainty. We aim at understanding how individuals estimate the truthfulness of (true or false) uncertain information and how their metacognitive sense of the accuracy of their estimations regulates their decision to acquire extra information that might resolve uncertainty. To do so, we test the relationship between the assessment of news truthfulness, metacognition and information seeking.

Motives for seeking or avoiding information can be classified as instrumental, hedonic, or cognitive (Sharot & Sunstein, 2020). The instrumental value of information seeking or avoidance resides in its perceived impact on the capacity of the individual to select subsequent actions that will increase rewards or avoid losses (Dana et al., 2007; Grossman et al., 2017; Charpentier et al., 2018; Serra-Garcia & Szech, 2019; Soraperra, van de Veen, Villeval & Shalvi, 2022). Its hedonic value resides in the expected affects – people usually seek positive affects and avoid negative ones, such as for emotion regulation during medical diagnosis (Persoskie et al., 2014) or in financial markets (Charpentier *et al.*, 2018). Its cognitive value resides in the capacity for an information to alter, positively or negatively, the understanding of the state of the world (Loewenstein, 1994; Sharot & Sunstein, 2020). Depending on their willingness to reduce or not uncertainty, individuals decide whether to approach information or not. They combine the estimated value of the three utilities to compute the estimated value of their information-seeking choice, and then decide to seek information when the estimated value is superior to that of remaining ignorant (C. A. Kelly & Sharot, 2021).

Being tied to its capacity to reduce uncertainty, the information value that the individual estimates is likely to stem from the perceived truthfulness of a given statement, its degree of

uncertainty, prior knowledge or familiarity about the content of the statement, and intentions (in the case of instrumental information). When the estimated value of additional information is deemed insufficient or repulsive, individuals are more likely to disregard such information. In the context of information that increases cognitive utility – which is the focus of the current study-, individuals are expected to favor information that is related to concepts meaningful to them and susceptible to minimize the difference between their internal models of the world and the external reality. For this category of information, information seeking aims at minimizing this difference by improving internal models so they fit better the reality (Kappes et al., 2020; Sharot & Sunstein, 2020). However, such a minimization requires truthfulness discernment to estimate the value of information.

False information can be described along a two-dimension typology: a level of facticity – the degree to which an information relies on facts – and a level of deception – the author’s immediate intention to deceive (Tandoc et al., 2018). False information is typically low in facticity and high in immediate intention to deceive. However, examples of naïve agents spreading mildly factive inaccurate information with no intention to deceive are common (see, *e.g.*, Serra-Garcia & Gneezy, 2021). On the contrary, content such as satire can bear low to high intention to deceive and be perceived as non-false information by an audience that is complicit with the information source, whereas non accomplices may be deceived. The higher is the uncertainty about the level of facticity of content or the intention to deceive of the source, the more difficult is the categorization between true *vs.* false information or between naïve agent *vs.* deceiver. In that grey area, receivers’ are at greater risk of interpreting information in a way that aligns with their prior beliefs (Budescu et al., 2012; Dieckmann et al., 2017). Individuals judge information with the help of heuristics, such as familiarity (Pennycook & Rand, 2020), or cues for credibility such as technical quality and grammatical correctness (Maier, 2005); processing fluency makes statements signalled as true and subsequently more believed (Unkelbach, 2007); illusory truth effect correlates with statements’ plausibility (Fazio et al., 2019). Such processes facilitate reaching a judgment decision but they can pull individuals away from correctness.

Critically, metacognition is expected to play a major role on the demand for further information. Indeed, it indexes beliefs under uncertainty and one’s estimated ability to choose the “good” or the “best” decision (Pleskac & Busemeyer, 2010). Confidence reflects accumulation of evidence in decision processes (De Martino, Fleming, Garrett, & Dolan, 2013;

Pleskac & Busemeyer, 2010) and perceived metacognition (Fleming & Lau, 2014). People report levels of confidence associated with their choices that correlate with performance, although the correlation between one's degree of confidence and the actual degree of accuracy is highly variable. The relationship between confidence in one's judgment and accuracy is typically high for perceptual decision tasks (Boldt & Yeung, 2015), lower for memory processes, such as witness identification (Brewer & Wells, 2006), and it varies with a task difficulty (Moore & Healy, 2008). Also, in a value-based learning environment, a low reported confidence about the estimated most rewarding option in a two-arm bandit task is associated with a higher tendency to explore the lower-value option (Boldt et al., 2019). Similarly, participants were more likely to seek information when weakly confident in a perceptual two-choices task, even when information-seeking was costly (Desender et al., 2018).

Actions, confidence in one's actions and information seeking have been tied together in models of cognitive processes. Drift-diffusion models incorporate information under the traits of evidence accumulation up to a subjective threshold leading to decision (Gold & Shadlen, 2001; Ratcliff & Rouder, 1998). Stemming from this literature, perceptual decision-making has been modelled as individuals inferring the state of the world via sampling partial information from sensory evidence. In such models, action and confidence both guide the need for sampling more evidence. Information-seeking is then predicted to be triggered at a threshold of confidence, given the cost of seeking relative to that of not seeking information (Schulz, Fleming & Dayan, 2021). When a sufficient threshold of confidence is reached, the perceptual decision is made. The likelihood of deciding to sample additional information before responding is also predicted by direct metacognitive judgments (Desender et al., 2018).

In the context of cognitive-utility information considered in this study, individuals are not expected to be, on average, better than chance at judging information truthfulness. Though prior knowledge can certainly help form a judgment and thus, decide about the need for further information, news typically exhibit varying levels of uncertainty. This uncertainty modulates the difficulty of assessing the truthfulness of the news and prompt biased interpretations. We expect both individuals' truthfulness assessment and accuracy in estimating truthfulness to be primarily affected by the level of precision of information content. A precise information content should lure individuals into interpreting information as true. Following rational accounts of decision-making, confidence in one's judgment should be based on beliefs and probabilities of events occurrence. Any systematic deviations from an objective standard in correctly assessing events, such as scoring rules for probability judgments, would be considered

a bias or suboptimal behavior (Arkes, 1991; Soll, Milkman, & Payne, 2013; Arkes, 1991). Biases in estimating probabilities would induce under-confidence or over-confidence regarding the actual accuracy, whereas well-calibration would translate into a confidence-accuracy correlation close to the objective standard (Fleming & Lau, 2014). In the case of the truthfulness estimation of non-ego-relevant news, we expect participants' success in estimating information truthfulness to be unbiased. Accounting for such performances, metacognitive abilities are nevertheless expected to be unreliable in the sense that participants' confidence is expected to be non-calibrated. However, we expect participants to use their confidence to primarily explain the demand for further information. This demand should increase when confidence in one's truthfulness assessment is at the lowest points.

To better identify the role of content precision in information truthfulness assessment and demand for further information, using various Bayesian mixed-effect models, we compare its effect with predictions from alternative accounts of failure to correctly evaluate information. This includes models with stickiness of prior knowledge, sociodemographic characteristics (age, sex, level of education, proximity to political organizations), cognitive reflection, and distrust toward expert sources of information. Lastly, we investigate the relationship between ambiguous information processing, confidence and information-seeking. We test for a mediating role of confidence in the effect of content precision on the demand for further information.

To test our predictions, we designed an incentivized within-subject experiment in which non-ego relevant information varied in content precision. The value of such information resides in its capacity to affect the understanding of the state of the world. We presented participants with a set of brief news from the press that could be either true or false. Participants were instructed to evaluate the truthfulness of each brief news and report their confidence in their judgment on a continuous scale, by using a probability elicitation method. Then, we elicited the demand for information by using a two-step procedure. Participants had first to decide on whether acquiring or not further information about this news. In the case they chose to receive more information, they had a chance to receive at the end of the task an investigation article from a fact-checking platform detailing content related to the brief news. Second, we elicited the willingness-to-pay of the participants to have their information-seeking choice implemented. In case of a choice to (not) receive more information, participants had to declare how much they were willing to pay to (not) receive the investigation article at the end of the

task. To control for prior knowledge, we elicited, prior to the behavioral task, participants' liking, familiarity and closeness of values with twelve political organizations in direct connection with the topics of the news communicated to the participants.

Our main findings show that participants' metacognition is not calibrated to estimate probabilities of truthfulness. Estimating truthfulness is best explained by a model integrating news content precision and veracity. Imprecision indeed signals falsity to participants and the more a news is imprecise, the more it is correctly categorized as false. Despite poor metacognition, it is the low level of confidence in truthfulness estimation that drives the desire to acquire extra-information. These demonstrate a key role of metacognitive monitoring in information evaluation.

## II. Methods

### II.1. Participants

269 participants with no history of neurological or psychiatric disorders participated in this online experiment on Testable.org. Data were collected in two waves, a first one with 80 participants in November 2020 and a second one with 189 participants spanning from December 2021 to January 2022. There were no differences in the design of the experiment between the two waves, except for additional questions in the final questionnaire, as mentioned below. Participants, mainly students in engineering and business, were recruited from the regular GATE-Lab subject-pool, Lyon, France. Two participants were excluded from the statistical analyses due to outlying response times ("RT") during news evaluation (one subject:  $RT = 51.79 \pm 26.35$ ; one subject:  $RT = 1.93 \pm 1.31$ ) compared to the mean response time ( $14.41 \pm 8.44$ ). Nine participants were excluded because they did not complete the final questionnaire. In total, 258 participants were included in the statistical analyses (127 males, mean age  $\pm$  SD =  $21.9 + 2.78$ ).

The study was approved by an internal ethics committee and complied with the European data protection regulation (GDPR). Written informed consent was obtained from all subjects prior to participation.



## II.2. Experimental Design

### II.2.1. Pre-Experimental Questionnaire

People worldviews have been shown to explain what people believe to be true (Tappin, Pennycook, & Rand, 2020). To have a proxy of such prior knowledge, in the first part of the experiment we instructed participants to rate various political organizations that were related to the different news domains. We selected 12 organizations active in the domains of ecology, democracy or social justice. Each organization was described by a 1000 character ( $\pm 20\%$ ) statement taken from the organization websites, with minimal manipulating of the original website content. Participants had to evaluate them with respect to six dimensions on a scale from 0 to 7 ([Supplementary II.1](#)). They indicated how familiar they were with the organization, how familiar they thought the organization was to their closed ones, how close were the organization's values to their own values, how close they thought the organization's values were to the values of their closed ones, how much they appreciated the organization, and how much they thought their closed ones appreciated the organization.

We computed the participants' adhesion to each organization (as a proxy of the knowledge of the domain) by aggregating their six responses in a score that was normalized on a scale from 0 to 100.<sup>3</sup> After rating the organizations, participants read the instructions on the task and filled in a comprehension questionnaire about these instructions.

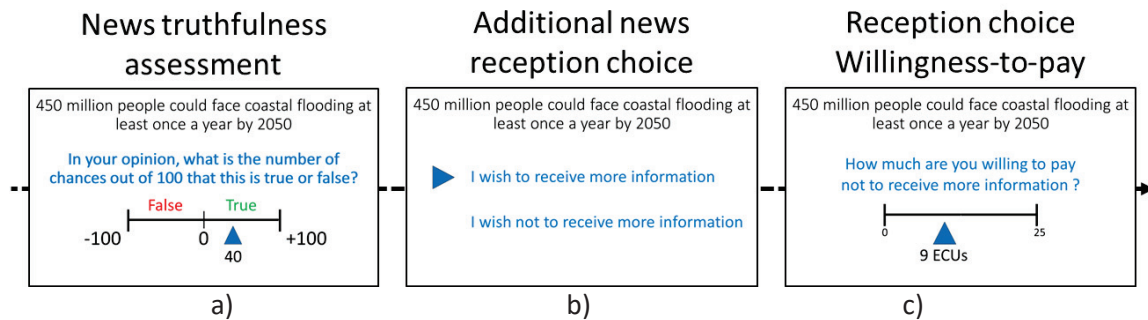
### II.2.2. The Tasks

The second part of the experiment consisted of two stages. The first stage included the truthfulness judgment task. Participants were divided into two groups that received 48 different stimuli each. Each of the 48 trials started with a fixation cross on the screen (Figure 1). Then, a brief news, either true or false, was displayed. Participants were asked to report what was, in their opinion, the number of chances out of 100 that this brief was true or false. Their response revealed their degree of confidence in their judgment. To respond, participants moved a slider either to the left (False) or to the right (True). Each move in a direction incremented their degree of confidence by 1%. Thus, the slider started at -100 on the left side and ended at +100 on the

---

<sup>3</sup> The higher the score, the more the participant was likely to adhere to the organization and be knowledgeable about its domain of activity. Among the democracy-related organizations, the Robert Schuman Foundation received the highest average score ( $59.62 \pm 16.6$  points), and Frexit the lowest score ( $37.27 \pm 14.88$ ). Among ecology-related organizations, the World Wide Fund got the highest average score ( $81.7 \pm 12.93$ ) and NIPCC the lowest ( $40.74 \pm 21.58$ ). Among the social justice-related organizations, the extreme were SOS Méditerranée ( $70.96 \pm 16.13$ ) and Generation Identitaire ( $41.58 \pm 18.06$ ).

right side. The elicitation of probabilities was incentivized, following the procedure of Karni (2009), as explained below.



**Figure 1: Description of the task.** Note: a) After a fixation cross, participants saw a brief news and were incentivized to report the probability that the news was true or false. A trial with a correct truthfulness evaluation was paid 50 ECU while an incorrect evaluation paid 0 ECU. b) Participants were then incentivized to choose between receiving more information about the news or not receiving more information. c) Given their choice, they had to indicate how much they were willing to pay to have this choice implemented. Depending on their bid, a BDM procedure determined whether their choice would be implemented and at which price, or not implemented.

The second stage corresponded to the elicitation of the demand for further information. After validating their assessment and while their screen was still displaying the brief news, participants were asked to choose between receiving more information or not receiving more information related to the same brief news. Finally, participants had to report how much they were willing to pay, between 0 and 25 Experimental Currency Units (ECU) of their 200 ECU initial endowment (with 100 ECU worth \$2), to have their receiving decision implemented (*i.e.*, to receive or not receive further information), using the Becker–DeGroot–Marschak (1964) method.<sup>4</sup> In the case participants opted for more information, they were eligible for receiving a debunk article investigating the content of the brief news in details. Debunk articles were taken from the French fake news debunk platforms *Les Décodeurs du Monde*, *AFP Factcheck* and *Libération Checknews* from the period 2017-2020. The additional information that was selected for reception was sent by email to the participants after the task. All these aspects were made common knowledge to the participants before they made their choices.

<sup>4</sup> Note that we elicited the willingness-to-pay to avoid receiving more information for the sake of symmetry of treatment with the elicitation of the willingness-to-pay to receive more information. We acknowledge that the monetary value of information avoidance cannot be estimated precisely in our design since it is always possible not to read the additional information received without paying for not receiving it.

### *II.2.3. Incentives*

At the end of the experiment, we randomly drew eight trials and rewarded correct truthfulness judgments in these trials. For each selected trial, one robot out of 100 robots was randomly drawn. To each robot was associated an accuracy level between 0 to 100, corresponding to the probability of this robot to provide the correct answer. Participants were aware that if the randomly drawn robot had an accuracy level higher than the subject's degree of confidence, we would take the robot's answer into account; otherwise, we would take the participant's answer into account. Each correct truthfulness judgment in these eight trials was paid 50 ECU.

We again randomly selected eight trials among the 48 to implement the BDM method. For each selected trial, if the participant's Willingness-to-Pay (WTP) was equal or above a randomly selected price between 0 and 25 (each price had an equal probability to be drawn), we deducted the selected price from his or her 200 ECU endowment and his or her decision was implemented. If the WTP was lower than the price, no deduction was operated and the option the participant did not choose was implemented.

At the end of the experiment, participants received the payoff from their performance in the estimation of the news truthfulness and the difference between an initial endowment and their payment for getting or avoiding additional information. They were paid on average \$15.92, including a \$9 show-up fee, for an experiment that lasted 46.38 minutes on average. Although subjects participated in the experiment from the French territory, the Testable.org platform rewards participation in experiments in US dollars. Participants were informed of this at the moment of their recruitment.

### *II.2.4. Stimuli*

To select the stimuli used in part 2, we followed the practical guide of Pennycook and colleagues for behavioral research on fake news and misinformation (Pennycook, Binnendyk, Newton, & Rand, 2021). We first designed a set of 210 true and false brief news (114 false news, 96 true news). The maximal length of each news was 140 characters, spaces included. We restricted the nature of these news to information with a cognitive utility, that is, factual information which content refers to concepts that individuals often think of with a capacity to alter their understanding of the state of the world. We avoided ego-relevant stimuli and stimuli that could elicit affects or have short term consequences on participants' daily decision-making.

The selected brief news described events or statements about ecology, social justice and democracy – three key themes that gained momentum as hot topics but did not directly concern participants' health (as would news related to the COVID-19 pandemic), nor personal economic situation. Some of the brief news were directly taken from the French fake news debunk platforms *Les Décodeurs du Monde*, *AFP Factcheck* and *Libération Checknews* from the period 2017-2020. Others have been fabricated from content on these platforms.

We then completed a pretest to ensure that our stimuli varied in content precision and met agreement regarding the themes of the news. We planned to keep 96 counterbalanced news after the pretest, with no statistical difference in content precision and capacity to make consensus on the related theme between the set of true news and the set of false news. The pretest ran on Testable.org and was rewarded \$7. Fifty-five independent raters (F=33, M=22; mean  $\pm$  SD age=26.2  $\pm$  4.78) were submitted the 210 true and false news. Five groups of 11 French-speaking raters evaluated each a set of 42 news out of the 210. To evaluate the stimuli content precision, desirability, consensuality and themes, raters had to answer the five following questions: 'On a scale of 0 to 10, how much would you say the content of this information lacks precision? (0 = not at all imprecise - 10 = very imprecise)'; 'On a scale of 0 to 10, how much would you like to know more about the content of this information? (0 = I would not at all want to know more at all - 10 = I would very much want to know more)'; 'On a scale of 0 to 10, how divisive and how likely to divide opinion do you think the content of this information is? (0 = not at all likely to divide opinion - 10 = very likely to divide opinion)'; 'Which theme do you think best fits the content of this information: Ecology, Democracy, Social justice, Health, Economy, Education, Identity, Security, Travel, Freedom, None of the above'; 'What other theme do you think would best fit the content of this information?'. Themes other than Ecology, Democracy or Social Justice were distractors in the forced-choice question.

For each theme and each truthfulness level, we kept the 16 news that reached the highest agreement on theme. As a result, the final set of stimuli included 96 counterbalanced true and false news that were categorized as either democracy-related, ecology-related or social justice-related. We computed the Intraclass Correlation Coefficient (ICC3k) for the measures of content precision, consensuality and desirability. The average fixed raters correlation coefficient was equal to 0.514 for content precision (CI [0.356, 0.65]), 0.8 for consensuality (CI [0.74, 0.86]) and 0.6 for desirability (CI [0.465, 0.71]) (all  $p < .001$ ). No difference was found between true (mean  $\pm$  SD= 5.53  $\pm$  1.24) and false (mean  $\pm$  SD =5.17  $\pm$  1.25) news content precision distributions (*ranksum*,  $p=0.09$ ), nor between true (mean  $\pm$  SD= 6.24  $\pm$  1.57) and

false (mean  $\pm$  SD= 6.55  $\pm$  1.55) news consensuality distributions (*ranksum*,  $p=0.39$ ). Moreover, content precision and consensuality were highly correlated ( $Rho=0.478$ ,  $p<0.001$ ). Overall, content precision was balanced between false and true news, and both were equally divisive among raters. We also found a (weak) correlation between the stimuli desirability, as evaluated by the raters, and the choices of the participants in the experiment to receive more information about the brief news ( $Rho=0.349$ ;  $p<0.001$ ).

### *II.2.5. Post Experimental Questionnaire*

At the end of the online experiment, participants had to fill in a set of questionnaires to measure notably their degree of curiosity. Epistemic curiosity may respond to two separate desires (Litman, 2008). One desire stems from an expectation to stimulate positive feelings of intellectual interest, while another one stems from an expectation to reduce undesirable states of information deprivation. To check the relationship between truthfulness estimation, the demand for information and epistemic curiosity, participants were administered the Litman questionnaire of Epistemic Curiosity (Litman, 2008). They also completed a questionnaire on exposition to information and four manip-check questions. Participants in the second wave of data collection received additional questions about their perceived share of fake news circulating on internet and the social medias. The objective was to check for a potential relationship between distrust toward channels of information and truthfulness estimations ([Supplementary III](#)).

### *II.2.6. Data Analysis*

To analyze our data, we ran null hypothesis statistical analyses on R, version 4.1.1 (R Core Team, 2021). Distribution comparisons have been conducted with ranksum tests. Truthfulness estimation, confidence degrees, success rates, demand for information and willingness-to-pay were analyzed with Mixed Linear Models (MLM). We used participants, order of stimuli, and year as random effects and fitted the models with the function *glmer* of the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). Post-hoc comparisons were conducted via simple slope comparisons with Bonferroni adjustments, using the *emmeans* package (Lenth, 2021). Intraclass Correlation Coefficients were computed with the *ICC* function from the *psych* package (Revelle, 2022). Mediation analysis was conducted with the *mediate* function from the *mediation* package (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014).

Bayesian analyses were conducted on RJAGS (Plummer, 2003) by modelling responses with beta-binomial or normal distributions. We set up non-informative Jeffreys priors and ran five Markov Chain Monte-Carlo (MCMC) to approximate the posterior distributions. We ran 12000 iterations, including 2000 warmup iterations. Hypotheses comparisons were conducted with R BRMS package (Bürkner, 2017, 2018, 2021) by comparing, for each dependent variable, Bayesian multilevel linear models. First, we centered and standardized each predictor to facilitate parameters interpretation. We then formulated MLM in accordance with models formulated for the statistical analyses. After checking that participants' behavior did not differ significantly across groups or sessions, we pooled all the data and kept subject as the only random effect to optimize computation time. Models were fitted using weakly informative priors (e.g., see Supplementary IV. 1).

Four MCMC were ran for each model to approximate the posterior distribution, including each 4000 iterations and a warmup of 1500 iterations. Models were compared with information criteria, specifically Widely Applicable Information Criterion (WAIC, Watanabe, 2010). The WAIC criterion provides a measure of predictive accuracy of the models for a new dataset – out of sample deviance of the model – and sanction models for their number of parameters.

We also used Bayesian stacking to average Bayesian predictive distributions. By weighting the predictions from the multiple models, based on their information criteria performance, model weights were computed that can be interpreted as the probability of the model to be the best in terms of out-of-sample prediction (Burnham & Anderson, 2002).

### III. Results

We hypothesized that individuals would not be, on average, better than chance at judging information truthfulness. Their estimation is expected to be primarily affected by the level of precision of information content. A precise information content should lure individuals into interpreting information as true, affecting accuracy in estimating truthfulness. Their confidence in their estimation is expected to weakly correlated to the accuracy of their estimation. However, we expect participants to use their confidence to primarily explain the demand for further information. This demand should increase when confidence is at the lowest points



We will first report on the participants' performance in the assessment of the brief news truthfulness and its determinants. Then, we will focus on the analysis of the participants' metacognitive abilities and their confidence-accuracy calibration. Finally, we will analyze the determinants of the demand for more or for less information.

### III.1. Performance in the Assessment of News Truthfulness

Before analyzing the extent to which participants were able to estimate the truthfulness of the brief news received, we checked that the data collected in the two waves did not differ, so that we can pool them together for the main analysis. No significant differences were detected between the first and second waves with respect to the average response time (RT) to estimate truthfulness ( $14.41 \pm 8.44$ s; *ranksum*  $p = 0.089$ ). We also estimated Bayesian beta-binomial models of the probability to successfully estimate the truthfulness of news for both datasets and both groups with Jeffreys priors via RJAGS. The delta of the posterior probabilities in the two datasets was equal to 0.002 (95% Credible Interval [-0.017, 0.009]), whereas the delta of the two groups was equal to 0.025 (95% Credible Interval [0.007, 0.043]). Therefore, there is no statistically significant difference in the success probability between the two datasets and a 2.5% difference between groups.

In line with the previous literature and our conjecture that individuals are not be better than chance at evaluating the truthfulness of true and false news, the participants' average success rate was  $51.57 \pm 6.65\%$  ([Supplementary IV.1](#)), with the lowest performance achieved for the democracy-related news (democracy:  $48.55 \pm 11.53$ ; ecology:  $52.16 \pm 11.25$ ; social justice:  $53.85 \pm 11.78$ ). A bootstrap Welch two-sample t-test (rep=10000, confidence level=.95) between participants' average successes and a randomly generated binomial distribution ( $n=258$ , size=48,  $p=.05$ ) returned non-significant ( $p=.068$ ). Accounting for truthfulness, participants were better at evaluating news that were true (true:  $64.03 \pm 11.88\%$ ) rather than false ( $39.1 \pm 12.03\%$ ) (see Figure 2A).

Interestingly, the estimation of a MLM model of repeated measures of success with main and interaction effects of truthfulness judgment and confidence value returned that confidence value was barely significant in explaining a participant's successes ( $p=0.041$ , odds-ratio = 1.002). Despite being significant, the effect of confidence was negligible. Indeed, each unit of confidence increased the probability of success by only 0.002 percent. Moreover, the

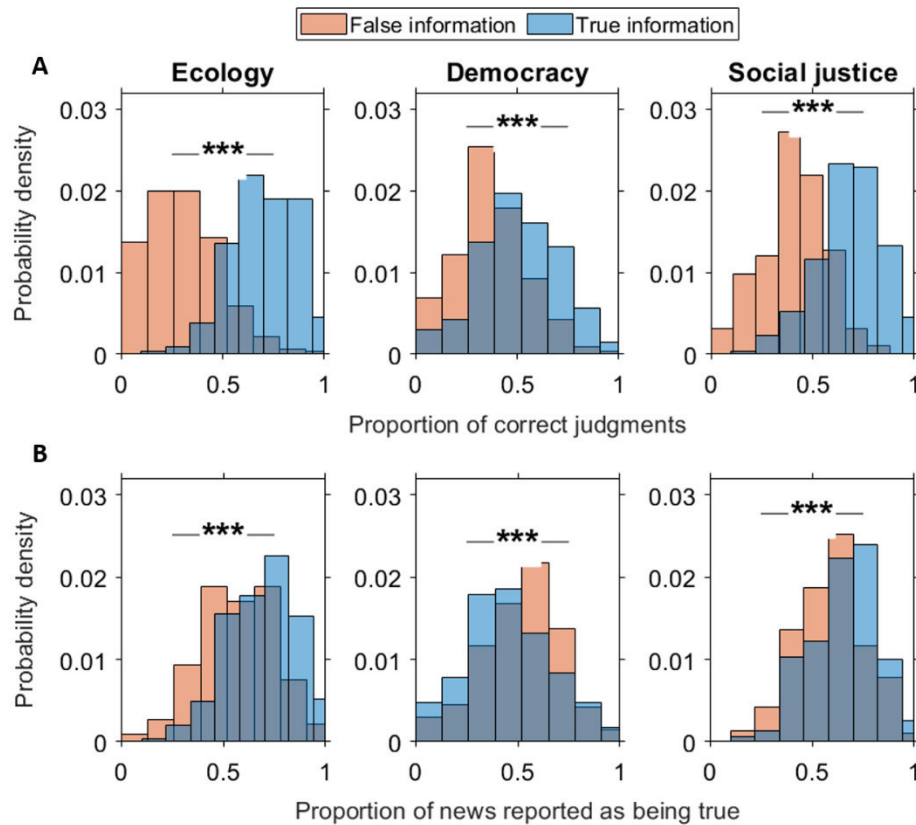
truthfulness judgment was never significant, regardless of whether we considered its main effect ( $p = 0.725$ ), or its interacting effect with confidence ( $p = 0.877$ ).

We also estimated a MLM model of success in which the independent variables include a dummy variable indicating whether the news was true or false and its interaction with the news' theme, controlling for the main effect of confidence. The truthfulness of the news returned highly significant ( $p < 0.001$ ; individual level clustering:  $SD = 0.00$ ), indicating that participants were better at predicting the truthfulness of true news than the falsity of false news. Its interaction with themes returned significant as well ( $p < 0.005$ ). Comparisons of estimated marginal means (emmeans) showed that the success rate was significantly higher for ecology-related and social justice-related news compared to democracy-related news (odds-ratio respectively at 1.16 and 1.23, all  $p < 0.005$ ) ([Supplementary IV.1](#)). In other words, discrimination between true and false information was more difficult for democracy-related information.

Such relatively higher ability to assess true news accurately can in fact be explained by a general tendency to declare information as true: participants on average declared  $59.54 \pm 10.6\%$  of news as true (true news judged as true:  $60.89 \pm 12.66$ ; false news judged as true:  $58.2 \pm 13.09$ ). We then estimated a model of success with the truthfulness judgment as independent variable, controlling for confidence and including main and interaction effects of news truthfulness with themes. The truthfulness judgment returned as highly significant ( $p < 0.001$ ). The success probability was 0.645 for true news and 0.393 for false news (true / false news odds-ratio = 2.55,  $p < 0.001$ ).

We also estimated a model of truthfulness judgment by including as independent variables the news truthfulness and its interaction with themes, controlling for confidence. Comparisons of estimated marginal means (emmeans) showed that a judgment of a news as true was significantly more likely when the news was related to ecology (prob. =  $0.649 \pm .01$ ) and social justice (prob. =  $0.612 \pm .01$ ) compared to democracy (odds-ratio at 1.61 and 1.37, respectively; all  $p < 0.001$ ) (see Figure 2B).



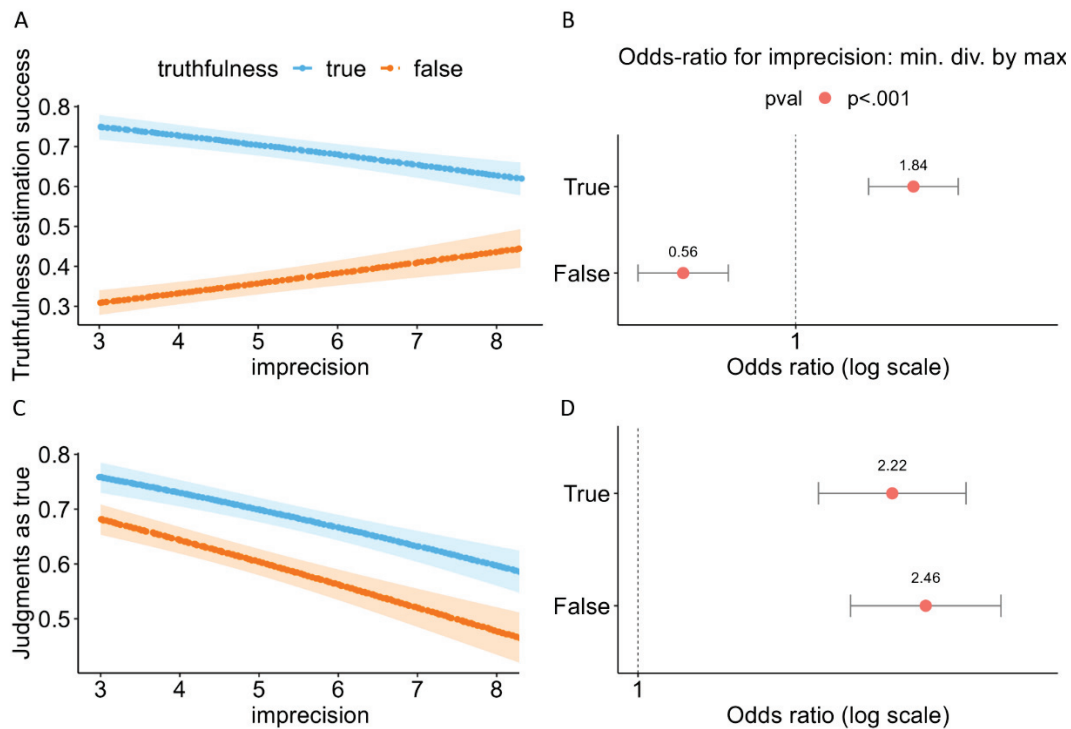


**Figure 2: Individuals were better at evaluating a news that was true than a news that was false, a result that is induced by a tendency to declare information as true.** We estimated models of prediction success and of judgment of a news as true, including among independent variables the news truthfulness (true vs. false) and its interaction with the news theme. A. Comparing emmeans between true and false news, participants consistently showed a higher success likelihood for news that was true. Pairwise comparisons indicate a lower effect size of the news truthfulness when this news was related to democracy (odds ratio: democracy = 1.72; ecology = 4.16; social justice = 3.09) (all  $p < 0.001$ ). B. Emmeans comparisons showed that participants were significantly more likely to assess a news as being true rather than false, with an exception for democracy-related stimuli (odds-ratio = 0.787) (odds-ratios for effect size of news truthfulness on ecology-related stimuli: 1.364; on social justice-related stimuli: 1.349 (all  $p < 0.001$ )).

If success in estimating a news truthfulness is affected by the tendency to declare a news as true, it might also depend on the precision of the news content, on prior knowledge, on sociodemographics characteristics, or on cognitive reflection. We tested for each of these possible determinants by designing several linear mixed-effect models, each one encompassing

the variable of interest, and we conducted post-hoc comparisons on the estimated marginal effects at the means. In each model, the binary dependent variable was the accuracy of the prediction at each trial. To test for the effect of the news content precision, we built a model with content imprecision (from 0 to 10, as assessed by the independent raters), the news truthfulness, and the interaction between the two variables. To test the role of prior knowledge, we included in the model the degree of adherence to organizations that were related to the news theme. To test for the effects of individual characteristics, we constructed a model including socio-demographics (age, sex, education and epistemic curiosity). To test for the role of cognitive reflection, we estimated a model including response time. To test for the effect of distrust toward expert sources of information, we estimated a model including the scores of distrust from the final questionnaire. In all these models, the independent variables also included the subject's confidence (the probability that the news was true or false), the news truthfulness (true vs. false), its theme, and interaction terms between the last two variables.

Neither the model with adherence to organizations, nor the model with socio-demographic characteristics returned any significant effects. The cognitive reflection model showed a positive significant effect of the response time ( $p = 0.488$ , odds-ratio = 1.003). Distrust toward the sources of information on social media significantly improved success in our expertise distrust model ( $p = 0.044$ , odds-ratio = 1.002). Finally, the content precision-model showed a highly significant effect on the success likelihood of the interaction term between the imprecision of the news content and the news truthfulness ( $p < 0.001$ , odds ratio = 0.785). Estimating the model separately for true and false news, we found that success was more likely for true news when their content imprecision was at its *minimum* (minimum/maximum, odds-ratio = 1.838) and it was more likely for false news when their content imprecision was at its *maximum* (minimum/maximum, odds-ratio = 0.558) (all  $p < 0.001$ ) (see Figure 3). Assuming that participants to the experiment would have assessed the content precision like the raters on average, we interpret this result as indicating that the news content precision was used as a marker of truthfulness, whereas a low content precision was perceived as a signal of falsity. Using the same estimation strategy with truthfulness judgments taken as dependent variable (instead of prediction success) reveals again a high significance of the news content imprecision ( $p < 0.001$ , odds-ratio = 0.847), whereas its interaction with the news truthfulness was not significant. This shows that a low content precision decreased judgments of a news as being true. Taken together, these results support the notion that content precision is taken as a signal of truthfulness of information and imprecision as a signal of falsity.



**Figure 3: The news content precision serves as a marker of truthfulness of that news, providing a useful cue during truthfulness assessment.** Interaction plots of news truthfulness and imprecision of its content for truthfulness prediction success (A) and likelihood of evaluating a news as true (B), followed by the ratio between response odds at minimum imprecision and response odds at maximum imprecision for truthfulness prediction success (C) and likelihood of assessing a news as true (D). Prediction success was more likely for true news when its content imprecision was at its minimum (minimum/maximum, odds-ratio = 1.838) and more likely for false news when imprecision was at its maximum (minimum/maximum, odds-ratio = 0.558) (all  $p < 0.001$ ). Despite a non-significant interaction with the actual truthfulness, imprecision decreases the likelihood of evaluating a news as true.

Ultimately, we looked at which model explains best participants' performances in evaluating the truthfulness of news by comparing models with Bayesian inference hypothesis testing. We included a non-informative beta response model to model randomness (Jeffrey priors,  $\alpha = 0.5$ ,  $\beta = 0.5$ ), a subject random-effect model, the response time model, the truthfulness judgment model, the news content precision model, and the model of conformity to prior knowledge. In the latter, we kept for each theme the organization with the best  $t$ -value

and modelled its interaction with the news theme. Models of socio-demographics and distrust toward the sources of information have not been kept in the Bayesian inference hypothesis testing due to the non-significance of their variables. This exercise revealed that the winning model was the model including the interaction term between the news truthfulness and its content precision (see Table 1).

*Model comparison with WAIC*

Model	$\Delta$ WAIC	$\Delta$ SE	WAIC	SE WAIC	pWAIC	Weight
<i>news content precision model</i>	0.00	0.00	16511.89	51.02	8.22	1
<i>response time model</i>	-321.62	25.51	17155.12	9.99	7.45	0.00
<i>non-informative beta response mo</i>	-323.85	25.30	17159.58	6.98	6.15	0.00
<i>conformity to prior knowledge moc</i>	-323.97	25.34	17159.84	7.46	7.44	0.00
<i>subject random-effect model</i>	-324.00	25.30	17159.91	6.68	6.66	0.00
<i>truthfulness judgment model</i>	-324.33	25.31	17160.55	7.29	7.05	0.00

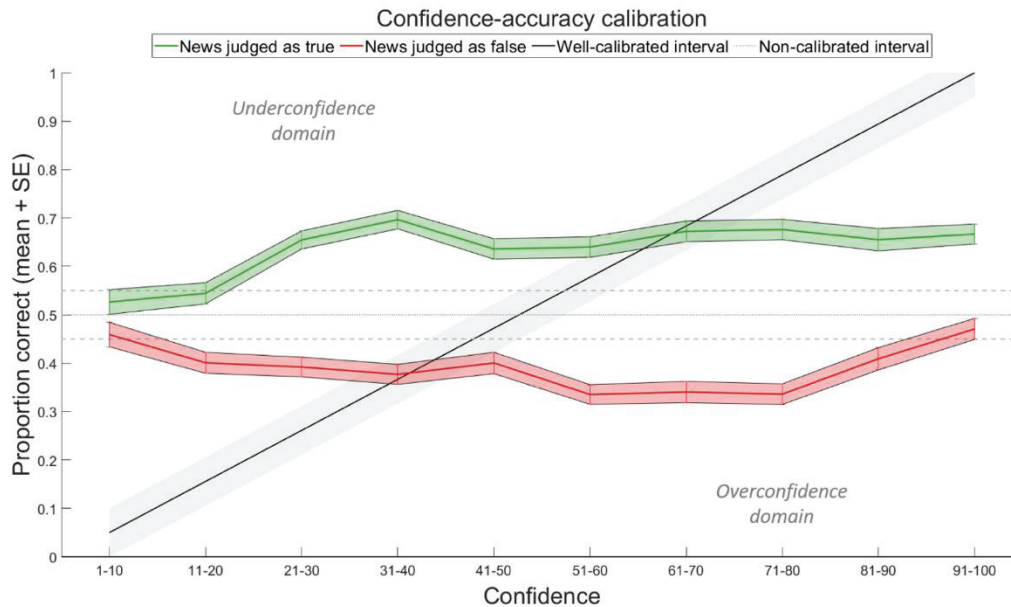
**Table 1: The news content precision model explains best participants' success in estimating the news truthfulness.** Model comparison ordered by WAIC. The best model has the lowest WAIC, showing best out-of-sample capacity, and higher weight, showing best prediction of in-sample data.

### III.2. Metacognitive Abilities

In this section, we examine the participants' calibration, that is their ability to accurately estimate the chances that news is true or false. We assume that this metacognition might influence the demand for further information that we will study in the next section. The confidence-accuracy calibration reflects, for given truthfulness judgments (the news is evaluated as true or false), the relationship between the continuous scale of confidence ([0,100]) and the binary outcome (true or false). In other words, this calibration indexes the extent to which confidence in one's judgment predicts the accuracy of this judgment. A perfect calibration is characterized by a linear confidence-accuracy function with 100% accuracy for 100% confidence, 90% accuracy for 90% confidence, etc. We sorted the individual confidence-accuracy relationships into five bins. We represented an area of well-calibrated estimation that spanned 20% (see Figure 4).

As the plot shows, participants' accuracy and their confidence were independent. This indicates that participants were neither well-calibrated, nor ill-calibrated for estimating probabilities. Figure 4 says more about this non-calibration, as values above the diagonal signal

underconfidence and values below the diagonal reveal overconfidence. Participants were on average overconfident about the accuracy of their judgment when they evaluated the news as true with a probability above 60%, but as soon as they evaluated the news as false with a probability above 40%.



**Figure 4: Participants were not calibrated for estimating the chances of news truthfulness.** The confidence-accuracy calibration plot displays the participants' accuracy in estimating probabilities that their judgment was correct, as a function of confidence degree. The plot shows that overall, the proportion of accurate truthfulness estimations did not increase nor decrease with their confidence. Well-calibrated estimated probabilities would intersect with confidence degrees in the grey area, meaning a 0-20 % confidence degree would lead to a 0-20 % accuracy in evaluating the news truthfulness.

We estimated the same models of conformity to prior knowledge, socio-demographic characteristics, response times, and we modelled the interaction effect of the news content precision with truthfulness judgments on confidence value. Consistently with the truthfulness estimates, adherence to the organizations and socio-demographic variables returned no significant effects, except sex that explained a higher confidence degree for males than females ( $p < 0.001$ , coefficient = -9.02). Response times were also significant ( $p < 0.001$ ), with each second decreasing confidence value by -0.077 unit. Most importantly, the news content precision interacted with judgment of the news as true ( $p < 0.001$ ) had a significant positive

impact on confidence. We compared confidence degrees between judgments of the news as true and judgments of the news as false for three levels of imprecision. We found that confidence was higher for judgments of the news as true than for judgments of the news as false when imprecision was at its lowest level (Cohen's  $d = 0.126$ ,  $p = 0.002$ ). When imprecision was at its highest level, confidence was higher for judgments of the news as false than for judgments of the news as true (Cohen's  $d = 0.132$ ,  $p = 0.005$ ). Finally, confidence was not significantly different between both types of judgments at the median imprecision level ( $p = 0.437$ ). These findings add evidence to the notion that the news content imprecision could be used by individuals as a signal of falsity, modulating their estimation of truthfulness. However, we acknowledge that the effect of content precision on confidence was small and did not affect the confidence-accuracy calibration.

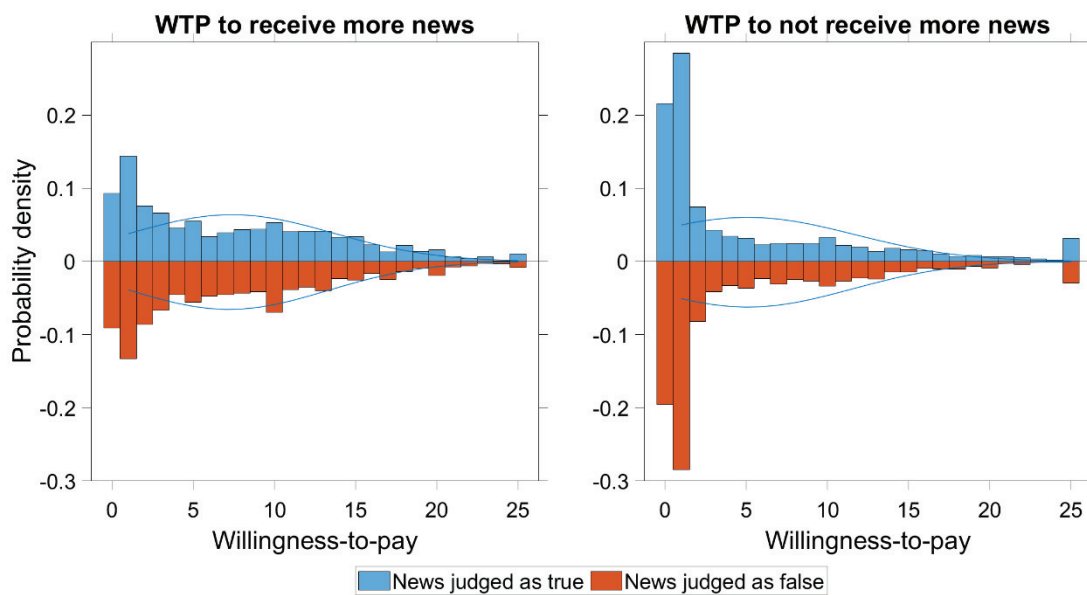
### III.3. Demand and Avoidance of Information

We can turn to the analysis of the demand for or avoidance of further information. The average decision time was  $1.56 \pm 1.12$ s for the binary decision to receive or not further information and  $2.84 \pm 0.88$ s for the Willingness-to-Pay elicitation. The average response times in the second wave of data collection were significantly smaller for both reception choices ( $1.44 \pm 1.08$ , *ranksum*  $p < .0001$ ) and WTP ( $2.78 \pm 0.91$ , *ranksum*  $p < 0.05$ ). We modelled the probability to demand more information and the associated WTP with Bayesian beta-binomial models (Jeffreys priors:  $\alpha = 0.5$ ,  $\beta = 0.5$ ) and Bayesian normal distribution models (Jeffreys priors:  $\mu = 0$ ,  $\sigma = 1$  from half-Cauchy distribution) with RJAGS, respectively. The delta of the two waves posterior probabilities of the demand for information was equal to 0.043 (95% Credible Interval [0.024, 0.062]), whereas the delta of the two groups was equal to -0.043 (95% Credible Interval [-0.061, -0.026]). Participants in the first wave were 4.3% more akin to choose to receive more information, whereas participants in the second wave were 4.3% more likely to demand further information (odds-ratios = 1.19).

Less than half of the subjects were willing to receive more information. The average frequency of choice to receive more information was  $38.03 \pm 30.55$  % (democracy:  $41.04 \pm 33.16$ ; ecology:  $43.27 \pm 33.67$ ; social justice:  $42.56 \pm 33.08$ ). We found no main nor interaction effects of news truthfulness and theme on the demand for more information. Indeed, after conducting Bayesian modelling between judgments, the delta of reception choice posterior probabilities for information judged as true and information judged as false was equal to 0.002



(95% Credible Interval [-0.023, 0.026]). Hence, truthfulness judgment did not account for the demand for further information. Unsurprisingly, with an average willingness-to-pay of  $6.02 \pm 5.41$  ECU, participants were more willing to pay to receive more information (mean:  $7.33 \pm 6.19$  ECU) than they were not to receive it (mean:  $5.07 \pm 6.51$  ECU) (see Figure 5). The delta of WTP posterior samples for choices to receive *vs.* choices not to receive more information was equal to 2.159 (95% Credible Interval [0.806, 3.489]) with Cohen's  $d$  of 0.39, meaning there was a moderate effect size of the reception choice on the WTP. This difference was lower for participants in the second wave (Cohen's  $d$  first wave = 0.48; second wave = 0.14).



**Figure 5: The willingness-to-pay was higher for receiving additional information than for avoiding it.** The willingness-to-pay (max: 25 ECU) to receive more information was centered around a higher moment (mean:  $7.33 \pm 6.19$ ) than that for avoiding information (mean:  $5.07 \pm 6.51$ ).

The demand or avoidance of further information about news judged as true or false could be motivated by the participants' confidence in their judgment, their familiarity with the theme, socio-demographic individual characteristics, and by the news' content precision. To identify these motivations, we estimated several separate MLM. The dependent variables were either the demand for more information or the WTP. Post-hoc comparisons were conducted on estimated marginal means. Each MLM encompassed the variables of interest, the random-effect variables, and controlled for the data collection wave. We modelled metacognition with a variable capturing the participant's degree of confidence and its interaction with the truthfulness

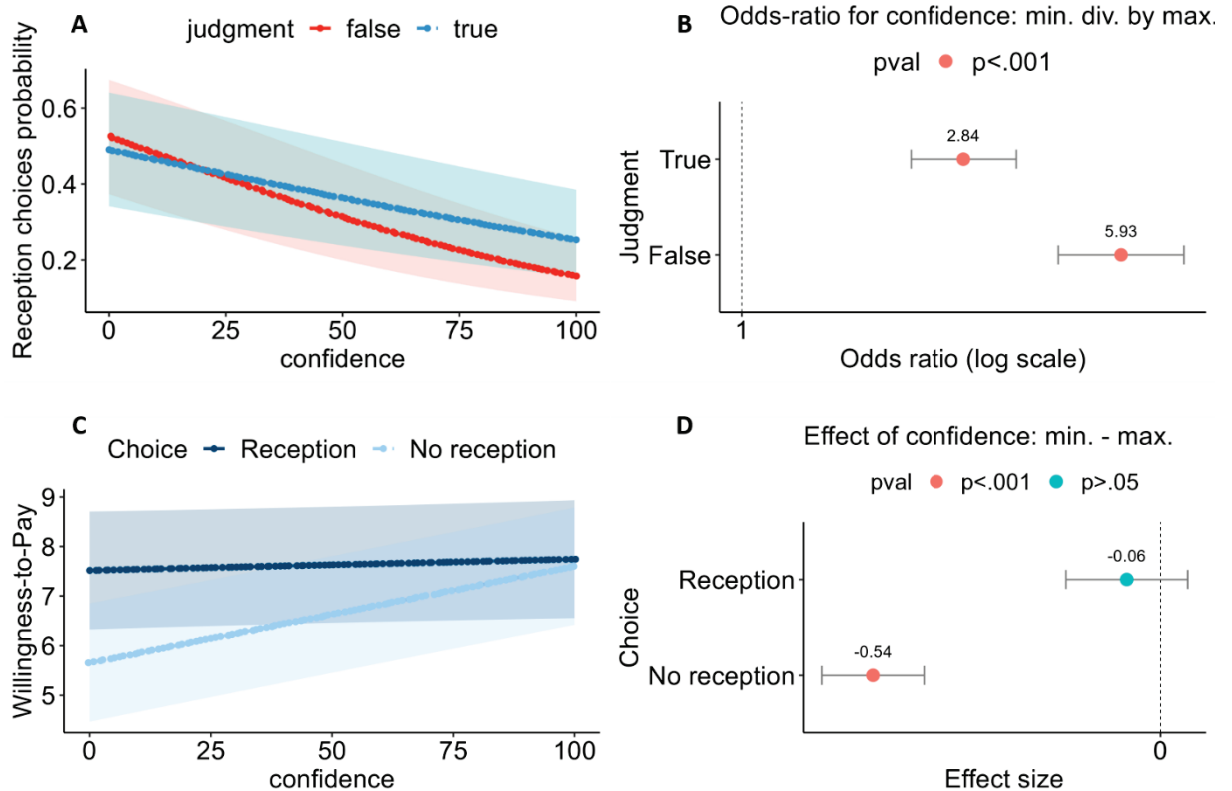
judgment. Familiarity with the theme models included the adherence to the relevant organizations. The model testing the impact of socio-demographic variables included age, sex, education and epistemic curiosity as fixed effects. To test distrust toward expert sources of information, we included in a separate model the scores of distrust given by the participants.

As predicted, confidence explains the demand for further information about the brief news ( $p < 0.001$ )<sup>5</sup>. Adding to the model an interaction term with truthfulness judgment also returned a significant negative effect of the latter variable ( $p < 0.001$ ; [Supplementary IV.2](#)). This shows that the probability of demanding more information decreases as confidence in one's judgment increases, especially when the brief news was judged as false (minimum/maximum confidence; judgment as false, odds-ratio = 5.93; judgment as true, odds-ratio = 2.84) (see Figure 6 A et B). These results are comforted by the regression analysis of the WTP, with the degree of confidence about the truthfulness of the news and its interaction with the decision to receive further information as independent variables ([Supplementary IV.2](#)). Confidence interacted with the demand for information (see Figure 6C). The effect size of confidence (minimum minus maximum confidence) for choices to receive more information was equal to -0.06 and was non-significant ( $p = 0.28$ ), whereas the effect size for choices not to receive was equal to -0.54 and significant ( $p < 0.001$ ) (see Figure 6D). To sum up, there is a significant inverse relationship between the demand for information and confidence, and this relationship is stronger for the news that participants judged as false. Moreover, participants are willing to pay more not to receive more information about what they think they already know.

---

<sup>5</sup> Models with adherence to organizations returned only one significant organization in the choices relative to ecology (Climato-Réalistes:  $p = 0.043$ ) and one in the choices relative to social justice (FEMEN:  $p < 0.001$ ). No organization has been found significant in any models of WTP. Socio-demographics and distrust models returned no significant variables.





**Figure 6: The likelihood of choosing to receive more information decreased as confidence increased.** It decreased faster for news judged as false (odds-ratio: judged as false = 5.93; judged as true = 2.84). Confidence significantly interacts with reception choices. The WTP to receive more information was not affected by the degree of confidence (effect size = -.006,  $p = 0.28$ ), whereas the WTP to avoid receiving more information increased with the degree of confidence (effect size = -0.54,  $p < 0.001$ ).

Finally, we investigated whether information uncertainty played a role in information-seeking by testing for a mediating role of confidence in the news content-precision effect on the demand or avoidance of further information. To do this, we estimated a linear model in which the decision to demand or avoid further information was the dependent variable and the interaction of content-precision with the judgment of the truthfulness of the news was the independent variable. We controlled for the wave and the mediator was the effect of confidence. Parameters were bootstrapped, using 10000 repetitions, and confidence intervals were computed using the adjusted bootstrap percentile method at 95% [2.5%, 97.5%]. The indirect mediation effect (average causal mediation effect) of confidence on the impact of content-precision on the demand for information was significant ( $p < 0.001$ ), yet the confidence interval includes 0 (estimate = -0.0018, 95% CI [-.0009, .00]). The total effect of independent and moderator variables on the demand for more information was non-significant ([Supplementary IV.2](#)).

## IV. Discussion

The present study investigated the relationship between individuals' ability to assess truthfulness in information, their metacognition, and their demand or avoidance of further information susceptible to resolve uncertainty in the news truthfulness. In a task manipulating non-ego-relevant information with cognitive utility, we expected that subjects would not be better than chance in estimating the news truthfulness. As a consequence of the uncertainty surrounding information, we expected that the news content precision would impact the assessment of its truthfulness, with imprecision signaling information falsity to individuals. In such a context, we expected that the participants' metacognition would be unreliable but would nevertheless guide the demand or avoidance of further information. Precisely, we conjectured that a low confidence about the news truthfulness assessment would increase the demand for information and the willingness-to-pay for it.

### IV.1. How Individuals Evaluate Uncertain Information

Our findings showed that participants were not better than chance at estimating brief news truthfulness, without extreme biases given the centeredness of data around a 50% success rate (mean =  $51.57 \pm 6.65\%$ ). There was a lower probability of declaring democracy-related news as true compared to other themes. Despite chance-equivalent skills, we found that participants suffered from a bias, shifting their judgments of the news toward truthfulness (mean =  $59.54 \pm 10.6\%$ ). This bias increased success in detecting true information but reduced the ability to detect false news. Such a bias has already been reported both in front of non-extravagant (Marsh, Cantor & Brashier, 2016) and extravagant content such as pseudo-profound bullshit (Pennycook et al., 2015). True news are more frequently believed than false news (Bago et al., 2020; Pennycook, Epstein, et al., 2021; Pennycook & Rand, 2019b). We cannot exclude the possibility that true information bears some features that helps detection but that we did not control for. However, this interpretation is unlikely because we identified no differences between the sets of true and false news with respect to the perception of news content precision or their capacity to make consensus. Participants' better performance at detecting true news could simply be explained by the fact that most information people encounter in their environment is true (Marsh et al., 2016; Pennycook et al., 2015).

We also found that participants' metacognitive abilities had a negligible impact on success in estimating truthfulness, as each unit of confidence increased the probability of success by only 0.002 percent ( $p = 0.041$ , odds-ratio = 1.002). Participants made inaccurate

estimations of the probability that they were correct. We computed a mean Brier Score of 0.32, a score above the chance level (which corresponds to a score of 0.25 in binomial tasks). However, Figure 4 indicated that participants' confidence-accuracy calibration was flat. They were mostly underconfident when judging information as true and overconfident when judging it as false. Interestingly, male participants' confidence values were higher than that of females ( $p < 0.001$ , coefficient = -9.02). We also found an effect of response times, with each additional second in reporting one's confidence level significantly decreasing confidence value by -0.077 unit. Such dissociation between confidence and the actual success rate has already been reported in lie detection (Bond & DePaulo, 2006; Serra-Garcia & Gneezy, 2021). People have a low ability to estimate how large is their ignorance (Ungar *et al.*, 2012).

Beyond their bias toward the truth, participants use how imprecise the content of information is as a cue to detect news falsity. Indeed, in the context of uncertainty individuals use heuristics to judge information. For example, the previous literature has reported use of familiarity (Pennycook & Rand, 2020), processing fluency, grammatical correctness, or cues such as repeated exposure (Maier, 2005). In this study, stimuli had varied levels of content precision. Our results indicate that the news content precision was used as a marker in the evaluation of the news truthfulness, with low levels of precision cuing for information falsity. The lower the precision of the news, the higher was the likelihood that the news was declared as false. The news content precision also impacted the estimation of accuracy. When precision was at its highest level, confidence was higher for the news judged as true than for the news judged as false (Cohen's  $d = 0.126$ ,  $p = 0.0018$ ).

We tested the effect of the news content precision on success in assessing the news truthfulness and alternative hypotheses by estimating linear mixed-effect models. Neither conformity between information and prior knowledge, socio-demographic characteristics, or distrust toward expert sources of information explained performance in assessing truthfulness. In contrast, response time, which could reflect cognitive reflection, significantly influenced performance. However, model comparisons showed that it was the news content precision that had the highest influence on success in the assessment of truthfulness. False news was more likely to be detected when their content precision was low, whereas true news was more likely to be detected when their content precision was high. The content precision model won the model comparison, whereas alternative models were all ranked within five WAIC units, meaning that they did not beat the subject's random-effect model.

## IV.2. Meta-Cognition Regulates the Decision to Acquire More Information

We offered participants the opportunity to receive or not to receive further information about the brief news that was susceptible to resolve uncertainty about their truthfulness and we elicited the corresponding willingness-to-pay for acquiring or avoiding this additional information. We found that the participants' average demand for further information was distributed widely and similarly across themes (mean =  $38.03 \pm 30.55$  %). The willingness-to-pay was higher for acquiring information than for avoiding it. We acknowledge that this observation is not surprising inasmuch as individuals are always able to ignore the information received for free simply by disregarding it.

What did determine the participants' demand for additional information? Although individuals held an inaccurate perception of their own knowledge, this perception was the most decisive dimension that guided information-seeking behavior in the task. Indeed, we found a highly significant and negative effect of confidence ( $p < 0.001$ ) on information-seeking. At each trial, following their binary assessment of the news truthfulness, the lower was the participants' confidence in their estimation, the more likely they were to ask for additional information about the news. This likelihood was higher for trials in which they estimated that the news was false. The same relationship was reflected in the WTP, whereby participants were willing to pay more not to receive more information about news that they thought they already knew to be false.

These results suggest that the decision to seek more information or to avoid it likely stems from the expected benefit of this additional information in terms of subsequent cognition and reduction of uncertainty about the state of the world. This interpretation is consistent with previous findings showing that individuals use uncertainty – reflected in their confidence in their judgment – in their representation of values to choose between exploration and exploitation (Boldt et al., 2019). Given the ambiguity regarding the brief news value, the choice to receive more information and the associated willingness-to-pay for it can be seen as reflecting a desire for exploration of the environment. From that respect, it is interesting to note that although the news content precision itself impacted confidence in one's assessment, it had no additional direct effect on the decision to ask for more information or to avoid it, nor on the willingness-to-pay for it. A mediation analysis showed that the news content precision impacted the likelihood of asking for more information only through its effect on confidence in one's assessment.

Overall, the analysis shows that individuals are not good at detecting false news, especially because they have a higher tendency to declare an ambiguous news as true than as false, and they use their metacognition to demand further information, although they also have a relatively weak perception of their knowledge. As a result, they are at risk not only of receiving undetected false information about the state of the world but also of exploring inefficiently their environment, in particular by paying too little for further information that would reveal the falsity of information.

### IV.3. Implications and Future Directions

In this study, we have chosen on purpose non-partisan, non-political stimuli to focus on non-ego relevant news that had mainly cognitive utility, that is, factual information that could help individuals to form more accurate beliefs about their environment or the state of the world, and that would neither threaten their identity nor affect their perception of how others would see them. The reason for that choice was to restrict as much as possible distortions in the demand for information that would result from motivated reasoning to protect one's image or identity. We cannot exclude that some news may have induced some identity concerns, for example among those related to ecology or social justice. This is why we controlled in our analysis for the feelings of the participants regarding different types of organizations. The results showed, however, that this had limited explanatory power in the demand or avoidance of information in our task. Introducing ego-relevant news in our task would be an interesting extension to measure to what extent the role of confidence in the demand for additional information would be weighted differently, as a result of possible motivated reasoning.

To explore further the nature of the cognitive processes engaged in the assessment of ambiguous news and in the demand for or avoidance of further information in this context, it would be needed to assess individuals' cognitive thinking style, attention capacities and depth of reasoning. In our study, we could investigate the participants' response time, which was unlimited in the task, as a proxy of cognitive effort. However, we found that the response time was not a prime predictor of participants' performance in predicting the truthfulness of the news. This may not be so surprising, as the previous literature has identified a correlation between cognitive reflection and disbelief in fake news or belief in true news, but mostly in contexts where the news content was most obviously implausible or obviously plausible (Pennycook & Rand, 2019b). This suggests that the role of the cognitive thinking style and in particular, the attitudes toward cognitive effort, should be investigated together with the degree

of information content precision (or ambiguity) to better understand both the individuals' performance in assessing the truthfulness of information and their willingness to invest in the acquisition of additional information about the true state of the world.

Previous studies have revealed that accuracy prompts were a good candidate to help individuals focus their attention on accuracy and achieve better performances in distinguishing true from fake news headlines (Pennycook, Bear, Collins, & Rand, 2020; Pennycook & Rand, 2021a). To a large extent, our environment was ideal to increase individuals' attention to the quality of the news since we rewarded them explicitly for providing accurate estimates of the truthfulness of the news and for their confidence that their prediction was accurate. In that sense, we certainly captured a lower bound of the distress of people facing misinformation since in everyday life, people may receive and treat news without paying the same attention to their truthfulness than in the lab. However, even in this "privileged" context, we found that the confidence-accuracy relationship was not calibrated. For example, when participants reported probabilities that the news was false above 80%, the predictions were accurate only about 45% of the time on average, but the same average success rate was observed when they reported probabilities that the news was false below 10%. We acknowledge, however, that we forced our participants to make a choice and we did not let respond that they "did not know". It is possible that, when not forced into a choice, people would formulate such a statement when confidence is not high enough, that is for confidence values around 50%.

Overall, our results showed that in the domain of non-ego relevant information with cognitive utility, the relationship between evaluation of information and demand or avoidance of further information susceptible to reduce the original ambiguity is mediated by metacognitive abilities and the calibration of probability estimation. Truthfulness judgments of news and confidence in these judgments are both best explained by a Bayesian model integrating news content precision and actual truthfulness. Low levels of confidence in one's judgment drove the demand for additional information, although the measured accuracy of these judgments is on average low. These findings characterize the computations required for the evaluation and search of information in an increasingly ambiguous world and demonstrate a key role of metacognitive monitoring in false news evaluation. This suggests that policy interventions should not only target the attention paid to the quality of the sources of information but also the formation of metacognitive skills.

# Chapter II

## Neurocomputational processes of inferring others' preferences for information and fake news

V. Guigon (ISCMJ/GATE), R. Philippe (ISCMJ), J. Benistant and M. C. Villeval (GATE)  
and J-C. Dreher (ISCMJ)

CNRS, Neuroeconomics lab, ISCMJ and CNRS, Groupe d'Analyse et de Théorie  
Economique (GATE) and Université Claude Bernard Lyon 1, Lyon, France

### Abstract

To reduce the devastating effects of misinformation online, recent interventions have proposed to fact check uncertain information or to encourage selective sharing of news to specific recipients. These two types of strategies are based, respectively, upon increasing our beliefs about the veracity of the news and upon inferences we make regarding potential receivers' preferences for acquiring more information. Yet, when deciding whether to share or additional information with others or not, it is unknown how the brain processes uncertain information according to one's own confidence of the reliability of information and depending on our beliefs concerning the preferences of the receivers. Here, we investigated how the brain integrates these two forms of beliefs (i.e., one's beliefs about a given piece of information, and one's inferences regarding the receivers' preferences). Participants in a scanner had to decide whether or not to match the preferences for information of receivers, while beliefs about these preferences were manipulated. When no information about receivers' preferences was provided, participants were more likely to send extra information, especially when their own beliefs about the truth of the news was low. In contrast, participants sent less information when they were informed that the receivers' preferences were at large social distance from the content of the news. This behavior was explained by a Bayesian model in which individuals weighed



beliefs about information truthfulness, beliefs about a population's preference for information and beliefs about a target agent's preferences. Neural signals computing increasing beliefs in the truthfulness of the item associated with increased activity in the Ventral Medial Prefrontal Cortex, Striatum and Dorsolateral Prefrontal Cortex. Second-order beliefs were associated with a signal increase in Temporo-Parietal Junction. These results highlight the brain systems engaged in the use of beliefs to infer preferences of others for acquiring extra information about news.

## I. Introduction

The growth of social media and messaging platforms has increased interest in understanding how misinformation spreads (Jackson, Malladi, & McAdams, 2022). Misinformation consists of information that can be false, inaccurate or misleading. Contrary to disinformation, it does not need to be created deliberately to mislead. The spread of misinformation in social media has devastating consequences at the societal level, increasing polarization and resistance to climate action and vaccines (Barreto et al., 2021; Rapp & Salovich, 2018; Tsfati et al., 2020; Van Bavel, Rathje, Harris, Robertson, & Sternisko, 2021). A piece of fake news that appears on a social network typically makes some claim about the world that is factually wrong. However, because there is often a high degree of uncertainty about the truth of posted information, users may decide to spread news items across their social network. A number of strategies have been proposed to fight misinformation, including fact-checking, capping the number of others to whom messages can be forwarded (Jackson et al., 2022), censorship or encouraging more selective sharing by individuals (Guess, Barberá, Munzert, & Yang, 2021; Traberg, Roozenbeek, & van der Linden, 2022), or directing attention to accuracy (Kozyreva, Lewandowsky, & Hertwig, 2020; Pennycook, Epstein, et al., 2021; Pennycook, McPhetres, Zhang, Lu, & Rand, 2020). Recent behavioral experiments and computational models of social learning have demonstrated that these approaches, or a combination of them, can be effective to reduce the spread of misinformation (Bak-Coleman et al., 2022; Globig, Holtz, & Sharot, 2022; Pennycook & Rand, 2021b).

The neurocomputational mechanisms underlying the decision whether to send fact-checking information to others remain unknown. Understanding these brain mechanisms has the potential to clarify why people share uncertain information even when they do not trust it (Pennycook, Epstein, et al., 2021; Ren, Dimant, & Schweitzer, 2021; Tappin et al., 2021). The

decision to send or not additional supporting or debunking information to friends or unknown people in one's social network depends upon two types of beliefs: the degree of confidence we have in evaluating the veracity of the information (i.e., metacognitive ability), reflecting first order belief about the news; and the beliefs we have about the willingness of the receiver of the information to effectively use this extra information. This latter type of beliefs can be formed based on observed proximity between the receiver's opinion and the content of the information (i.e., second order beliefs). Here we aimed to understand how the brain integrates our own beliefs about uncertain information and the beliefs we have about the willingness of receivers to effectively receive the news, when deciding whether or not to share extra information with others.

To investigate the neurocomputational mechanisms of inferring others' preferences for receiving additional supportive or debunking information, we designed a new model-based fMRI study using a modified coordination game experiment. Prior to the experiment, participants rated various organizations related to 3 topics (i.e., ecology, social justice and democracy). Participants in the scanner were then incentivized to match the choices of Receivers from a previous behavioral experiment in which Receivers indicated their willingness to receive more information about a news item. That is, participants decided whether to send extra information about the news to the Receiver if they believed she wished to receive the extra information, or not to send it if they thought the receiver preferred not to receive such information. The receiver was drawn from a pool of 20 possible receivers. Importantly, in the informative 'Cue' condition, participants were given information to reduce their uncertainty about receiver's willingness to receive extra information. This cue indicated the receivers' social proximity to one of the organizations, whose theme was congruent with the news. In the control condition, participants were provided with no information about the receivers. At the end of each trial, participants received feedback about whether they matched or mismatched the Receiver's choice.

We examined the neural signals that compute the decision to share with others extra information that can reduce uncertainty about news. To do so, we tested a number of computational models developed to account for coordination of actions during dyadic social interactions to achieve a common goal. To be successful, the Sender (participant) has to integrate her own beliefs concerning the truthfulness of the information and her beliefs about the Receiver's preferences to receive extra information (i.e. the two players are rewarded only

if they both decide to send and receive, or not send and not receive the extra information). We compared a model of Bayesian inferences against alternative models, including Q-learning and heuristic models. We found that inferences of others' preferences for information were best predicted by a Bayesian model in which participants weigh beliefs about information truthfulness, beliefs about the simulated population preference and beliefs about the Receiver's preferences for information.

We modelled three alternative classes of decision-making that rely on different assumptions about beliefs. A first class makes the assumption that participants have fixed beliefs about Receivers: a random-biased model served as a baseline, representing a heuristic of random choice with a bias corresponding to prior beliefs about the Receivers' preferences; a Win-Stay/Lose-Switch model defines a heuristic based on the principle that an agent keeps performing an action until it is not rewarded anymore. Here it represented the heuristic of maintaining a belief about a Receiver's decisions until they receive contradictory evidence. A second class of reinforcement learning model makes the assumption that participants update beliefs about a Receiver's preferences over time via the history of decisions (Sutton & Barto, 2018). A third class of expected utility models makes the assumption that participants perform a simple non-probabilistic link between beliefs and a Receiver's preferences.

We predicted increased activity in information valuation brain areas associated with beliefs about others' preference given information truthfulness. Finally, we predicted that beliefs about others' preferences given a participants' beliefs about others' social distance from a given set of opinions will be associated with signal increases in TPJ and DMPFC. The latter hypothesis was supported by findings that increased activity in DMPFC and TPJ have been reported in subjects when considering sharing news articles with others compared to selecting them for themselves (Baek et al., 2017).

## II. Methods

### II.1. Participants

35 right-handed participants participated in this study. 3 participants were excluded due to non-exploitable fMRI data. Artifacts appeared on one participant's MRI images at level of upper neck; behavioral data acquisition failed for one participant during the scanning session; one participant forgot to bring their glasses and was not satisfied with the vision goggles the

CERMEP provided. 32 participants in total (F=16, mean  $\pm$  SD age:  $24.64 \pm 4.02$ , max: 35, min: 18) completed the study with exploitable fMRI data.

Participants were all recruited by advertisements on the Institut des Sciences Cognitives Marc Jeannerod facebook page. The study was approved by an ethics committee (CPP) under CNRS promotion (INSB) in accordance with 1° or 2° of article L1121-1 of the public health code and not concerning a product mentioned in article L. 5311-1 of the public health code. The study was approved by CNRS delegation for data protection (GDPR). Subjects reported no history of psychiatric or neurological disorders, and no current use of any psychoactive medications. Written informed consent was obtained from all subjects prior to participation according to the guidelines of the local ethics committee.

Upon completing task instructions and prior to the experiment, participants completed a task comprehension questionnaire ([Supplementary V](#)).

Participants all answered a set of questionnaires at the end of the scanner session. To assess social value orientation and depth of reasoning, questionnaires included 11-20 level-k reasoning (Arad & Rubinstein, 2012), Murphy et al. slider SVO (Murphy, Ackermann, & Handgraaf, 2011). Also, we included 4 experiment-evaluation questions ([Supplementary V](#)). After the experiment, participants were debriefed about true and fake news and provided access to debunk webpages and news and received payment (70€).

## II.2. Experimental design

### II.2.1. Organizations and training

Prior to the task, participants rated 12 political organizations related to ecology, democracy or social justice, on 6 dimensions ([Supplementary II.1](#)). After reading instructions and before going to the fMRI scanner, participants answered a comprehension questionnaire then performed a 20 stimuli training task on a computer that was identical to that performed in the scanner. Stimuli, themes and B participants were different from the scanner task. Subjects were informed that B participants answers were randomly chosen.

### *II.2.2. Task*

Each trial in the scanner started with a fixation cross (Figure 1). Briefly after that, participants viewed on their screen a short news item that was either true or false. They were asked to rate the percentage probability that it was true or false by indicating their degree of confidence as described previously (Karni, 2009). Probability elicitation provided a quantitative proxy for elicitation of their personal beliefs. To do so, participants pushed a response button either to the left (False) or to the right (True). Each push towards a particular direction incremented their degree of confidence by 10%. Participants had 15 seconds to read the news and provide their answer. The 15 s time window was determined as a trade-off between reducing fMRI acquisition time and providing enough time for participants to respond. Participants from the Receiving News task online behavioral study had taken  $14.46 \pm 8.23$ s to provide an answer. Participants validated their assessment by pushing a third button, and a blue feedback confirmed they had answered in time. Participants then viewed a third screen displaying the news stimulus and asking them to choose between sending more information to the Receiver or not sending any information. Participants had 8 seconds to provide their answer, in which case they received a blue feedback to confirm they answered in time. Then on a fourth screen, participants were given feedback concerning whether the Receiver had accepted the supplementary information. The Participant won if their decision to send matched the Receiver's decision to receive or if their decision not to send matched the Receiver's decision to refuse the supplementary information. Success was indicated by a green feedback and failure by a red feedback. When participants failed to answer during the time limit, they were shown a warning that the stimulus would be presented again at the end of the task. The task was divided in two fixed conditions: a first block of 48 stimuli under the Control condition and a second block of 48 stimuli in the Informative Cue condition.

In the Control condition, participants had to decide whether to send the supplementary information but were provided no information about the Receiver. In the Informative Cue condition, we provided an estimated social distance between the Receiver and a political organization congruent with the stimulus. Cues were a proxy for elicitation of the others' beliefs. Cues provided with democracy-related news were social distance to France FREXIT (an organization campaigning for France to leave the European Community), cues provided with ecology-related news were social distance to NIPCC and Cues provided with social justice-related news were social distance to FEMEN, a radical feminist organization.

Participants were told they played with data from Receivers that participated in a previous experiment. Behavioral data from the previous experiment was implemented in the task to account for Receivers receiving choices and cues provided to the participants. The Receivers' choice dataset was chosen to maximize both the directionality of the Receivers' desirability for news and for the correlation between the Receivers' choices to accept further information and their social distances to the organizations indicated by the informative cues ([Supplementary II.2](#)).

In this behavioral protocol, receivers had evaluated 12 organizations similar to those the senders evaluated. Then, in the task, receivers were elicited beliefs regarding true and fake news before declaring their Willingness-to-Pay (Becker, DeGroot, and Marschak, 1964) to receive or avoid receiving more information about the news. We calculated Receiver's social distance to each organization by aggregating their responses to six dimensions with which they evaluated each organization. The aggregated score was normalized on a scale from 0 to 100. The closer the score was to 0, the lower the social distance between that Receiver and the organization.

### *II.2.3. Receivers task*

As participants in the current study, prior to the task, Receivers rated 12 political organizations related to ecology, democracy or social justice, on 6 dimensions. During the task, at each trial after a fixation cross, Receivers viewed a true or false brief news. This news was similar to the present task. They were elicited to declare in their opinion the number of chances out of 100 that this brief was true or false. They were then asked in a third screen to choose between receiving more information or not receiving any more information. In a fourth screen, they had to choose how much they were willing to pay between 0 and 25 Experimental Currency Units (ECUs) to have their reception decision implemented.

In the case Receivers chose to receive more information, they were eligible for receiving a debunk article investigating the content of the news stimulus in details. Debunk articles were taken from French fake news debunk platforms *Les Décodeurs du Monde*, *AFP Factcheck* and *Libération Checknews* from the period 2017-2020.

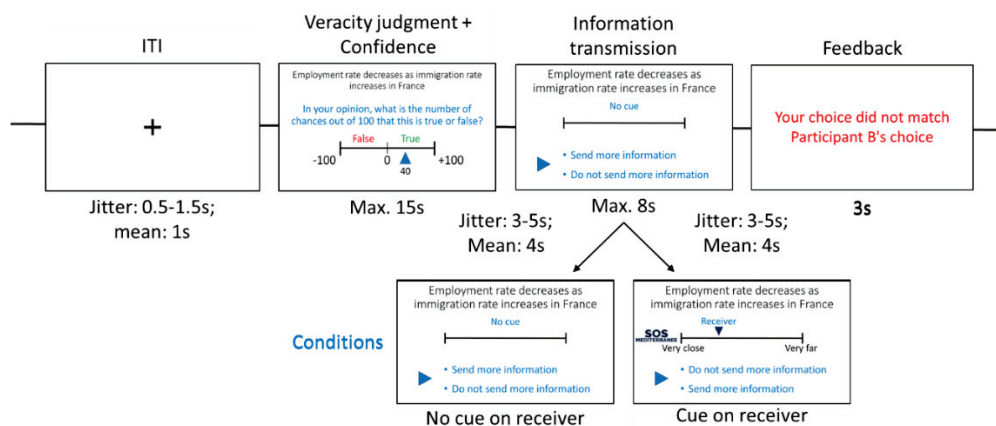
### *II.2.4. Incentivization*

Participants were told to believe that their payoff depended on their successful trials. Participants received a show-up fee of 60€. They were told they could earn more depending on

their success in the task. At that at the end of the experiment, sixteen trials would be randomly drawn to reward their truthfulness judgment. For each successful trial, one robot out of 100 robots would be randomly drawn. Each robot had an accuracy level from 1 to 100 which was equal to their probability of providing the right answer. Participants were aware that if the randomly drawn robot had an accuracy level superior to their their own degree of confidence, we would take the robot's answer into account. Else, we would take the participant's answer into account. For each trial with a successful truthfulness estimation, participants were told they would be rewarded with 50 Experimental Currency Units (ECUs). A non-successful trial would be rewarded 0 ECUs. They were told that an additional sixteen trials would be randomly drawn to reward their sending decisions. For each successful trial, participants would earn 50 ECUs and non-successful trial would be rewarded 0 ECUs.

### II.3. Organizations

We choose 12 organizations, categorized as either democracy-related, ecology-related or social justice-related ([Supplementary III](#)). We provided to participants a 1000 characters ( $\pm 20\%$ ) long description for each organization, taken from the organizations websites. Descriptions were construed minimalizing manipulation of website content. Subjects were asked to rate each organization on a scale of 1 to 10 with respect to 6 dimensions, measuring familiarity, the degree to which they agree with the organization and its values and proximity of its values to those of the participant and their family and friends ([Supplementary II.1](#)). We computed the distance of the participants to each organization on the basis of their mean score over the 6 dimensions normalized on a 100 points scale.





**Figure 1:** Description of a trial of the task: after a fixation cross, participants saw a news that were true or fake and were incentivized to provide a probability that the news was true or false. If their response was provided in time, they were incentivized to choose on the next screen between sending more information to the receiver or send nothing. If their response was provided in time, a green feedback appeared if the sending choice matched the receiver's desirability while a red feedback marked a failure to do so. Failures to provide a response in time led to a screen notifying the trial would be presented again at the end of the task. If the sending choice matched the receiver's desirability, the trial was worth 50 Experimental Currency Units (ECUs). Otherwise, it was worth 0 ECUs. A trial with a correct truthfulness estimation was worth 50 ECUs while an incorrect evaluation was worth 0 ECUs.

#### II.4. Stimuli

Our process of selecting stimuli was in accordance with Pennycook and colleagues' practical guide to doing behavioral research on fake news and misinformation (Pennycook, Binnendyk, et al., 2021).

First, we designed a set of 210 true and fake news items (114 fake news; 96 true news). Maximum length was 140 characters (spaces included). We chose to manipulate non-ego-relevant information. We restricted as much as possible each news item's short term impact on participants' daily decision making or affect elicitation. We focused on brief news describing events or situations concerning ecology, social justice and democracy – three key domains that gained momentum as hot topics but didn't directly concern the participants' health or individual personal situation. For instance, we avoided news related to COVID-19. We sought to choose information with mainly cognitive content concerning concepts meaningful to participants that had the capacity to alter their understanding of the state of the world. Some news items were directly taken from the French fake news debunk platform *Les Décodeurs du Monde*, *AFP Factcheck* and *Libération Checknews* from the period 2017-2020. We produced several abridging or taking excerpts from longer items on those platforms.

Next, we ran an online pre-test on Testable.org on the news items intended to serve as stimuli in both the participants' task and the Receiver's task (previous behavioral study). We controlled them on four dimensions: ambiguity, desirability, capacity to achieve consensus ('consensuality') and theme. We sought to identify 96 counterbalanced news items

after the pre-test, with maximum agreement on each dimension and no statistical difference in ambiguity and consensuality between the set of True news and Fake news. Fifty-five independent raters (F=33, M=22; mean  $\pm$  SD age=26.2  $\pm$  4.78), who were rewarded a fixed amount of 7\$, evaluated the 210 True and Fake news. 5 groups of 11 French speaking raters each evaluated a set of 42 news items from the 210. For each theme we found the 16 True and 16 Fake news items with the highest agreement concerning theme. We computed the Intraclass Correlation Coefficient (ICC3k) for the measures of content precision, consensuality and desirability. Average fixed rate correlation coefficient was .514 for ambiguity (CI [.356, .65]), .8 for consensuality (CI [.74, .86]) and .6 for desirability (CI [.465, .71]) (all  $p < .001$ ). We found no difference between True (mean  $\pm$  SD= 5.53  $\pm$  1.24) and Fake news (mean  $\pm$  SD =5.17  $\pm$  1.25) ambiguity distributions (*ranksum*,  $p=.09$ ); True (mean  $\pm$  SD= 6.24  $\pm$  1.57) and Fake news (mean  $\pm$  SD= 6.55  $\pm$  1.55) consensuality distributions (*ranksum*,  $p=.39$ ); or True (mean  $\pm$  SD=6.34  $\pm$  1.18) and Fake news (mean  $\pm$  SD=6.33  $\pm$  1.24) desirability consensuality distributions (*ranksum*,  $p=.92$ ).

## II.5. fMRI acquisition

Imaging was conducted on a Siemens 3 Tesla Magnetom Prisma scanner at the CERMEP – Imagerie du vivant (Bron, France) using EPI BOLD sequences and T1 sequences at high resolution. We performed single-shot EPI, TR / TE = 1600/30, flip angle 75°, multiband acquisition (accelerator factor of 2), in an ascending interleaved manner with 52 slices interlaced 2.40 mm thickness, FOV = 210 mm. We used iPAT mode with an accelerator factor of 2 and GRAPPA method reconstruction. Total length of functional images acquisition depended on participants' response time and number of missing trials. First acquisition was made after stabilization of the signal (3 TR) based on manufacturer's standard automatic 3D-shim procedure. A whole-brain high-resolution T1-weighted structural scan (voxel size: 0.9\*0.9\*0.9mm<sup>3</sup>) was also acquired for each subject. We co-registered them with their mean EPI images and averaged across subjects to permit anatomical localization of functional activations at the group level. Field map scans were acquired to obtain magnetization values that were used to correct for field inhomogeneity. We acquired cardiac and respiratory traces generated by CMRR MB sequences. Subjects performed 2 runs, one run per condition, in which each stimulus was seen once. Stimuli were fixed in each condition, order of conditions was fixed and order of stimuli within each condition was randomized for each subject to optimize further signal deconvolution. All subjects were given written instructions prior to the scanning

session, familiarized themselves with the cognitive task during test trials before scanning and general instructions were orally repeated before scanning.

## II.6. Behavioural analysis

Questionnaire responses, behavioral measures and response times were collected for each of the 32 subjects during the fMRI sessions. Statistical analyses were run on R version 4.1.1 (R Core Team, 2021). Null hypothesis significance testing of distribution comparisons was conducted with Wilcoxon tests. Confidence values, truthfulness judgments, sending choices, successes and reaction times were analyzed with Mixed Linear Models (MLMs). We used subjects, order of stimuli throughout the task and conditions as random effects and fitted the models with the function `glmer` of the package `lme4` (Bates et al., 2015). Reported continuous  $\beta$  are standardized coefficients (centered mean, one standard deviation unit). Post-hoc comparisons were conducted via simple slopes comparisons with Bonferroni's adjustment method, using the *emmeans* package (Lenth, 2021). Intraclass Correlation Coefficients were computed with ICC function from the *psych* package (Revelle, 2022). Bayesian statistical analyses were conducted on RJAGS (Plummer, 2003) by modelling responses with beta-binomial or normal distributions. We set up non-informative priors and ran five Markov Chain Monte-Carlo (MCMC) to approximate the posterior distributions. Priors for Bayesian beta-binomial models were  $\alpha=0.5$ ,  $\beta=0.5$ . Confidence values were modelled with Bayesian normal distribution models and priors were  $\mu$  following a Normal distribution and  $\sigma=1$  from half-Cauchy distribution. We ran 12.000 iterations including 2.000 warmup iterations. Following approximation, we computed deltas between posterior distributions.

## II.7. Modelling

Consider a round of the 'Sending extra information' task. At the time of the sending decision, the participant faces one Receiver drawn from the pool of 20 possible Receivers. The participant has to choose an action (to send or not to send extra information) that matches the action she estimates the Receiver has chosen in the previous behavioral experiment (to receive or not to receive more information for a specific news item). In the case the participant chooses the action that matches the Receiver's action, the trial is worth 50 monetary experimental units

(MEUs). In the opposite case, the trial is worth 0 MEUs. This situation can be modelled with a single parameter  $\theta$ , the probability of the Receiver's action being to receive more information, expressed as a binomial distribution. Receivers' actions are unknown until feedback; hence the participant must estimate  $\theta$ .

Our main hypothesis concerning estimation of others' preferences for information regards the generative model participants are the more likely to use. We assume participants represent their beliefs about others' preferences by estimating a probability distribution over  $\theta$ . To test our hypothesis, we compared the results of computational models split into 4 families. The first family is based on heuristics and includes a Random Biased decision model and a Win-Stay/Lose-Switch. The second family includes with Reinforcement Learning. The third class includes non-Bayesian utility models. The fourth family is Bayesian.

### II.7.1. Heuristics models

The Random Biased model generates via a softmax function the probability of selecting action  $a$  at round  $t$  with a constant rate  $x$  and  $\tau$  a free parameter representing prior volatility.  $b$  represents a bias towards selection of  $a$ .

Equation 1.

$$p_t = \frac{1}{1 + e^{-x*\tau + b}}$$

The Win-Stay/Lose-Switch model is a heuristic learning strategy. It generates the probability of selecting action  $a$  at round  $t$  based on the outcome of the previous round. If it resulted in a success the participant *stays*, selecting the same action  $a$ ; otherwise, the participant *switches*, selecting the alternate. We implement it with two Q-values,  $V^{stay} = 1$  and  $V^{switch} = -1$ . A softmax function then gives the probability of choosing action  $a_t = a_{t-1}$ .

Equation 2.

$$p^{stay} = s(V^{stay} - V^{switch})$$

Priors on mean and standard deviation values of free parameters for both models were set at  $\mu = 1$  and  $\sigma = 1$ .

### II.7.2. Reinforcement Learning models

A utility table  $U(a^{self}, a^{other})$  represents the participant's payoff given the participant's action  $a^{self}$  and the receiver's action  $a^{other}$ . In our experiment both players make a binary choice which can take as values 0 or 1. In case both actions match, the reward takes a single value  $V = 50$  MEUs in case both actions match and  $V = 0$  otherwise. Classically, agents are considered trying to maximize their expected value  $V = E[U(a^{self}, a^{other})]$  via a softmax function as decision rule. In our case, we assume the decision rule is the following:

Equation 3.

$$p(a^{self} = 1) = s\left(\frac{V}{\beta}\right)$$

$p(a^{self} = 1)$  is the probability of the participant choosing action  $a = 1$ . The sigmoid function  $s$  converts the value into the probability distribution.  $\beta$  its free parameter controls for the magnitude of behavioural noise.

One of the simplest strategies is to choose on each trial the action that in the recent past gave the most reward. This strategy is referred to as reinforcement learning (RL) and approximates the optimal solution for many different types of decision problem in nonstrategic contexts (Hampton et al., 2008).

We represent learning with Reinforcement Learning (RL), updating value with the Rescola-Wagner rule. We assume  $V$  is described by a learning rate  $\eta$  defined by a sigmoid and a reward prediction error defined as the difference between the participant's value for action  $a$  at trial  $t-1$  and the actual reward. The learning rate expresses the priority of the most recent events over all events.

Equations 4.

$$V_t = V_{t-1} + \eta * \delta_{t-1}$$

$$\delta_{t-1} = R_t - V_t^a$$

### II.7.3. Utility models

To express the value  $V$  of action  $a$  based on truthfulness estimation, we link negatively  $V^i$  to the participant's  $p^{truth}$  that the information is true or false. We assume  $p^{truth}$  is described

by  $\frac{confidence}{100}$ . The relationship between confidence levels and first-order judgments such as value have been shown to be quadratic (Lebreton, Abitbol, Daunizeau, & Pessiglione, 2015). Given its three terms, the equation has the advantage of expressing a linear relationship if the first term  $a = 0$ . Accordingly, we describe  $V^i$  with the following term:

Equation 5

$$V^i = a(1 - p_i^{truth})^2 + b(1 - p_i^{truth})$$

Value  $V$  of action  $a$  based on the cue is represented with a negative link between  $V^i$  and the participant's  $p^{distance}$ . We assume  $p^{distance}$  is described by  $\frac{cue\ value}{100}$ , representing the estimated distance between the receiver and the round's organization. For control condition, we forced  $p^{distance}$  at 1.

Equation 6

$$V^i = a(1 - p_i^{distance})$$

Priors on values of free parameters were set at  $\mu = 1$  and  $\sigma = 1$ .

#### *II.7.4. Bayesian models*

We assume participants represent their beliefs about others' preferences by estimating a probability distribution over  $\theta$ . Using this probability distribution, the participant can calculate which action to choose. The decision process takes the distribution over parameter  $\theta$  and converts it into an action probability (to send or not to send) by integrating the probability distribution. Each participant holds a prior belief about  $\theta$  at each round. Over successive rounds, based on receivers' actions, this belief is updated. Beliefs update can be computed with Bayes' rule, inverting the probabilistic relationship between  $\theta$  and others' action.

To account for prior beliefs, the beta distribution is appropriate for both estimating  $\theta$  and conjugating with the binomial distribution. It is determined by two hyper-parameters  $\alpha$  and  $\beta$  that represent probability over  $\theta$  and described in Equations 7.

Equations 7

$$\theta \sim Beta(\alpha, \beta)$$

$$Beta(\alpha, \beta): P(x | \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

The parameter  $\alpha$  represents belief that the Receiver's action is to receive while  $\beta$  represents beliefs that the action is not to receive. The larger  $\alpha$  over  $\beta$  is, the larger the probability favour of choosing to send. The larger the sum of values  $\alpha$  and  $\beta$  is, the larger the sample size is, meaning the larger the evidence is in favour of the probability.

### **Beliefs integration**

We hypothesized that estimating an individual's preferences for information relies on weighting and updating beliefs about information, beliefs about population desirability and beliefs about the individual's beliefs. A first Bayesian model represents this hypothesis.

For a given round, deciding on the action is determined by observable parameters. Parameters we hypothesized the participant includes to estimate  $\theta$  are 1) the participant's estimation of information truthfulness, 2) the history of receivers' past actions and 3) in the Informative Cue condition, the value provided as informative cue that states the distance between the Receiver and the organization. We included the participant's bias regarding the estimated Receivers' probability to choose to receive as a fourth parameter. Here, we assumed that for each parameter the participant has a prior belief  $Beta(\alpha_i, \beta_i)$  over  $\theta$ . Each prior belief is differently weighted by the participant. We considered that the probability distribution over  $\theta$  given the history of Receivers' past actions is the unique updated probability distribution. Probability distribution over  $\theta$  given the truthfulness estimation of the round's information and that over  $\theta$  given the Informative Cue concerning the round's Receiver are considered independent from one round to another.

The truthfulness estimation beta distribution expresses the probability of the Receiver choosing an action  $a$  given the probability  $p_{truth}$  that the information is true or false. The relationship between confidence levels and first-order judgments being quadratic, we describe prior beliefs over  $\theta$  given  $p_{truth}$  in round  $t$  with a quadratic function of confidence (Equations 8). Hyper-parameters  $\alpha$  and  $\beta$  index the negative relationship between confidence and the probability to send.  $\theta$  is the estimated probability that the Receiver chooses  $a = \text{'to receive'}$  where  $\alpha^{truth}$  and  $\beta^{truth}$  are the probability of the information being either true or false. Considering the differences in send probabilities between truthfulness judgments,  $\theta$  for both truthfulness judgments would be different (Equations 8). We exponentialized hyper-parameters to constraint them in the positive domain.



Equations 8

$$\theta^{truth} \sim Beta(\alpha_1^{truth}, \beta_1^{truth})$$

$$\alpha_{1_t}^{truth} = a_t(1 - p_{truth}^2) + b_t(1 - p_{truth})$$

$$\beta_{1_t}^{truth} = a_t(p_{truth}^2) + b_t p_{truth}$$

The learning beta distribution tracks the probability of the population of Receivers within the pool choosing an action  $a$ . The prior beliefs are dynamically updated with the choice history. The value of parameters  $\alpha$  and  $\beta$  can be understood as representing evidence in favor of an action, with increasing value representing increasing evidence. Consequently, we describe learning by an increase in the posterior probability parameters  $\alpha$  or  $\beta$  given the receiver's action. In the case the receiver chose at the previous round to receive,  $\alpha$  is increased by 1. In the alternate,  $\beta$  is increased by 1. A discount rate  $\eta$  is added as a sigmoid so it is constrained in the interval  $[0,1]$  (Equations 9). Assuming participants learn different Receivers' probability of choosing action  $a$  depending on the information estimated truthfulness, we included two branches to track Receivers action history for both estimated truthfulness values.

Equations 9

$$\theta^{learning} \sim Beta(\alpha_2, \beta_2)$$

$$\alpha_{2_t} = \eta * \alpha_{2_{t-1}} + action_{receiver}$$

$$\beta_{2_t} = \eta * \beta_{2_{t-1}} + (1 - action_{receiver})$$

$$\eta = \frac{1}{1 + e^{-x}}$$

The Cue beta distribution expresses the probability of the Receiver choosing an action  $a$  given their distance to the organization, where the organization is associated in theme with that of the information. The negative relationship between the social distance and send probability is described here by a linear function. We assume that for a same value of Informative Cue, participants may interpret differently the distance between the Receiver and the organization. To account for participants' interpretation of the Informative Cue,  $\alpha$  and  $\beta$  have been modelled with different constants (Equations 10). Hyper-parameters are then exponentialized to constraint them in the positive domain.

Equations 10

$$\theta^{distance} \sim Beta(\alpha_3, \beta_3)$$

$$\alpha_{3_t} = a_t(1 - cue) + b$$

$$\beta_{3_t} = a_t cue + c$$

The participant's belief about the receivers' reception bias is described by constant terms, differentiated for  $\alpha$  and  $\beta$  and exponentialized:

Equations 11

$$\theta^{bias} \sim Beta(\alpha_4, \beta_4)$$

$$\alpha_{4_t} = a$$

$$\beta_{4_t} = b$$

### Setting the priors

At the first round, a participant begins the game with prior beliefs that may be based on prior life experience or expectations about the behavior of receivers. Our Bayesian models represent such prior belief for each parameter with Jeffreys priors  $\mu = 0.5$  and  $\sigma = 3$  on all hyper-parameters  $\alpha$  and  $\beta$ . These priors represent non-information prior states of knowledge over  $\theta$ , limiting the constraints on the posterior. They are updated through convergence to reach the estimated participant's prior knowledge. Priors on participants' free parameter were set at  $\mu = 1$  and  $\sigma = 3$ .

## II.8. Model analysis

To test for estimation of others' preferences for information, we compared 8 models between conditions. Two Bayesian model accounted for our hypothesis. The first Bayesian full model (Bayes) included Beta distributions over  $\theta$  as is whereas the second Bayesian Non-learning model (Bayes\_noRL) tested the assumption that participants do not track the probability of the population of receivers. Instead it considers participants have a fixed model of the population's probability to choose action  $a$  with no discount over time. A full utility model (U\_full) accounted for our hypothesis without the assumption that participants estimate a probability distribution over  $\theta$ . It is described by a RL term, a second term for truthfulness estimation and a third for the cue, each term in  $V^i$  weighted by a free parameter. We included three control utility models, a simple RL model (U\_RL), a truthfulness estimation utility model

(U\_truth) and a RL model with a second term for truthfulness estimation (U\_RL\_truth). Finally, we included two heuristics models to account for strategies that do not model the receiver's action, a Random Biased model (RB) and a Win-Stay/Lose-Switch model (WSLS).

We performed Bayesian Model Selection (BMS) with the VBA toolbox (Variational Bayesian Analysis) in a random effect analysis. We relied on the free energy for model evidence and used Protected Exceedance Probability measurements (PEP) to select the most frequent model in our sample.

## II.9. fMRI analysis

### II.9.1. Data pre-processing

Image analysis was performed using SPM12 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK, [fil.ion.ucl.ac.uk/spm/software/spm12/](http://fil.ion.ucl.ac.uk/spm/software/spm12/)) in MATLAB R2019b (Mathworks, Inc.). Time-series images were registered in a 3D space to minimize any effect that could result from participant head-motion. Once DICOMs were imported, functional scans were realigned to the first volume, corrected for slice timing and unwarped to correct for geometric distortions. Inhomogeneous distortions-related correction maps were created using the phase of non-EPI gradient echo images measured at two echo times (5.20 ms for the first echo and 7.66 ms for the second). Finally, in order to perform group and individual comparisons, they were co-registered with structural maps and spatially normalized into the standard Montreal Neurological Institute (MNI) atlas space (152 spaces) using the DARTEL procedure implemented in SPM12 (Ashburner, 2007; Ashburner & Friston, 2009), resulting in a voxel size of : 2.625 x 2.625 x 2.64 mm for the statistical analysis. Then images were spatially smoothed with an 8 mm isotropic full-width at half-maximum (FWHM) Gaussian kernel using standard procedures in SPM12. After preprocessing, we scrubbed volumes that could include movement artefacts with ArtRepair v5b3 (<https://cibsr.stanford.edu/tools/human-brain-project/artrepair-software.html>) (Mazaika, Hoeft, Glover, & Reiss, 2009).

### II.9.2. Whole-brain analysis

We created voxel-wise statistical parametric maps from functional data using the general linear model. An fMRI design matrix was created for each subject with truthfulness

estimation, send choices and feedback modeled as separate categorical regressors. Regressors were based on individual response periods. Due to our event-related design and interdependence between onsets, we convolved them with the canonical hemodynamic response function and its temporal and dispersion derivatives. Response times, truthfulness judgment, theme, action and distance between the participant and the Receiver with respect to the trial's organization were modeled as separate parametric modulators. Realignment parameters from data preprocessing were included as regressors of no-interest to correct for neural activity associated with subject motion.

To investigate model-based brain activations, we applied Bayesian model shape parameters of each  $\theta$  distribution as parametric modulators of interest. Shape parameters describe beta distributions. The mode ( $\frac{\alpha-1}{\alpha+\beta-2}$ ) represents the peak of the beta distribution, representing its most probable value and can be interpreted as the belief the participant has about the reception probability. The higher it is, the more probable it is that the participant chooses action  $a$ . The variance ( $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ ) is the second moment, centered on the mean, and a measure of statistical dispersion. It can be interpreted as the participants' confidence in their beliefs. The higher it is, the more beliefs are dispersed, hence the less the participant is confident. We computed both moments and investigated how they modulate brain activity. We controlled for the second moment (variance if the parametric modulator of interest is mode; mean/mode otherwise).

Each participant's design matrix was then regressed against fMRI data to yield parameter estimates for each participant. Statistical analyses were performed at individual and group level. Contrasts of interest were performed on individual subject data in the first level analysis. For the second level group analysis, contrast images from all subjects were included in a second level random effect analysis. Data were submitted to an event-related general linear model analysis, fitting a reference hemodynamic response function to an impulse response function for each event in the observed time series data.

We examined activation differences in brain regions by applying one sample t-tests at onsets of interest within conditions at the HRF + derivatives level. We used two sample t-tests to investigate direct comparisons between conditions at onset of interest. When a parametric modulator of interest was included, t-tests were operated on HRF + derivatives in correlation with the parametric modulator. Additionally, we computed F-tests for main effects of onsets of

interest. Statistical parametric brain maps were generated from which we displayed the T-value in signal intensity of each voxel. Statistical maps were superimposed on a 152-subject, T1-averaged image from SPM12. We report results for voxels that survive a  $p < .001$  uncorrected voxel-level threshold and  $p < .05$  FWE-corrected cluster-level threshold. To plot Statistical Parametric Maps of different contrasts on a same average image, we applied an additional cluster-size thresholding at 100 voxels.

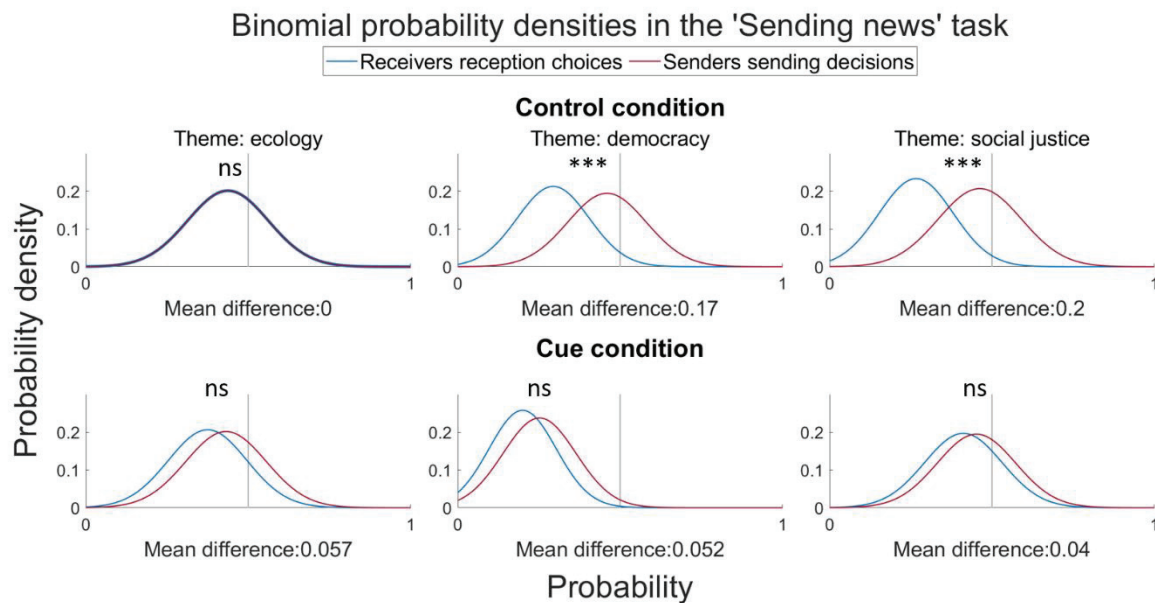
### III. Results

#### III.1. Behavior of others' preferences estimation

In the Cue condition participants received a value indicating the distance of the trial's Receiver to the relevant organization at the beginning of each trial. This organization was congruent with the news item theme. The higher the score, the greater the social distance between the individual and the organization. The mean social distance between participants and the Receivers' social distances from the congruent organizations, for each theme, was  $28.26 \pm 17.76$  (French FREXIT:  $31.28 + 15.82$ ; NIPCC:  $22.81 + 17.01$ ; FEMEN:  $30.73 + 18.88$ ). The mean score for all cues was  $43.35 \pm 21.51$ , with  $64.73 \pm 10.94$  points for the democracy-related organization (FREXIT),  $30.94 \pm 23.02$  points for the ecology-related organization (NIPCC) and  $36.18 \pm 10.96$  points for the social justice-related organization (FEMEN).

Participants had a window of 8 s to make their sending decision, their mean response time (RT) was  $2.92 \pm 1.25$ s (control RT=2.72, cue RT=3.13). Receivers' frequency of choice to receive was 33.3% for both conditions. In the Control condition, participants mean sending frequency was  $45.31\% \pm 12.95\%$  and mean success rate was  $50.2\% \pm 8.16\%$ . In the Cue condition, participants reduced their sending frequency to  $38.28\% \pm 13.27$  which increased their success rate to  $57.68\% \pm 8.65$  (Figure 2). An MLM analysis was performed to investigate Receivers' reception choices by modelling reception probability with Receivers' behavior associated with the reception choice. Independent variables were main and interaction effects of Receivers' truthfulness estimation and the social distance to the trial's organization (i.e., the Cue value provided to the participants). Only the social distance had a significant effect on reception probability (odds-ratio = .511, 95% CI [.297, .879],  $p < .05$ ) which indicates that the more Receivers were close to the organizations, the more they chose to receive. The effect of

confidence was at the margin of statistical significance (odds-ratio=.979, 95% CI [.958, 1.002],  $p=.069$ ).



**Figure 2: Participants' send probability was closer to the Receivers' reception probability in the Cue condition.** We plotted probability densities of send decisions against reception choices, for each theme in both conditions. In the Control condition, distributions between reception decisions and send decisions are significantly different for two themes. In the Cue condition, distributions are no longer different in any theme, indicating the increased ability of participants to match Receivers' decisions. Wilcoxon signed rank tests between probability distributions. ns  $p > .05$ , \*\*\* $p < .001$ .

To investigate the effects of stimulus variables, qualitative variables and truthfulness estimation on participants' choices to send, we modelled the send probability in Mixed Linear Models (MLM) with subject, condition and time (discrete value increasing at each trial) as random effects. First we looked into effects of stimuli by modelling the main and interaction effect of truthfulness with theme. We found no significant effect of variables (all  $p > .05$ ) therefore we didn't include them as controls in the following models.

To account for the use of truthfulness estimation in send choices, we modelled send probability with main and interaction effects of truthfulness judgment with confidence and controlled for the effect of conditions. Results show participants' probability to send decreased by 22.7% as confidence increased by one standard deviation (odds-ratio = .723, 95% CI [.634, .824],  $p < .001$ ) (Figure 3). This main effect interacted with truthfulness judgment ( $p = .04$ ).

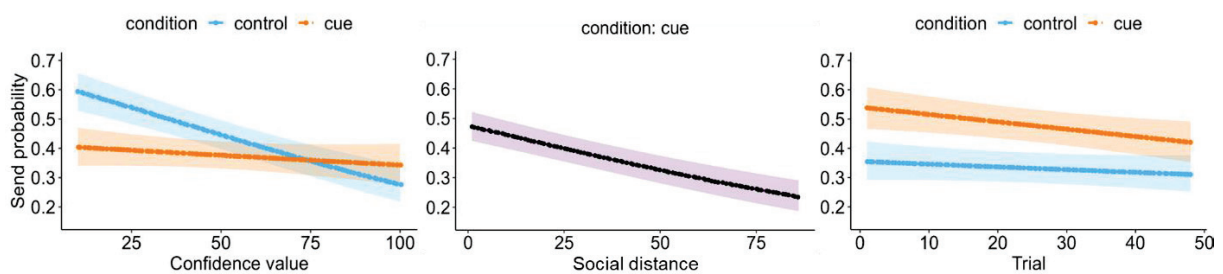
The comparison of estimated marginal means (emmeans), suggests that participants' increased their probability to send more, when their confidence was lower for news items they judged to be fake (odds-ratio = 3.27,  $p < .001$ ) than for news items they judged to be true (odds-ratio = 1.78,  $p = .0033$ ). Condition, (Control vs. Cue), had a highly significant effect ( $p < .001$ ) in this model. The interaction of condition with confidence, controlled for judgment and variable main effects, showed that the interaction effect was very highly significant (all  $p < .001$ ). Emmeans comparisons showed that the increase in the probability to send, when confidence decreased, was very highly significant for Control condition (odds-ratio=3.83,  $p < .001$ ) but not significant in the Cue condition ( $p = .225$ ). Thus, participants in the Control condition use their confidence in estimation of the new items truthfulness to decide when to send, and their use of confidence differs according to whether they believe the news item is true or fake. However, in the Cue participants do not use confidence anymore.

Next, we analysed whether participants used the distance values provided in the Cue condition. In an MLM controlled for judgment and confidence values, the send probability decreased significantly by 44.2% when cue value increased by one standard deviation to indicate the Receiver was at a greater social distance from the organization (odds-ratio=.558, 95% CI [.485, .641],  $p < .001$ , AIC=1936.8) (Figure 3). We ran the same model with the difference between the participant's distance from the organization as cue and the Receiver's distance from the organization (i.e., the cue value). This variable was also significant, however, the effect-size was lower and the AIC higher compared to the when the cue value alone was used (odds-ratio = 1.295, 95% CI [1.136, 1.48],  $p < .001$ , AIC=1994.4). This odds-ratio can be interpreted as the probability to send additional information increases by 29.5% when the difference in social distance between the participant and the organization and the Receiver and the organization increases by one standard deviation and the Receiver is closer to the organization than the participant.

Finally, as proxy for learning, we investigated the interactive effect of condition with trial number in each condition. This interaction was significant (odds-ratio=.939, 95% CI [.893, .988],  $p < .001$ ), and showed a decreasing probability to send additional information for each standard deviation sized increase in trial number. The decrease in probability to send was steeper in the Cue condition (odds-ratio=.654, 95% CI [.465, .921]) than in Control condition (odds-ratio=.809, 95% CI [.682, .959]) (Figure 3). Most importantly, testing for the interaction of trial number with the truthfulness judgment, we find that participants differentially learn



receivers' choices given their judgment. The relationship between the send probability and the increase in learning by one standard deviation is null for information judged as false (odds-ratio=1) and negative for information judged as true (odds-ratio=.692, 95% CI [.498, .962]).



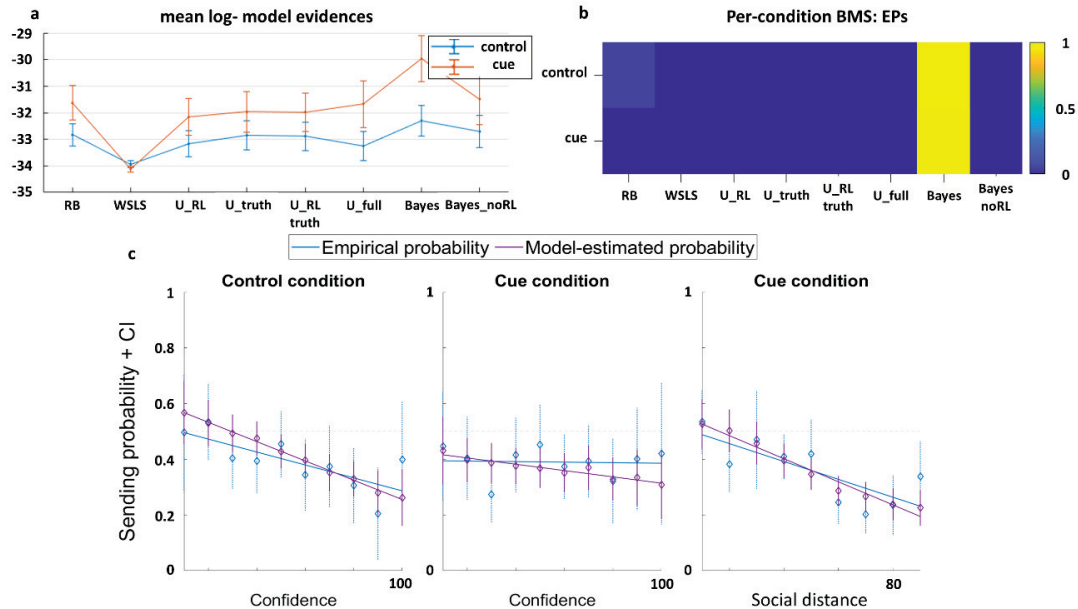
**Figure 3: Linear fixed effects of confidence value, social distance and trial on sending probabilities for Control and Cue conditions ( $\beta$  from Mixed Linear models, not standardized).**

We performed an MLM with socio-demographics (i.e., age, sex, diploma), the response in the 11-20 game and the participant's Social Values Orientation (SVO) scores. We found that send probability significantly decreased as SVO increased (odds-ratio = .981, 95% CI [.967, .996],  $p < .05$ ) and the sending probability of female subjects was higher than males (odds-ratio = 1.55, 95% CI [1.18, 2.03],  $p < .005$ ).

Overall, MLMs indicate that participants' beliefs regarding Receivers' reception choices are modulated by the participant's use of their own beliefs regarding information truthfulness. During the task the participants learned to decrease their rate of sending additional information and adapted to the cue provided concerning the Receiver's social distance to the relevant organization ([Supplementary VII.2](#)).

### III.2. Modelling estimation of others' preferences for information

Mean log-evidences show that the model that fits data best in both conditions is the Bayesian full model (Bayes) (Free energy: control=-32.2, cue=-29.97, Figure 4a). The protected exceedance probabilities (PEP) indicate that it explains decisions in the task better than the other models (between conditions PEP = .99, Figure 4b). Average classification accuracies for this model were equal to .64 in the Control condition and .71 in the Cue condition.



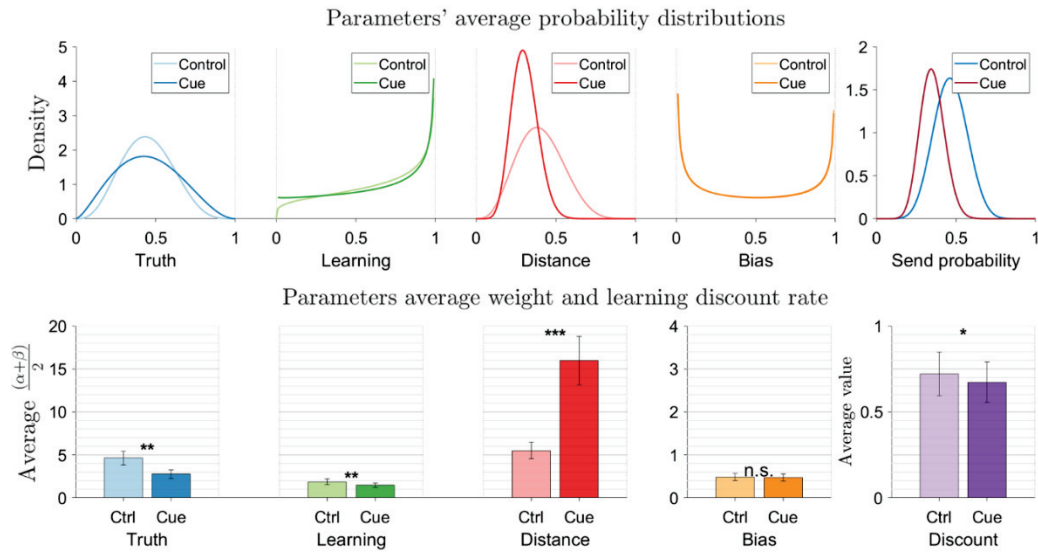
**Figure 4: Bayesian Model Selection shows the Bayesian model with learning fits the data best.** Models optimized with Free energy criterion. **a)** The mean log- evidences show that data is fitted best by the Bayesian full model (Bayes) in both conditions. **b)** Exceedance Probabilities (EP) were allocated between the Random Biased model (RB) and the Bayes model in the Control condition, as displayed in the Bayesian models selection (BMS). The Bayes model overtook the EP in the Cue condition. **c)** Average empirical vs model-estimated probability to send for each confidence value or social distance value, aggregated in bins. Bayes model estimated probability to emit action  $a$  was close to the empirical probability. Models: Random Biased (RB), Win-Stay/Lose-Switch (WSLS), Reinforcement Learning Utility model (U\_RL), truthfulness estimation Utility model (U\_truth), RL with truthfulness estimation (U\_RL\_truth), Utility full model (U\_full) and Bayesian with fixed population probability (Bayes\_noRL). Between conditions Protected Exceedance Probability = .99.

Participant's behavior in the task can be understood by interpreting properties of the Bayesian model distributions. In our study, beta distribution hyper-parameters  $\alpha$  and  $\beta$  represent the participants' beliefs about  $\theta$ , the probability of the Receiver to choose action  $a =$

to receive. We described  $\theta$  with four parameters, each with its own set of hyper-parameters. From these hyper-parameters the posterior distribution of beliefs about  $\theta$  can be determined. Figure 5 (a) illustrates how our different  $\theta$  parameters multiply to retrieve the participant's send probability, as well as the differences between conditions. Probability distributions can be investigated via their properties, such as central tendency and statistical dispersion. These properties give insights into participants' beliefs during the task.

In our study, the mean ( $\frac{\alpha}{\alpha+\beta}$ ) of a beta distribution represents the participant's mean weighted belief in the interval  $[0,1]$ . The variance ( $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ ) represents the spread around the mean weighted belief. It can be interpreted as the participant's confidence in their own beliefs. The mode ( $\frac{\alpha-1}{\alpha+\beta-2}$ ) is another measure of central tendency. It can be unaligned with the mean of the beta distribution in the case of a skewed distribution and corresponds to where the distribution reaches its tallest height. It can be interpreted as the most probable belief in the distribution. The contribution of each beta distribution to the final probability is given by the weight of this distribution ( $\frac{\alpha+\beta}{2}$ ). The weight tells which parameter is most powerful in determining decisions to send or not. The learning discount rate  $\eta$  is another value of interest. It represents the extent to which past observations weigh in the learning process compared to new observations.

After fitting the Bayesian model to participants' behavior, we examined the mean, variance and weight for  $\Theta$  parameters:  $\theta^{truth}$ ,  $\theta^{learning}$  and  $\theta^{distance}$ . The value of the weight across subjects was situated between .5 and 16 (Figure 5 b), suggesting that some priors could be considered as un-informative on their own, however contributed, nevertheless, to the estimation of  $\theta$ .



**Figure 5: The winning model takes beliefs about the Receivers action  $a =$  to receive for four parameters,  $\theta^{truth}$ ,  $\theta^{learning}$ ,  $\theta^{distance}$  and  $\theta^{bias}$ . Between conditions participants update their beliefs, resulting in a decrease in the probability to send. By integrating the distributions of all four parameters, at each trial the model outputs the participant's predicted action  $a$ . a) The figure displays the mean beta distributions for all four parameters and the differences in shape between conditions. b) Average weights of beta distributions. Hyper-parameters  $\alpha$  and  $\beta$  show that participants in Cue condition rely significantly less on the truthfulness estimation and learning whereas they rely significantly more on the social distance (all  $p < .01$ ). The decrease in discount rate indicates that in cue condition participants allocate more weight to the most recent observations ( $p < .05$ ).**

In the Control condition, we found an average mean across subjects equal to  $.48 \pm .17$  for  $\theta^{truth}$ ,  $.5 \pm .07$  for  $\theta^{learning}$  and  $.43 \pm .2$  for  $\theta^{distance}$  (Supplementary VII). In this condition we fixed the participants prior for  $\theta^{distance}$  at  $Beta(0.5,0.5)$ , which is equivalent to  $p = .5$ . The VBA method approximates participants' true priors, explaining their shift away from the starting value at  $.5$ . Hence, participants had prior beliefs on Receivers' distance to organizations (average:  $\alpha = 4.46$ ,  $\beta = 6.64$ ). Variability was great among participants regarding Receivers' action  $a$  given the truthfulness estimation. Average variance across subjects was equal to  $.03 \pm .03$  for  $\theta^{truth}$ ,  $.06 \pm .02$  for  $\theta^{learning}$  and  $.02 \pm .01$  for  $\theta^{distance}$ . These values inform us that participants were less confident in their beliefs about Receivers' action  $a$  when  $a$  was estimated with the learning. Its value remains small when compared to that of  $\theta^{bias}$  variance, equal to  $13 \pm .003$ . Average weight across subjects was equal to  $4.63 \pm 4.01$  for  $\theta^{truth}$ ,

$1.86 \pm 0.8$  for  $\theta^{learning}$  and  $5.49 \pm 2.55$  for  $\theta^{distance}$ . In other words, the parameters that influenced the participants' estimation of  $a$  the most were the truthfulness estimation and their prior beliefs concerning the Receivers' distance to the organizations. We found no significant difference between the two ( $p = .25$ ).

In the Cue condition, the mean value across participants was barely higher for  $\theta^{learning}$  ( $m = .52 \pm .06$ ,  $p = .49$ ) and significantly lower for  $\theta^{distance}$  ( $m = .32 \pm .18$ ,  $p = .032$ ) ([Supplementary VII](#)). Participants' mean beliefs increased towards action  $a = not\ to\ receive$ , with great variability between participants. This is corroborated by mean weights across participants. Weights significantly decreased for  $\theta^{truth}$  ( $m = 2.77 \pm 2.96$ ,  $p = .007$ ) and  $\theta^{learning}$  ( $m = 1.48 \pm 0.4$ ,  $p = .007$ ) but significantly increased by three-fold ( $m = 16.95 \pm 18.26$ ,  $p < .001$ ) for  $\theta^{distance}$ . The contribution of the truthfulness estimation and learning to the participants' probability to send, decreased. The contribution of the distance was nearly multiplied by three (average:  $\alpha = 9.35$ ,  $\beta = 21.72$ ). There was considerable variability in the extent to which participants updated their use of their beliefs in the Cue condition. Lastly, the mean variance across participants was significantly higher ( $m = .05 \pm .02$ ,  $p = .012$ ) for  $\theta^{truth}$  and significantly lower ( $m = .01 \pm .01$  for  $\theta^{distance}$ ,  $p = .008$ ), which means that they were now more confident in their beliefs when  $a$  was estimated with the social distance of the Receiver to the organisation. When provided with -the social distance Cue for the Receivers, participants relied less on their private information and more on information about others, resulting in an improved approximation of the receivers' preferences.

With mean hyper-parameters ( $\alpha = .48$ ,  $\beta = .5$ ) and ( $\alpha = .45$ ,  $\beta = .49$ ) for Control and Cue condition biases, we interpret the bias as non-informative compared to the other parameters. The differences in the discount rate  $\eta$  show that subjects granted significantly more weight to the most recent observations, in the Cue condition ( $\eta = .67 \pm .07$ ) compared with the Control condition ( $\eta = .72 \pm .11$ ) ( $p = .016$ ). We analysed the contribution of learning to fitting the participants' data. Statistics were in favor of the Bayes model compared to the Bayes noRL in the Control condition (Bayes mode, PEP = .863, estimated frequency of model = .966) and Cue condition (Bayes model, PEP = .992, estimated frequency = .977). Despite lower hyper-parameter values for  $\theta^{learning}$  in regard of other parameters,  $\theta^{learning}$  significantly contributed to estimate the participants' behavior.

To understand what effect the most probable belief given by each parameter had on the probability to send, we ran a mixed linear model with each parameter's mode as independent variable. We controlled for response times, information theme and truthfulness judgment and included subject random effect. In the Control condition, the mode of  $\theta^{truth}$  significantly increased the probability to send (odds-ratio = 3.22, 95% CI [2.37, 4.38],  $p < .001$ ), the mode of  $\theta^{learning}$  didn't affect the probability ( $p = .096$ ) and most surprisingly, the probability to send significantly increases with the mode of  $\theta^{distance}$  (odds-ratio = 6.53, 95% CI [4.27, 10],  $p < .001$ ). The intercept was significant as well (odds-ratio = .16, 95% CI [.1, .25],  $p < .001$ ). In Cue condition, the intercept remained significant (odds-ratio = .23, 95% CI [.13, .41],  $p < .001$ ). Most importantly, the mode of  $\theta^{distance}$  was the only remaining significant parameter (odds-ratio = 4.29, 95% CI [1.66, 11.08],  $p < .001$ ), confirming the results on the measures of central tendency and weight of parameters. Finally, we introduced in the MLM the mode of  $\theta$  beta distribution, with hyper-parameters from all four parameters. This mode represents the most probable belief given  $\theta^{truth}$ ,  $\theta^{learning}$ ,  $\theta^{distance}$  and  $\theta^{bias}$ . We find its effect highly significant in both Control and cue Condition (all  $p < .001$ ; [Supplementary VI.4](#)).

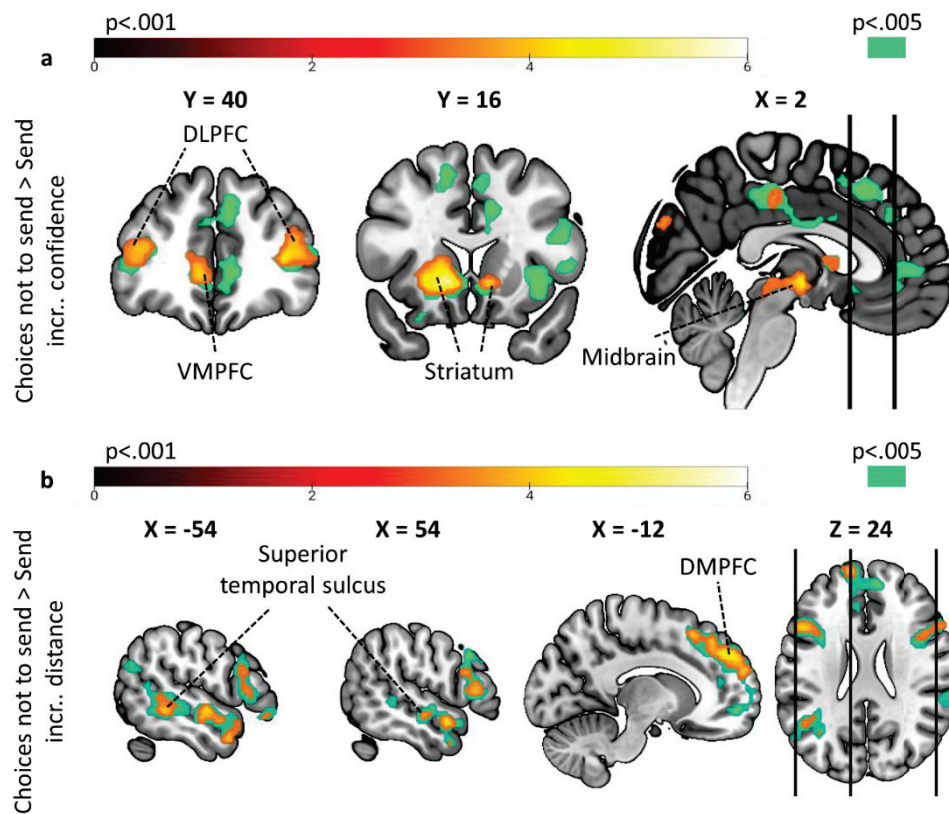
### III.3. Neural bases of others' preferences estimation

We analysed the fMRI data at the time participants made the decision to send to identify the neurocomputational bases of inferring others' preferences for non-instrumental information. The behavioral and modelling results supported a computational account of participants' beliefs about others' preferences based on beliefs about information truthfulness and beliefs about the Receiver's beliefs. The first component, the information truthfulness, is linked to the participant's confidence in their evaluation of the information. The higher the confidence, the higher the belief that the information is true (or false). The second component, the belief about the others' beliefs, is elicited by the social distance from the receiver to the organization. The probability to send additional information decreased when either of these two components increased. We expected activation in brain areas involved in information valuation on one hand, and social processing on the other. We integrated in our fMRI Generalized Linear Models (GLMs) the decisions to send and the decisions not to send as well as parametric modulators. We included as control variables to the GLMs the participants' response times, truthfulness judgments and the information theme.

First, we examined brain activations underlying sending decisions modulated by participants' behavioral variables. We looked at the modulating effect of confidence on brain activity in the Control condition and that of the Receivers' social distance in the Cue condition. Participants learnt during the task to lower their sending rates. Accordingly, we compared brain activity that anti-correlated with the variable of interest during decisions not to send to that during decisions to send. Then we investigated brain activity that anti-correlated with beliefs about Receivers' preferences to receive extra information. We looked at the modulating effect of beliefs about information truthfulness and beliefs about the Receiver's beliefs during decisions not to send. We represented participants' beliefs about Receivers with the mode of Bayesian parameters. The mode of a beta distribution represents the most probable belief about Receivers' preferences. As such, the mode for  $\theta^{eval}$  represents the most frequent probability to send, given the probability that the information is true. In the case the mode  $Mo$  is above .5, the participant's most probable belief about the Receiver is a preference to choose to receive extra information with probability =  $Mo$ . Therefore, the mode indexes both the probability to send and the behavioral parameter that underlies it.

The first whole-brain analysis showed that an increased confidence in truthfulness estimation during decisions not to send, compared with those to send, elicited activity in the information valuation brain areas. Elicited activity included midbrain, striatum, ventromedial prefrontal cortex (vmPFC) and dorsolateral prefrontal cortex (dlPFC) (all p voxel-level <.001 uncorrected, cluster-level <.05, Figure 6a). Whereas, the cue, during decisions not to send, relative to decisions to send, increased activity in brain areas involved in social processing. We found corresponding heightened activity in the superior temporal sulcus and DMPFC (all p voxel-level <.001 uncorrected, cluster-level <.05, Figure 6b) ([Supplementary VIII](#)).

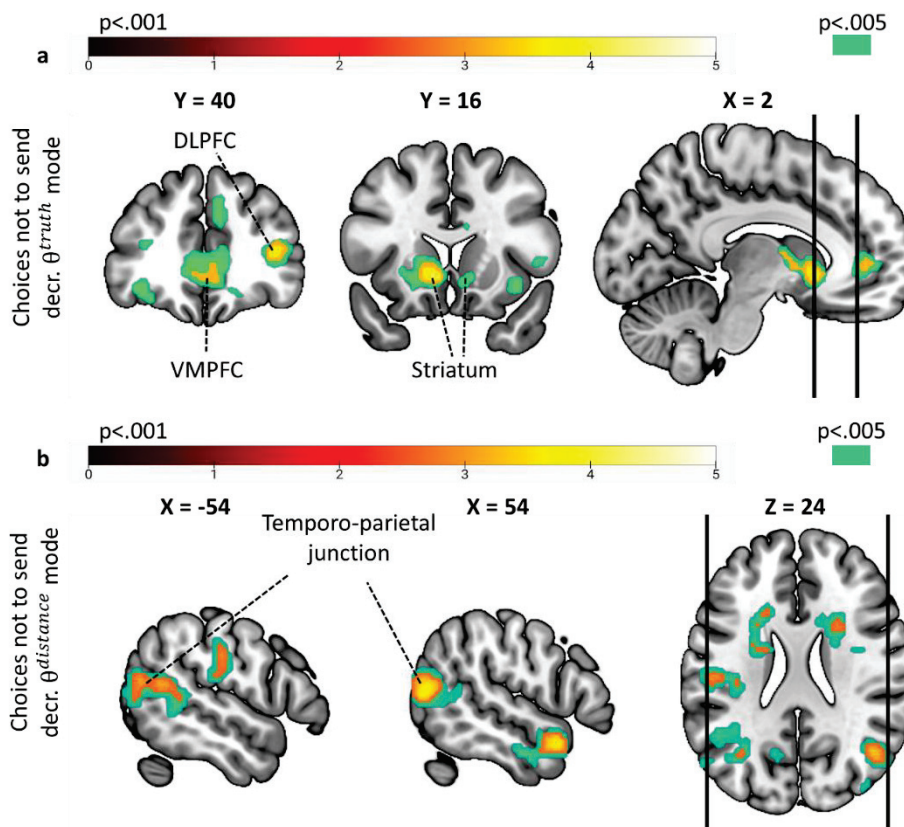




**Figure 6: The confidence in truthfulness estimation and the Receivers' social distance to the relevant organisation elicited corresponding activity in an information valuation network and a social processing network.** Throughout the task participants learned to decrease their probability to send extra information. **a)** When participants had no cue about Receivers, an increase in confidence in the news item truthfulness judgement was associated with a decrease in the probability. This increase in confidence modulated the BOLD signal in vmPFC, dlPFC, midbrain and striatum, with greater effect during choices not to send. **b)** When they received the social distance cue, it anti-correlated with the probability to send additional information. The Receiver's social distance to the relevant organization modulated the BOLD signal in STS and DMPFC, with greater effect during choices not to send. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to  $p < .001$  and  $p < .005$  (cluster-level thresholds at  $p < .05$  uncorrected).

To test the hypothesis that beliefs about Receivers' preferences based on truthfulness probability elicited activity in the information valuation areas, we performed a whole-brain analysis with the mode of  $\theta^{truth}$  as parametric modulator. Results showed that a decrease of

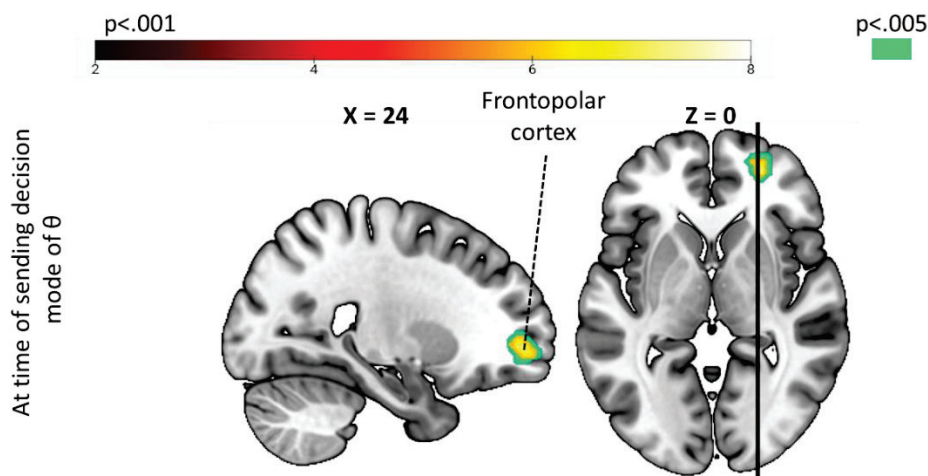
the mode during decisions not to send anti-correlated with brain activity in striatum, vmPFC and dlPFC (all  $p$  voxel-level  $< .001$  uncorrected, cluster-level  $< .05$ , Figure 6b). In summary, when the mode of beliefs that Receivers' are less willing to receive additional information decrease, due to information truthfulness probability, activity in the information valuation brain areas increases. We then performed a whole-brain analysis with the mode of  $\theta^{distance}$  as parametric modulator. This analysis tested the hypothesis that beliefs about Receivers' preferences given Receivers' social distance to the organization provoke an increased activity in the social processing areas. Results show that brain activity in the temporo-parietal junction (TPJ) anti-correlated with the mode during decisions not to send (all  $p$  voxel-level  $< .001$  uncorrected, cluster-level  $< .05$ , Figure 7b). Taken together with the behavioral parameter, this shows that beliefs about others' preferences, when based on probability of others' social distance to organizations, elicit activity in social processing areas ([Supplementary VIII](#)).



**Figure 7: Beliefs about Receivers' preferences based on confidence-elicited brain activity in the information valuation network. Social distance-elicited brain activity in the TPJ. a)** Beliefs about Receivers' preferences, based on truthfulness probability were represented by the mode of  $\theta^{truth}$ . During choices not to send, a decrease in the mode

indicated increasing belief in the Receiver's preference not to receive additional information, and were associated with heightened BOLD signal in vmPFC, dlPFC and striatum. **b)** Beliefs about preferences based on social distance to the relevant organisation were represented by the mode of  $\theta^{distance}$ . During choices not to send, decreases in the mode were associated with heightened BOLD signal in TPJ. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to  $p < .001$  and  $p < .005$  (cluster-level thresholds at  $p < .05$  uncorrected).

The Bayesian models comparison highlighted that participants took into account the learning to form beliefs about Receivers' preferences. However, the weight of the learning, represented by the weight of  $\theta^{learning}$ , was very low. Thus, it is likely that it was not taken into account to take sending decisions. Modelled in an MLM, the mode of  $\theta^{learning}$  along with the modes of the other parameters brought evidence that learning was not a variable significantly accounting for the participants' probability to send additional information. Consequently, we didn't look at brain activity that correlates with the mode of  $\theta^{learning}$ . However, we analysed the modulator effect of the mode of  $\theta$  on brain activity at the time of the sending decision. This mode represents the participants' beliefs about the receivers' preferences based on the four parameters. It includes the hyper-parameters from all four parameters. We found a signal increase in the right frontopolar cortex (rFPC) corresponding to a change in the mode of  $\theta$  (figure 8). This suggests that self-related and others-related beliefs are integrated in the rFPC when inferring others' preferences for information ([Supplementary VIII](#)).



**Figure 8: When participants integrate self-oriented beliefs and other-oriented beliefs, the estimated receivers' preferences was associated with increased activity in the lateral frontopolar cortex.** The mode of  $\theta$  represents the participants' beliefs about the Receivers' preferences based on the four parameters We looked at whole-brain activity in both conditions, that vary with that mode at the time of the sending decision and found BOLD signal increased in the right lateral frontopolar cortex. For illustration purposes, we represented whole-brain activities at voxel-level thresholds equal to  $p < .001$  and  $p < .005$  (cluster-level thresholds at  $p < .05$  uncorrected).

#### IV. Discussion

We investigated the brain computations engaged in decisions to share extra information about uncertain news items with others. Preferences of others are largely hidden, making it difficult to predict whether a given person in a social network will be willing to receive or not debunking or confirmational information (FeldmanHall & Shenhav, 2019). However, when people are given information regarding others' preferences, humans can learn another person's opinions to predict their choice behavior. This is true when information has instrumental value (i.e. it helps action selection to acquire future rewards for oneself or others), such as when choosing on behalf of others (Nicolle et al., 2012), when predicting others' likes (Kang et al., 2013) or when predicting others' decisions (Suzuki et al., 2012). Here, we showed that a similar process is engaged when information about others' has no instrumental value, but only reduces one's uncertainty (Cohen et al., 2007; Golman et al., 2021; Jezzini et al., 2021; Kobayashi & Hsu, 2019). Indeed, when provided information regarding social proximity between Receivers' preferences and the nature of the news, participants learnt to match the Receivers' choices. The decision to send extra information was at chance level when no information about Receivers' preferences was provided.

When no information about the Receivers' opinion (i.e., social proximity between Receivers' preferences and the nature of the news) was provided, participants were at chance level to send extra information. They overestimated the Receivers' preference to seek extra information. In this Control condition, the higher their confidence that the news was true or false, the less likely they were to send extra information to others. When participants judged the news to be false but were weakly confident about their judgment, and were more likely to

send extra information. This increased willingness to send extra information to others when the news is not trusted suggests an inferred desire to see fake news debunked.

Comparatively, the frequency with which extra information was sent was less affected by a lower confidence in the truthfulness of news items judged to be true. This observation reflects an asymmetrical impact of the valence of information truthfulness. A similar effect of valence has been observed for acquiring information with instrumental utility. The more likely participants were to win a lottery, the more they wanted to know its outcome; the more likely they were to lose, the less they wanted to know the outcome (Charpentier et al., 2018).

In the Control condition, participants relied only upon their confidence in the truthfulness of the news. Unlike this condition, in the Cue condition receiver's preferences for extra information could be inferred based on the cue to their opinion. In that condition, the confidence in one's own judgment of the news items truth no longer determined the decision to send extra information. Instead, it was determined by the cue indicating a closer proximity between Receivers' preferences and the nature of the news. When provided with receivers' opinion, participants improved their estimation, their average send rate fell to 38% for a corresponding success rate of 58%. The average probability densities of sent choices and reception choices for each theme were not significantly different in the Cue condition, while they were in the Control condition.

Our Bayesian model explained how people integrate their own confidence and the social distance between the receiver's opinion and the content of information. They learnt preferences about the group of Receivers throughout the task. Results showed that participants only learnt to decrease their sending for information judged as true. Looking at social distance, we found in the Cue condition that the separation between the participant's and the Receivers' social distances to the organizations congruent with the news items significantly explained the probability to send. Precisely, the probability to send increased by 30% when the separation of the two social distances increased by one standard deviation, and the Receiver was closer to the organization than the participant. However, we found that the receivers' social distance to the organization explained still better the probability to send, decreasing by 44% as social distance increased by one standard deviation. It may be that participants did not estimate similarity between the Receivers and themselves but used the social distance between the Receivers and the organization to estimate receivers' knowledge and preference.



Supporting our hypothesis for shifts in beliefs, when participants were provided with cues about Receivers, the significance of independent variables varied across conditions. In the Cue condition, the effect of truthfulness estimation was no longer significant while the learning effect became steeper. These first results indicate that participants revised their beliefs about similarity between themselves and Receivers.

#### IV.1. Computational mechanisms of estimating others' preferences for information

At the computational level, our Bayesian model considered that participants' beliefs about Receivers' preferences were based on estimation of a probability distribution of receivers' action (i.e., to receive or not to receive). The model integrates four components: the truthfulness estimation of the information, the history of Receivers' past actions, the social proximity between the Receiver's opinion and the content of the news items and a bias in estimating Receivers' probability to choose one action. This model showed that participants effectively weighed beliefs about information truthfulness, beliefs about the simulated population preference and beliefs about the target agent's beliefs. The probability to send in the Control condition was only explained by beliefs about truthfulness of the news. This was no longer true in the Cue condition. Beliefs about Receivers' social distance to the theme of the news items were significant, as predicted in the Cue condition and surprisingly, in the Control condition too. This provides evidence that participants' knowledge of information seemed to guide their sending behavior. This also confirms that participants shifted the weight they grant to beliefs during the task. Likely, this shift helped participants bring the estimated Receivers' preferences towards the empiric Receivers' preferences. However, while participants did learn the Receivers' population preferences during the task, the linear modelling showed that participants didn't significantly use that social information in their inference process.

Our computational model extends previous Bayesian models accounting for beliefs formation and inferences of others' preferences. In the latter models, individuals integrate the estimated value of rewards to others to their own subjective value (Huber et al., 2013; Jayles et al., 2017; S. A. Park et al., 2017; Toelch & Dolan, 2015; Wu et al., 2016). It is also consistent with Bayesian inference models that show confidence in a choice reflects the probability that

the choice is correct (Aitchison, Bang, Bahrami, & Latham, 2015; Fleming & Daw, 2017; Ma & Jazayeri, 2014; Meyniel, Sigman, & Mainen, 2015; Pouget et al., 2016).

#### IV.2. Neurocomputational bases of the decision whether to send extra information to others

Our Bayesian framework allowed us to distinguish three brain systems engaged in computing the decision to send extra information. The information valuation brain network was engaged in supporting self-oriented beliefs regarding confidence in information truthfulness. A second brain network, consisting in the dmPFC and rTPJ, was engaged in supporting other-oriented second-order beliefs (i.e. beliefs about others' preferences). Third, the FPC integrated the different signals of our model as a signal reflecting the most probable belief. Below we discuss these three main findings.

The valuation brain network, in particular the vmPFC and ventral striatum, is tied to information valuation (Bromberg-Martin & Monosov, 2020; Charpentier et al., 2018). Converging evidence points to a central role of the vmPFC in the computation of decision confidence (Bang & Fleming, 2018; Gherman & Philiastides, 2018; Pereira et al., 2020), the mapping of this internal estimate onto explicit reports (Bang, Ershadmanesh, Nili, & Fleming, 2020; Fleming et al., 2012), and the integration of single-trial confidence estimates into a long-run estimate of task performance (Rouault & Fleming, 2020). Information value is tied to the ability of information to reduce uncertainty. When they were uncertain (i.e. not confident) in their evaluation of information truthfulness, participants estimated that Receivers chose to receive extra information. When confidence in information truthfulness decreased, we observed an increase of signal in brain valuation areas, notably the vmPFC and striatum, as well as in the dlPFC. These brain regions are also known to be engaged in information seeking (Bromberg-Martin & Monosov, 2020; Charpentier et al., 2018; White et al., 2019). The computational analysis of the behavioral data showed that beliefs about others, based on truthfulness estimation, led to an increased BOLD signal in the valuation brain network. In the context of our study, participants had to decrease their beliefs about Receivers' preferences to be successful. When confidence in the most probable beliefs decreased, we found a heightened activity in striatum and vmPFC. The striatum and vmPFC have also been found to represent



both instrumental and non-instrumental motives such as motivation to acquire information (Kobayashi & Hsu, 2019).

The increased brain activity in TPJ, DMPFC and STS, anti-correlating with Receivers' social distance to organizations, parallel our behavioral results. Participants had higher preference for information when Receivers social distance to the organizations increased. The computational bases of this finding was identified by our model-based fMRI analysis. The neural signal of the TPJ reflect computations of the beliefs about others, when beliefs were based on the social distance. The signal in this area anti-correlated with the most probable beliefs. The dmPFC, TPJ brain network is involved in others-related preferences (Joiner et al., 2017) and mentalizing (Molenberghs, Johnson, Henry, & Mattingley, 2016). Recent perceptual decision making studies have also shown that computations of confidence in another player's decisions are performed by combining distinct estimates of player ability and decision difficulty. This process engages an interaction between brain systems implicated in decision-making, such as lateral intraparietal sulcus, and theory of mind, such as TPJ and dmPFC (Bang, Moran, Daw, & Fleming, 2021).

Using a Bayesian model and fMRI, we show that the lateral frontopolar cortex (LFPC) is the only region that varies with the most probable belief ( $\theta$ ), integrating the different parameters of the model. This region integrated self-oriented and others-oriented beliefs in the inference process. One of the general roles ascribed to the FPC, based on individual decision making studies, is the integration of information. This integrative role is supported by anatomical features of FPC (Jacobs, Driscoll, & Schall, 1997; Jacobs et al., 2001), by the fact that this region evolved part of the frontal lobes (Semendeferi, Armstrong, Schleicher, Zilles, & Van Hoesen, 2001) and that it is a late cortical structure to reach maturation (Flechsigs, 1901; Gogtay et al., 2004). At the functional level, different theories of LPFC have placed this region at the top of a prefrontal cortical hierarchy in terms of cognitive control capacities, such that some regions influence others more than vice versa (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Donoso, Collins, & Koechlin, 2014; Dreher et al., 2008; Zajkowski, Kossut, & Wilson, 2017). For example, the lateral FPC tracks the relative advantage in favor of switching to a foregone alternative (Badre & Nee, 2018; Boorman, Behrens, & Rushworth, 2011; Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999; Ramnani & Owen, 2004), independently of whether such switches are externally instructed or voluntary, and independently of the complexity of the pending behavior (Boorman, Behrens, Woolrich, & Rushworth, 2009;

Boschin, Piekema, & Buckley, 2015; Mansouri, Freedman, & Buckley, 2020). Yet, in the social domain, the role of the FPC remains unclear and it is unknown whether it integrates knowledge about others' ability with our own assessment when computing confidence in others' choices. A recent study reported that the lateral FPC may contextualize an internal sense of confidence for explicit report in accordance with task requirements when using an interactive task which required subjects to adapt how they communicated their confidence about a perceptual decision to the social context (Bang et al., 2020). Our study supports and refines this hypothesis, by showing that the LFPC integrates private beliefs with beliefs attributed to others. The current findings suggest that managing competing goals in the cognitive domain may be only a subfunction of a more general function of the FPC. The structure might manage and integrate different types of beliefs, such as individual and social beliefs, in the case of the present study (S. A. Park, Sestito, Boorman, & Dreher, 2019).

### IV.3. Conclusions

We identified the neurocomputational processes of inferring other's preferences for information via a Bayesian model in which participants weighed different types of beliefs (i.e., first order and second order beliefs). This Bayesian model accounted for sending decisions by weighing beliefs about information truthfulness, the simulated population preference and the target agent's social proximity with the nature of the information. Our behavioral manipulation, explicitly asking confidence level in the news, has practical implication for social media platforms. For instance, recent social media experiments where people could use 'trust' and 'distrust' buttons, slowed down the spread of misinformation more so than the commonly existing engagement options (Globig et al., 2022). In consequence, the incentive structure of sharing was associated to the veracity of the information instead of the social approbation. This type of intervention has the potential to reduce violence, vaccine hesitancy and political polarization, without reducing user engagement. Finally, our neuroimaging results identified the brain systems engaged in updating the beliefs about others' preferences for acquiring extra information about news.

# Chapter III

## Testosterone Causes Decoupling of Orbitofrontal Cortex-Amygdala Relationship While Anticipating Primary and Secondary Rewards

Valentin Guigon<sup>1</sup>, Simon Dunne<sup>2,3</sup>, Agnieszka Pazderska<sup>4</sup>, Thomas Frodl<sup>2,5,6</sup>, John J. Nolan<sup>4,7</sup>, Nicolas Clairis<sup>8</sup>, John P. O'Doherty<sup>2,3,9</sup>, Jean-Claude Dreher<sup>1,2</sup>

### Affiliations

<sup>1</sup> CNRS, Neuroeconomics lab, ISCMJ, Lyon, France, and Université Claude Bernard Lyon 1, Lyon 69100, France

<sup>2</sup>Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

<sup>3</sup>Computation and Neural Systems, California Institute of Technology, Pasadena, CA

<sup>4</sup>Department of Endocrinology, St James's Hospital, Dublin, Ireland

<sup>5</sup>Department of Psychiatry, School of Medicine, Trinity College Dublin, Dublin, Ireland.

<sup>6</sup>Department of Psychiatry and Psychotherapy, Otto von Guericke University of Magdeburg, Germany

<sup>7</sup>Steno Diabetes Center, Gentofte, Denmark

<sup>8</sup>ICM Institute for Brain and Spinal Cord, Paris, France

<sup>9</sup>Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CAA

### Abstract

Correlational evidence shows that levels of testosterone are positively related to reward sensitivity in humans. Yet, studies of the direct effects of exogenous testosterone administration

on the reward system in human males are scarce. We sought to investigate the effects of testosterone injection on behavior and brain activations while participants were anticipating erotic or monetary rewards. Healthy young male participants (N=40) performed an incentive delay task with cued erotic and monetary rewards in a between-subjects, double-blind, placebo-controlled design. We hypothesized that testosterone administration may: 1) increase posterior lateral orbitofrontal cortex activity, previously observed to be engaged more with erotic as compared to monetary rewards in healthy young men; (2) decrease the functional coupling between the medial part of the orbitofrontal cortex and the amygdala while anticipating rewards. Results show testosterone specifically increased incentive behavior related to erotic stimuli as compared to monetary rewards. Participants were faster to respond when administered testosterone for cues announcing erotic compared to monetary rewards. This behavioral interaction effect was associated with higher association between relative motivational value for erotic as compared to monetary cues in the ventral striatum. No changes were observed after testosterone injection in the posterior lateral orbitofrontal cortex while viewing erotic rewards. However, testosterone injection reduced the functional coupling between the ventromedial prefrontal cortex and the amygdala while anticipating both primary and secondary rewards, showing testosterone affects limbic-prefrontal connectivity during reward processing.

## I. Introduction

Hormones play a strong role in modulating adult behavior. Testosterone is one of the most significant, with concentration levels associated with fluctuations in motivation (Aarts & Van Honk, 2009), sexual motivation (Travison, Morley, Araujo, O'Donnell, & McKinlay, 2006), mood (Anderson, Bancroft, & Wu, 1992), aggression (Archer, 2006; Carré, McCormick, & Hariri, 2011; Nelson & Trainor, 2007) and social decision-making (P. A. Bos, Panksepp, Bluthé, & Honk, 2012; Sapienza, Zingales, & Maestripieri, 2009). Studies have pointed its implication in dopaminergic pathways (Aubele & Kritzer, 2012; Dimeo & Wood, 2006; Frye, Rhodes, Rosellini, & Svare, 2002; Nagypál & Wood, 2007; Packard, Schroeder, & Alexander, 1998; Teresa Arnedo, Salvador, Martínez-Sanchis, & Pellicer, 2002), prefrontal cortex (Sinclair, Purves-Tyson, Allen, & Weickert, 2014; Wood, 2008), amygdala (Radke et al., 2015; G. A. Van Wingen et al., 2009) and striatum (Hermans et al., 2010a). These observations suggest testosterone might affect brain networks responsible for incentive processing but no clear evidence so far supports this assumption. In this study, we report a modulating effect of

testosterone on the reward system activity and connectivity associated with anticipation of both primary and secondary rewards.

Studies about testosterone effects on sensitivity to punishment and reward dependency (Van Honk et al., 2004) or risk-taking (Apicella, Carré, & Dreber, 2015; Stenstrom & Saad, 2011) advocate for testosterone affecting reward processing. For long, testosterone induced effects on motivated behavior during BOLD fMRI studies have been investigated through social or emotional stimuli processing. A first link established between testosterone and modulation of reward processing investigated adolescents performing a card-guessing task. Manipulating monetary incentives, authors reported association between pubertal development and dorsolateral prefrontal cortex (DLPFC) activity and connectivity with nucleus accumbens (NAcc) (Poon, Niehaus, Thompson, & M., 2019). However, primary (e.g., drink, food, sex) and secondary (e.g., money, power, social approval) rewards elicit different processes and overlaps. Values associated to secondary rewards are abstract and both reward types processing show a distinct abstract-to-concrete organization in the OFC (Li et al., 2015). This raises questions regarding the reward brain network during testosterone-modulated reward processing and the presence of areas common to all anticipated rewards.

When making a choice, the rewarding properties one anticipates are encoded in regions such as ventral striatum (VS) and ventromedial prefrontal cortex (VMPFC), representing a common neural currency (Montague & Berns, 2002; Rangel, Camerer, & Montague, 2008; Sugrue, Corrado, & Newsome, 2005). These regions are part of a wider network. For instance, when experiencing value of reward, regardless of their type, brain activity is heightened in a common reward network. It is composed of ventromedial prefrontal cortex (VMPFC), ventral striatum (VS), amygdala, anterior insula (AI) and medio-dorsal thalamus (Sescousse, Li, et al., 2013; Sescousse et al., 2010). Evidence points towards a similar network when anticipating rewards, with high consistence across studies in VS activation (Bartra et al., 2013; Diekhof et al., 2012; Knutson et al., 2001; J. P. O'Doherty et al., 2002; Oldham et al., 2018; Wilson et al., 2018). Oldham and colleagues (2018) successfully highlighted the role these regions in monetary incentive delay tasks (Oldham et al., 2018). Over various studies, areas part of this network have proven to be targeted by testosterone in various motivated behaviors, specifically VS, VMPFC and amygdala.

In humans, administration of exogenous testosterone interacts with mesolimbic dopamine pathways, regulating incentive sensitivity (Wood, 2008). VS, a part of the

mesolimbic dopaminergic pathway, is subject to increased testosterone-induced effects in adults (Hermans et al., 2010b) and adolescents (Op De MacKs et al., 2011) during rewarding tasks. Amygdala, known for the large number of nuclear androgen receptors in its neurons (Simerly, Swanson, Chang, & Muramatsu, 1990), responds more to emotional stimuli when testosterone levels increase (G. A. Van Wingen et al., 2009; G. van Wingen, Mattern, Verkes, Buitelaar, & Fernández, 2010). With prefrontal cortex (PFC), its activity and connectivity is affected by its levels in social emotional behaviour (P. A. Bos, Hermans, Ramsey, & Van Honk, 2012; Radke et al., 2015; Volman, Toni, Verhagen, & Roelofs, 2011). Testosterone might affect regulation of neural activity within circuitry between both regions (Bialy & Sachs, 2002; Hermans, Ramsey, & van Honk, 2008; G. A. Van Wingen et al., 2009). During risk-taking testosterone induces OFC suppression, causing increased desire for monetary rewards and decreased sensitivity to punishment (Stanton, Liening, & Schultheiss, 2011; Stanton, Mullette-Gillman, et al., 2011). It also provokes a decoupling between OFC and amygdala in testosterone-induced participants (G. A. Van Wingen et al., 2009; G. van Wingen et al., 2010).

Building on these considerations, the present study used BOLD-fMRI to investigate the effects of exogenous testosterone on the reward system with an incentive delay task. We investigate the brain responses to two rewards, money and erotic pictures, at the anticipation phase. These two rewards present significant evolutionary differences likely to be reflected at the cerebral level. Money is a secondary reward which appeared recently in human history. Its abstract value needs to be learned by association with primary reinforcers. Erotic stimuli can be considered as primary rewards due to their innate value, satisfying biological needs. They are rewarding (Hamann, Herman, Nolan, & Wallen, 2004; Hamann et al., 2014) probably because sexual attractiveness evolved to become an important cue for mate choice (Rhodes, 2006), triggering higher VS activity (Redouté et al., 2000). We hypothesized common brain structures support general hedonic representations independent of reward type.

Healthy participants performed a version of the incentive delay task (Sescousse et al., 2010) in a double-blind protocol with injection of testosterone, controlled by placebo. Rewards were event-administered and depended on participants' performance in a perceptual discrimination sub-task, that required an effort to obtain the rewards. The task also included an active delivery mode with probabilistic and variable reward magnitude. The paradigm aimed to explore variations in the activation patterns of the brain regions involved in anticipating rewards regardless of their type. We were interested in the modulatory effect of testosterone on

participants' behaviors, brain functional activity and functional connectivity, as a function of reward probability and magnitude.

Behavioral analyses, functional analyses of fMRI data and functional connectivity analyses have been conducted. We expected the task to yield activations from a neural network comprised of VS, VMPFC and amygdala at the reward anticipation. Because plasma T modulates both sexual arousability and the response of various brain areas involved in sexual behavior, we expected a larger effect of testosterone on trials with erotic rewards, compared to monetary. We reasoned that demonstrating the T-dependency of the response of brain areas to VS would be evidence that these responses were related to sexual arousal and not merely to a state of general motivational arousal. Furthermore, we expected a significant decoupling to arise between the amygdala and the VMPFC among testosterone participants.

## II. Methods

### II.1. Participants

Forty-five right-handed heterosexual men (mean  $\pm$  SD age:  $21.25 \pm 2.97$  years) with no history of neurological or psychiatric disorders participated in this study. Two of them failed screening, one due to a high score on the Anxiety/Depression scale and the other because of low testosterone level (total testosterone = 11.8 nmol/L ; free fraction T = 0.224) at the initial visit. Three subjects completed the screening visit were then unable to continue their participation in the study. Forty subjects completed the fMRI study.

All subjects answered a set of questionnaires including the Sexual Arousal and Desire Inventory (SADI) (Toledano & Pfaus, 2006) and the Behavioral Activation Scale (BAS) (Carver & White, 1994). These enabled us to measure that participants showed a similar basal state of motivation before testosterone (or placebo) injection. Questionnaires were missing for one placebo and one testosterone participant. We compared placebos (N=18) and testosterone (N=20) subjects using an unpaired two-sample Student's t-test. One additional placebo subject was excluded as a result of the SADI analysis (N=17 for placebos) after a Grubbs' test ( $p < 0.01$ ). To further ensure similar states of motivation to see erotic stimuli between participants, they were asked to avoid any sexual contact during a period of 24h prior to the scanning session. To enhance their motivation for money, subjects were told their financial compensation for participation would depend on their winnings during the task.



Participants were all students recruited by advertisements at Trinity College Dublin and St. James Hospital, Ireland. The study was approved by two local ethics committees (Trinity College Dublin and St. James Hospital) in accordance with the Declaration of Helsinki. Written informed consent was obtained from all subjects prior to the examination. Participants were debriefed and received payment (50€). They were screened for inclusion and exclusion criteria by a professional endocrinologist.

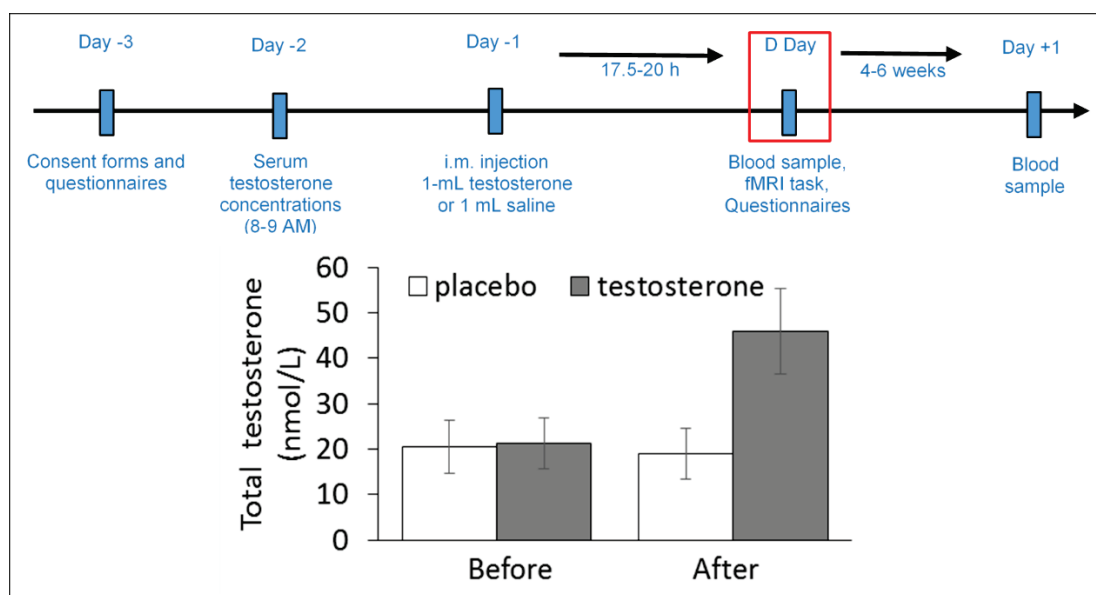
## II.2. Testosterone injection and blood samples

Subjects were injected intramuscularly with either a placebo (1mL of saline solution) or testosterone (1mL dose of 250mg testosterone enanthate, Androtardyl/Testoviron Depot) at St. James Hospital, approximately 12 hours before the scanning session. We chose to measure blood testosterone levels because salivary testosterone measures have proven to be a less reliable measure of the biologically active, non-protein-bound proportion (free testosterone) than testosterone in blood serum (REF). The pharmacokinetics of testosterone enanthate yield supraphysiological testosterone levels in serum as early as 2 h following injection, reaching peak levels 4 to 5 times above basal between 8 and 24 h after injection (Nieschlag & Behre, n.d.; Schürmeyer & Nieschlag, 1984; Snyder & Lawrence, 1980). We measured testosterone concentrations from blood samples twice: once on the day of the second visit at St James hospital before intramuscular injection of 250 mg of testosterone enanthate (Androtardyl/Testoviron Depot) or placebo (sesame oil); and once in the morning of the scanning session, before scanning and approximately 12 hours after injection of the drug or placebo.

We determined serum concentrations of total testosterone (TT) by the electrochemiluminescence immunoassay (ECLIA) kit on a cobas e analyzer (Roche Diagnostic Systems). Serum concentrations of sex hormone-binding globulin (SHBG) were measured by ECLIA kit on a cobas e analyzer. Serum albumin was measured by colorimetric assay (ALB2; cobas e analyzer). Apparent concentrations of free testosterone (FT) and available testosterone (AT) were calculated from values of TT, SHBG, and albumin using the method described and validated by (Vermeulen, Verdonck, & Kaufman, 1999). Estradiol was measured by ECLIA (cobas e analyzer). Blood samples were not obtained from one participant at Appointment 2 and one participant at Appointment 4 because of experimenter error. Those participants are

omitted from figures and analyses involving measurements of testosterone at the respective time points.

Total serum testosterone concentrations were compared with a 2-way repeated measures ANOVA that included the sampling time (before/after the injection) as within-subject factor and group (placebo/testosterone) as between-subject factor. Three subjects were excluded from the ANOVA: pre-injection measures were missing for two placebo-subjects and post-injection measures were missing for one testosterone-subject. A post-hoc Tukey's honestly significant difference (HSD) test was used to compare groups by pairs.



**Figure 1: Timing of pharmacological procedure.** 40 young right-handed healthy men were included (ages:18-30 y.o; M=21.25, SD=2.97). 21 were injected 1-mL dose of testosterone enanthate (250 mg; Androtardyl/Testoviron Depot). Placebo participants were administered 1mL saline solution. Testosterone was above-basal level between 8 to 24 h after injection. There was no significant difference between the two groups before administration of testosterone. However, 17,5-20 hours post injection of testosterone or placebo the blood levels of testosterone of the group receiving testosterone injections (N=20) is significantly higher than those receiving placebo (N=17) ( $p < 0.0005$  Tukey's HSD test).

### II.3. Questionnaires

Subjects completed the validated versions of the following questionnaires: Sexual Arousal and Desire Inventory (Toledano & Pfaus, 2006), Beck Depression Inventory

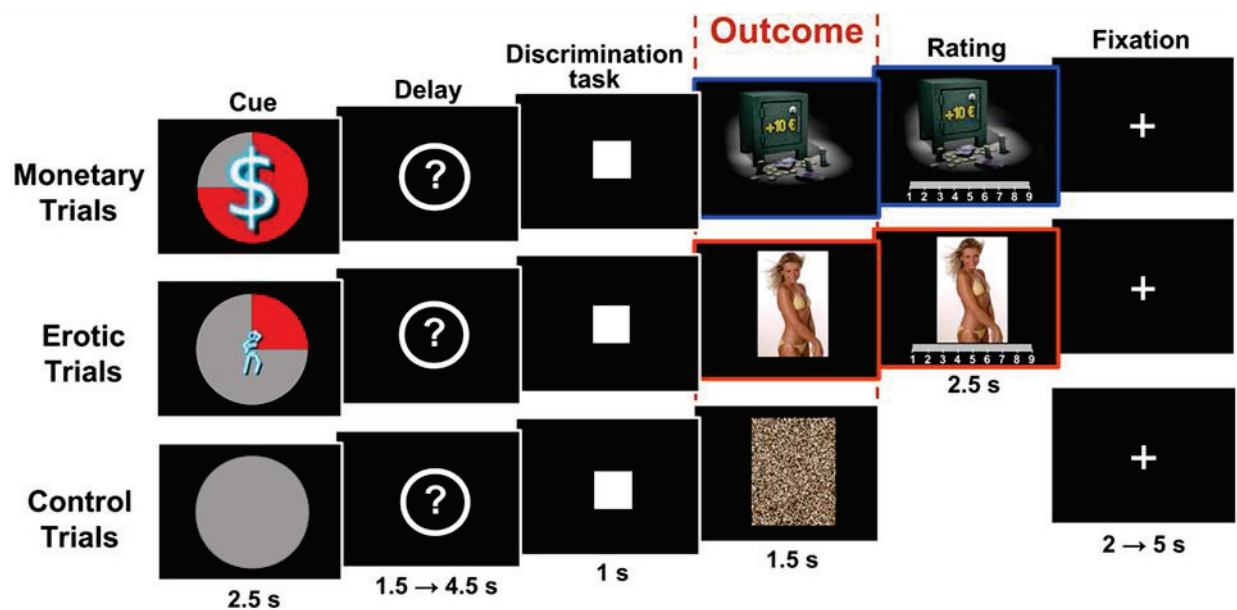
(13 item-BDI; Beck & Beck, 1972), Beck Anxiety Inventory (BAI; Steer & Beck, 1997), International Personality Item Pool (L. R. Goldberg, 1999; Lewis R. Goldberg et al., 2006), Profile of Mood Status (POMS; McNair, 1971), Machiavelli personality Test (MAIV; Christie & Geis, 1970), Eysenck personality questionnaire short scale (EPQR-S; Eysenck, Eysenck, & Barrett, 1985), Barratt Impulsiveness Scale (BIS-11; Patton, Stanford, & Barratt, 1995) Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983).

These questionnaires were administered twice to both groups of men: once at the time of screening (prior to scanning) and once on the day of the fMRI scan, after scanning (having received the injection of placebo or testosterone on the day before) to investigate the effect of testosterone administration on sexual arousal, behavior and mood states of subjects. Behavioral data obtained from the subjects were statistically analyzed using a mixed design ANOVA to compare differences between a between subject factor (Testosterone, placebo) and within subject factor (Before, After scanning). We administered a post-hoc survey after scanning regarding expected and experienced effects of injection.

#### II.4. Task

Each trial was divided into two phases: Reward Anticipation and Outcome. Reward Anticipation consisted of cue presentation, a delay period, then a discrimination task (Figure 2). Each cue carried three types of information regarding the upcoming reward: probability (25/50/75%), intensity (low/high) and type (monetary/erotic) (Figure S1). A total of 12 different cues, plus a control condition associated with no chance of winning, were used in this task. After a variable delay period (question mark representing a pseudo-random drawing depending on probability), subjects were asked to perform a discrimination task, in which they had to respond correctly to a target within a maximum time of 1000 ms. The shape of the target was drawn at random on each trial and could be either a triangle (left button press) or a square (right button press). Success on this discrimination task (indicated by a magnified target) allowed the subjects to view the outcome of the pseudo-random drawing, whereas erroneous or slow response (indicated by no change in target size) led to no reward. Success in the discrimination task rewarded subjects with an outcome presentation: an erotic picture or the picture of a safe including the amount of money won. Nudity being the main driver of erotic stimuli reward value, we separated stimuli into a “low intensity” and a “high intensity” group. The “low intensity” group displayed women in underwear or bathing suits. The “high intensity” group displaying naked women in suggestive postures. Each erotic picture was presented only once

during the course of the task. We introduced a similar element of surprise for monetary rewards by randomly varying the amounts at stake. The “low amounts” group was composed of 1€, 2€ or 3€ rewards. The “high amounts” group was composed of 10€, 11€ or 12€ rewards. Following each reward outcome, subjects were asked to provide a hedonic rating by moving a cursor along a 1-to-9 continuous scale (very low=1; very high=9). In non-rewarded and control trials, the subjects were presented with “scrambled” version of rewarding pictures, hence they contained the same information in terms of chromaticity and luminance. A blank screen was finally used as an inter-trial interval of variable length.



**Figure 2: The task during the fMRI scan.** Subjects first saw a cue for 2.5s indicating the reward type (pictogram), the reward intensity (size of pictogram) and the rewarding probability (red portion of pie chart). An empty circle indicated a control trial with no reward. After a delay period (1.5-4.5s), subjects performed a discrimination task within a 1000ms time window. A successful discrimination led to the outcome presentation for 1.5s.

The outcome presentation was consistent with its cue (reward type, magnitude and probability). Unrewarded and control trial outcomes were scrambled pictures. After each trial, subjects provided a hedonic rating of the reward on a Likert scale from 1 (very low) to 9 (very high). Inter-trial intervals consisted in blank screens.

## II.5. Behavioural analysis

Response accuracy and response time measures were collected for each of the forty subjects during the fMRI sessions. Hit rates and reaction times of correct answers and hedonic ratings were analyzed with mixed linear models using reward type, probability and intensity as within-subject factors, group as between-subject factor and subject as random effect. Statistical analyses were run on MATLAB (Mathworks Inc.).

## II.6. fMRI acquisition

Imaging was conducted on a Philips 3 Tesla whole-body scanner, with an eight-channel head coil, at the Trinity College Institute of Neuroscience (Trinity College Dublin, Ireland). Functional imaging consisted of blood-oxygen-level-dependent sensitive images acquired during performance of the behavioral task, using a gradient-recalled echo-planar sequence (repetition time=2s; echo time=78ms; matrix=80x80; flip angle=90°). Thirty-nine contiguous horizontal slices of 3.25mm thickness were acquired, encompassing the whole brain (field of view=240\*240mm<sup>2</sup>; voxel size=3\*3\*3.25mm<sup>3</sup>, gap=0.35mm). A whole-brain high-resolution T1-weighted structural scan (voxel size: 0.9\*0.9\*0.9mm<sup>3</sup>) was also acquired for each subject.

The scanning session was divided into three runs. Each functional run consisted of 370 volumes. Each of them included four repetitions of each cue, (except the control condition), repeated nine times. Within each run, the order of the different conditions was pseudo-randomized and optimized for further signal deconvolution. Testosterone levels are prone to variations during the day therefore all scanning sessions were scheduled in the morning (9am-10h15am). All subjects were given oral instructions and familiarized themselves with the cognitive task prior to scanning, during a short training session.

## II.7. fMRI analysis

### II.7.1. Data pre-processing

Functional data were preprocessed with MATLAB (Mathworks Inc.) and SPM8 (<http://www.fil.ion.ucl.ac.uk/spm/spm8.html>). Images were realigned to the mean volume within each run to correct for motion artifacts using a 6-parameter rigid body affine transformation and resliced to correct for differences in slice acquisition timing, resulting in a minimal sum of squared differences with the reference image. The images were then normalized

to a standard stereotaxic space (Montreal Neurological Institute (MNI) Template Montreal, Quebec, Canada,) using a 12-parameter affine/non-linear transformation. Image intensity was scaled to the mean global intensity of each time-series. Structural images were coregistered to the mean functional image then segmented into different tissue types with SPM8 segmentation tool. These grey and white matter images were used as an initial template for DARTEL registration and normalization to the standardized stereotactic space. Functional data sets were spatially smoothed using an isotropic Gaussian kernel with a full-width-at-half-maximum of 8 mm. Data were submitted to an event-related general linear model analysis, fitting a reference hemodynamic response function to an impulse response function for each event in the observed time series data.

### *II.7.2. Whole-brain analysis*

We created voxel-wise statistical parametric maps from functional data using the general linear model. An fMRI design matrix was created for each subject. Each cue stimulus type (sex, money, control) and each outcome stimulus type (sex, money, control) was modeled as a separate categorical regressor. Regressors were based on individual response periods convolved with the canonical hemodynamic response function and its temporal derivative. Parametric modulators were included to investigate the effect of different levels of intensity and probability of stimuli in the reward anticipation and the effects of different levels of probability and hedonic ratings on outcome presentation phase. Realignment parameters from data preprocessing were included as regressors of no-interest to correct for neural activity associated with subject motion. Each participant's design matrix was then regressed against fMRI data to yield parameter estimates for each participant. Statistical analyses were performed at individual and group level. Contrasts of interest were performed on individual subject data in the first level analysis. For the second level group analysis, contrast images from all subjects were included in a second level random effect analysis.

We examined activation differences in brain regions by applying one sample t-tests for each group for each condition. We used two sample t-tests to investigate direct comparisons between testosterone and placebo subjects. For each contrast at group level, statistical parametric brain maps were generated and displayed the T-value in signal intensity of each voxel. Statistical maps were superimposed on a 152-subject, T1-averaged image from SPM8.

### *II.7.3. Identification of common brain regions*

To identify common brain regions activated by both erotic and monetary cues and rewards in both groups, we used the SPM8 conjunction analysis function. Bilateral amygdala and bilateral striatum regions of interest (ROIs) were extracted from Hammer's Individual Adult Brain Atlases (Hammers et al., 2003). Bilateral anterior cingulate cortex, frontal medial orbital and rectal gyrus ROIs were extracted from Automated Anatomical Labelling (AAL) Atlas (Tzourio-Mazoyer et al., 2002). All ROIs were combined with SPM toolbox MarsBaR (Brett, Anton, Valabregue, & Poline, 2002). We ran an SPM8 conjunction analysis to generate the conjunction of the following main effect contrasts: erotic cue (EC) – control cue (CC) & monetary cue (MC) – CC. This conjunction included both groups (testosterone and placebo). The conjunction analysis was inclusively masked with the combined ROIs. Finally, eventual differences were examined with a repeated measures ANOVAs including reward type (erotic/monetary) as a within-subject factor and group (placebos/testosterone) as a between-subject factor.

## *II.8. fMRI analysis*

### *II.8.1. Regions of interest*

Based on our results from the conjunction analysis, we selected from a meta-analysis left amygdala ROIs functionally connected to medial orbitofrontal cortex (Zald et al., 2014). We chose brain regions reported as the most frequently coactivated with medial and/or lateral portions of the OFC. Two ROIs were eventually selected for statistical tests: a left amygdala ROI (-22, -8, -18) showing consistent co-activations with mOFC and a left amygdala ROI (-18, -2, -12) showing consistent co-activations with the overlapping of mOFC and lOFC activations.

### *II.8.2. Identification of brain regions*

We implemented GLMs from SPM8 in MATLAB CONN toolbox (<http://www.nitrc.org/projects/conn>; Whitfield-Gabrieli & Nieto-Castanon, 2012) to run the functional connectivity analysis. To take into account the confound effects of participants'



movements without affecting intrinsic functional connectivity, we used the CompCor method (Behzadi, Restom, Liao, & Liu, 2007) implemented in the CONN toolbox. This method identifies principal components associated with white matter and cerebrospinal fluid. They were entered, along with realignment parameters, as confound variables in the first level analysis with a low-pass filter in frequencies below .008 Hz.

The CONN toolbox extracts the temporal evolution of whole brain activations, adjusted for the design matrix specified in SPMs. We calculated an interaction term between temporal series and both the reward type and the group (testosterone/placebo). ROIs temporal series and the term of the resulting psychophysiological interaction (PPI) have been entered in a bivariate correlational model for seed-to-voxels analysis. We identified treatment effects on functional connectivity between our ROIs and the areas of interest. The generalized PPI (gPPI) analysis was used to create voxel-based SPMs, which were then parceled according to the AAL atlas. Activation differences in cerebral regions were examined by applying t-tests to paired samples with the reward type (erotic, monetary) as within-subject factor and group (placebo, testosterone) as between-subject factor. For every contrast at group level, brain SPMs were generated, showing the T value in signal intensity for every voxel. Every contrast was performed with a significance threshold at voxel-level of  $p < .005$  uncorrected and a significance level at cluster-level at  $p < .05$  FDR-corrected for cluster size.

### III. Results

#### III.1. Testosterone injection and motivation

Before injections of testosterone or placebo the participants showed similar results on the BAS questionnaire, which reflects general motivation to obtain rewards (Student's t-test:  $p = 0.985$ ), however the group of testosterone subjects scored slightly lower than placebos in the SADI test (Student's t-test:  $p = 0.034$ ). There was a robust group, injection and group\*injection effect ( $p < .0001$ ) in the 2-way repeated measures ANOVA for the total testosterone concentration. This effect was solely driven by the testosterone group after the testosterone injection (Tukey's HSD test:  $p < .0005$ , in comparison to each of the other groups;  $p > 0.7$  between the other groups). This confirmed that there was no significant initial difference in the serum total testosterone concentrations before the administration of the testosterone and that the testosterone injection was effective (Figure 1).

### III.2. Questionnaires

Testosterone did not influence sexual lust, bodily arousal, and genital arousal as measured with the Sexual Arousal and Desire Inventory (SADI), state anxiety as measured with the Beck Anxiety Inventory (BAI) nor depression as measured with Beck Depression Inventory (BDI). It did not influence Machiavellianism as measured with Machiavelli Personality Test (MACH-IV), anxiety as measured by the BIS-BAS (Behavioral Inhibition System (BIS), Behavioral Activation System (BAS)), nor impulsivity as measured by the Baratt questionnaire or personality traits as measured by the IPIP questionnaire. (all  $p > .05$ ). Measures from BAI, BDI, MAIV, BIS anxiety, Baratt and IPIP shown no differences in responses to the questionnaires between the day of screening and the day of scanning in each of the two groups, demonstrating that the single injection of testosterone does not affect the responses to these questionnaires.

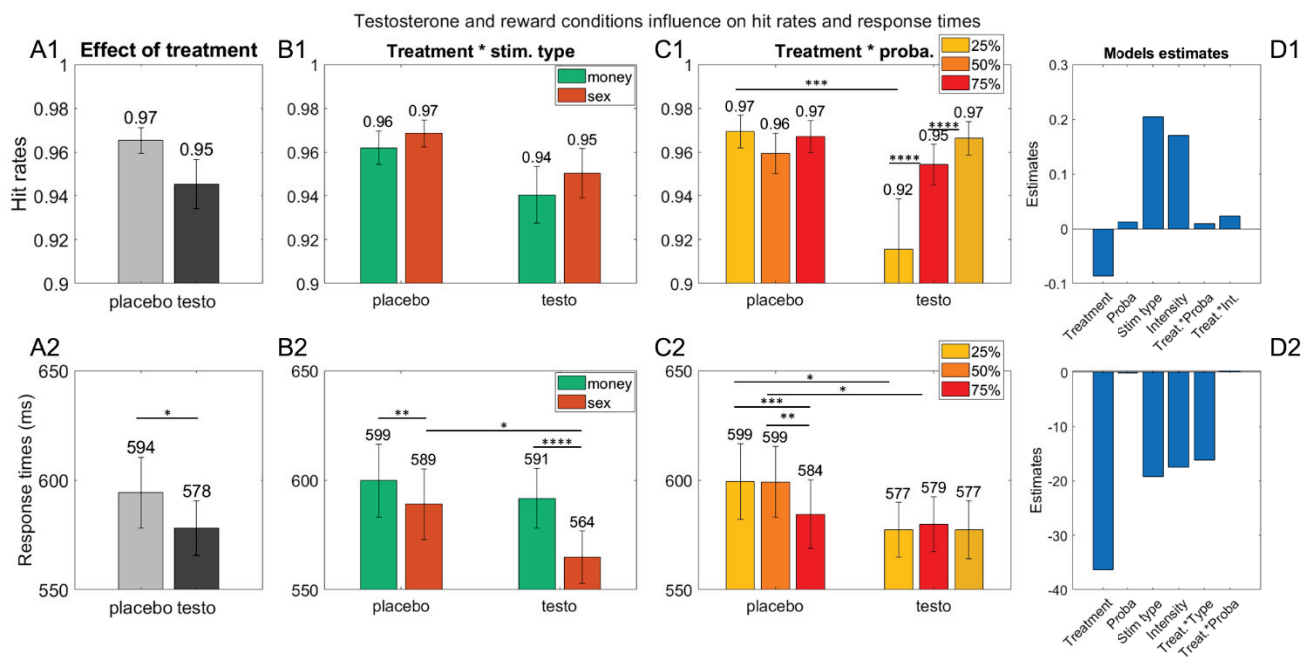
### III.3. Behavior

Hit rates and reaction times (RT), obtained at the time of the motor task, as well as hedonic ratings, obtained at the time of outcome, were analysed in separate generalized linear mixed-effects models (GLME) with reward type, probability and intensity as within-subject factors, treatment (placebo/testosterone) as a between-subject factor and subject as random effect. Further analyses with testosterone levels as between-subject factor were conducted. One subject was excluded from these models as data post-injection was missing for him. Treatments were treated as categorical (i.e., placebo or testosterone) and we included the number of missed trials as fixed effect. Probability percentages were treated by default as continuous and type and intensity as categorical.

GLME analyses were performed on repeated binomial (i.e., success or failure) hit rates measures. A first additive model revealed that the two treatments performed at ceiling and showed comparable mean hit rates on the discrimination task (placebo: 96.5%; testosterone: 94.5%; odds-ratio=.66,  $p=.095$ ). We didn't find any effect of reward type on hit rates (odds-ratio=1.21,  $p>.14$ ). Hit rates improved by one percent with each increase in unit of reward probability (odds-ratio=1.01,  $p<.001$ ) and an interaction model revealed a first-order interaction between reward probability and treatment (odds-ratio=1.02,  $p<.001$ ). This interaction was due to the fact that hit rates decrease in performance with the lowest probability stimuli in the testosterone group, but not the placebo group. Hit rates were not influenced by intensity (odds-

ratio=1.19,  $p=.18$ ) but there was a first-order interaction between reward intensity and group (odds-ratio=1.99,  $p=.01$ ). We added measures of bioavailable testosterone to further investigate the effect of treatment (Supplementary II, table S1).

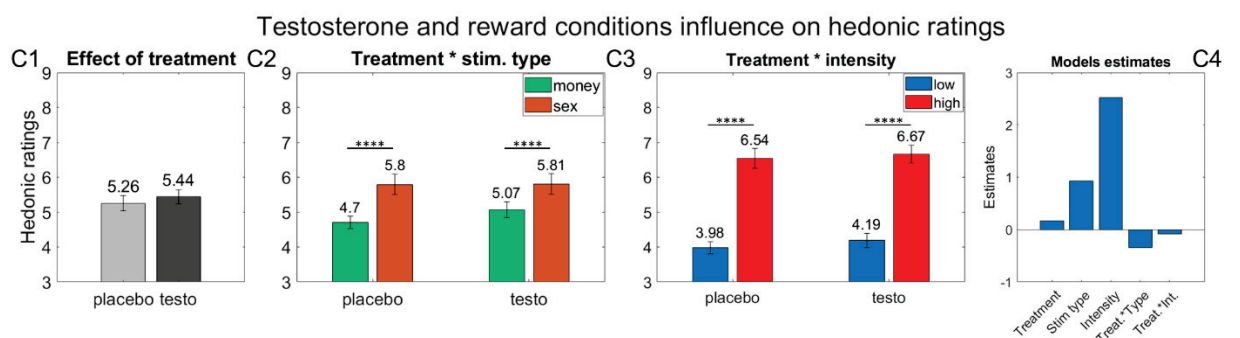
We performed GLME analyses on response times (in ms) data that were filtered for outliers via a Tukey fence procedure with a factor of 3. Treatments were treated as categorical (i.e., placebo or testosterone) and we included the number of missed trials and the number of outlying trials as fixed effects. An additive model showed that testosterone participants responded faster than placebos (est.=-36.3,  $p=.035$ ). We also found a reward type effect on RT during the discrimination task (est.=-19.2,  $p<.001$ ) and an interaction model revealed a treatment  $\times$  reward type interaction on RT during the discrimination task (est.=-16.2,  $p=.005$ ). These interactions reveal that participants respond differently to sexual stimuli, with faster RT. The latter interaction was driven by faster RT in the testosterone group for erotic compared to monetary rewards (Figure 3), suggesting that the testosterone group was more strongly motivated by erotic pictures than monetary gains in our experiment. There were also effects of intensity (est.=-17.5,  $p<.001$ ) and probability (est.=-.14,  $p=.044$ ) on RT, indicating that higher reward intensity and more likely rewards increased motivation. Moreover, there was a treatment  $\times$  probability interaction (est.=.3,  $p=.036$ ), reflecting that RT decreased more in the placebo than testosterone group with reward probability (Supplementary II, table S2).



**Figure 3: Testosterone and reward conditions influence on hit rates and response times.** A1) Treatment had no effect on subjects hit rates but A2) we found a difference

between placebo and testosterone response times ( $T=-2.11$ ,  $p=.035$ ). B1) There was no effect of reward type nor interaction effect of reward type and treatment on hit rates but B2) there was a reward type main effect ( $T=-6.64$ ,  $p<.001$ ) and an interaction effect with treatment ( $T=-2.8$ ,  $p=.005$ ) on response times. C1) We found an effect of probability ( $T= 3.88$ ,  $p<.001$ ) and an interaction effect of probability and treatment on hit rates ( $T= 3.47$ ,  $p<.005$ ). C2) We found an effect of intensity on response times ( $T=-6.04$ ,  $p<.001$ ) but no interaction with treatment. D1) The graph represents, respectively for D1) hit rates and D2) response times, estimates for treatment, probability and stimulus type in the additive model and interaction terms in the interaction models. Graph represent the mean  $\pm$  SEM.

An additive model revealed that the mean ratings were not different between the two treatments ( $p=.37$ ). However, there was an effect of reward type ( $est.=.93$ ,  $p<.001$ ) and an interaction model revealed a significant treatment  $\times$  reward type interaction ( $est.=-.34$ ,  $p=.003$ ), suggesting that the difference between erotic rewards valuation and monetary rewards valuation was higher for the placebo group (Figure 4). There was an effect of reward intensity on the ratings in the additive model ( $est.= 2.52$ ,  $p<.001$ ) albeit no treatment  $\times$  intensity interaction ( $p=.46$ ). This shows that the two intensity categories chosen a priori (high versus low) were effectively perceived by the subjects, and that this perception did not differ between groups (Figure 4). Finally, we observed no effect of probability on the ratings in the additive model ( $p=.43$ ), and no treatment  $\times$  probability interaction ( $p=.7$ ). This result illustrates an absence of decrease of hedonic ratings with increasing probability for both rewards in both groups. Overall these results show that the rating patterns of the testosterone and placebo groups did differ (Supplementary II, table S3).

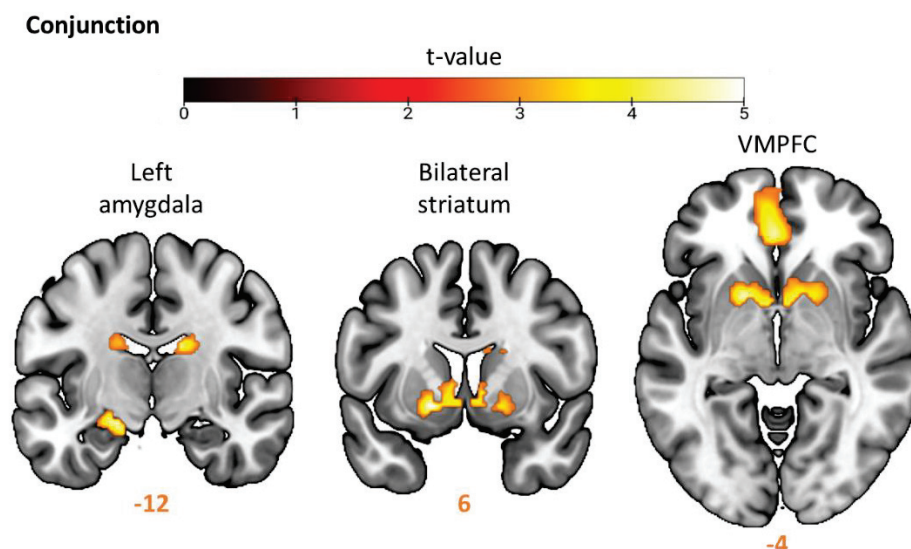


**Figure 4: Testosterone and reward conditions influence on hedonic ratings. A.** There was no difference between placebo ( $N=19$ ) and testosterone ( $N=21$ ) subjects' hedonic

ratings ( $p=.5464$ ). B. There was an effect of reward type ( $T=16.1$ ,  $p<.001$ ) and an interaction effect of reward type and treatment ( $T=-2.95$ ,  $p=.003$ ). C. We found an effect of intensity ( $T=43.9$ ,  $p<.001$ ) but no interaction effect of intensity and treatment ( $p=.46$ ). D. The graph represents estimates for treatment, stimulus type and intensity in the additive model, interaction between treatment and stimulus type and the interaction between treatment and intensity in the interaction models. Graph represent the mean  $\pm$  SEM.

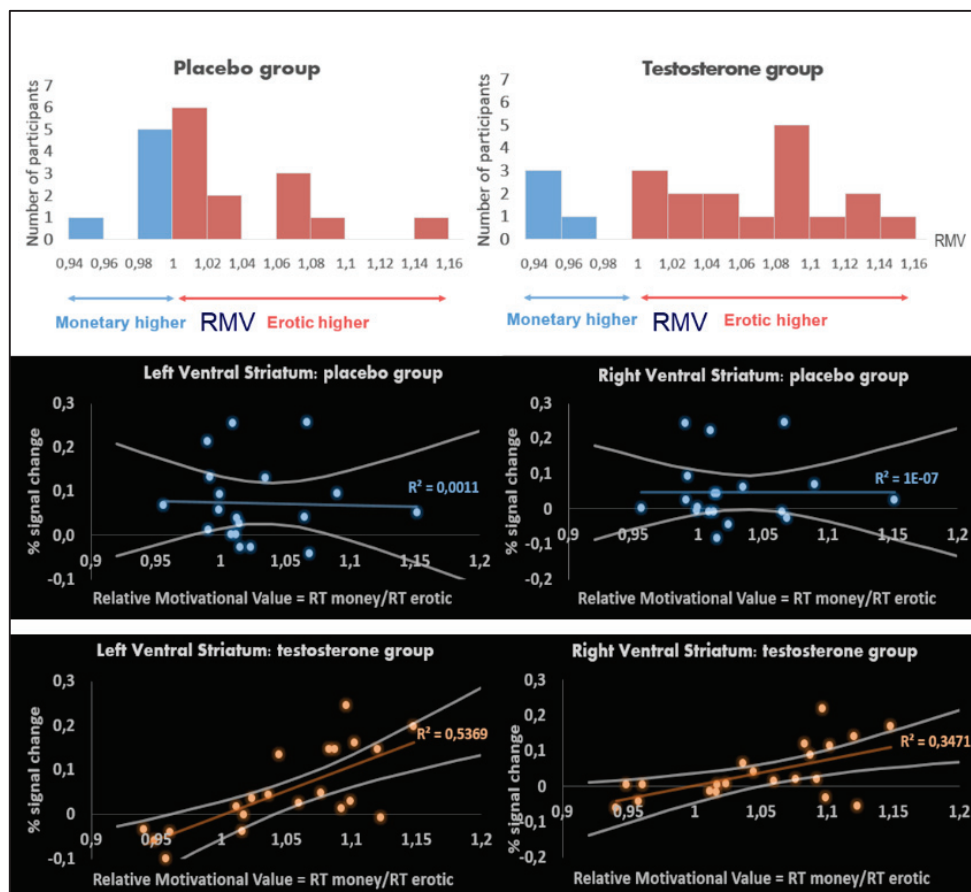
### III.4. Reward networks

The conjunction analysis (IE - IC with IM - IC) inclusively masked with bilateral amygdala, bilateral striatum and OFC ROIs showed that anticipating rewards significantly activated left amygdala (-14, -12, -18,  $k=75$ ,  $T(1,76) = 4.15$ ), right VS (16, -18, 22,  $k=85$ ,  $T(1,76) = 4.11$ ) and VMPFC (-2, 38, -4,  $k=113$ ,  $T(1,76) = 4.18$ ) in both treatment groups for both reward ( $p<0.001$  uncorrected) (Figure 5). The conjunction analysis yielded reward-type insensitive areas activated during anticipation of rewards in the lateralized portions of amygdala and VS. It didn't trigger insula nor thalamus activations.



**Figure 5: Brain regions that responded to both erotic and monetary rewards in both groups.** The regions yielded by the conjunction analysis ( $p<.001$  uncorrected) correspond to the regions common to the 4 different contrasts (E - C and M - C), regardless of treatment (display:  $p<0.005$  uncorrected t-values).

Although we couldn't find any direct treatment effect on specific regions under reward anticipation, calculating an index of the relative value of rewards yielded results regarding inter-individual differences. The Relative Motivation Value (RMV, Sescousse et al., 2014) indexes the reactivity for a type of reward ( $\frac{\text{monetary RTs}}{\text{erotic RTs}}$ ). A value greater than 1 reflects a higher relative value of erotic indices. The RMV correlated particularly well with the percentage of signal change in the bilateral VS [(G : (-12, 2, -8) ; D : (12, 0, -10)] in testosterone participants, while this is not the case in placebo participants. The RMV explains a significant proportion of the variance of the signal change variance in the left ( $R^2=0.5369$ ,  $F(1,19) = 22.02$ ,  $p < 0.001$ ) and right ( $R^2=0.3471$ ,  $F(1,19) = 10.1$ ,  $p < 0.005$ ) VS in testosterone participants, meaning a higher cue reactivity for erotic reward in the ventral striatum for the testosterone group. Comparatively, the RMV does not explain a significant proportion of the variance of signal change variance in the bilateral SV among placebo participants (Figure 6).



**Figure 6: Relative Motivation Value (RMV) reflects a higher cue reactivity for erotic reward in the ventral striatum for the testosterone group. The higher motivation**



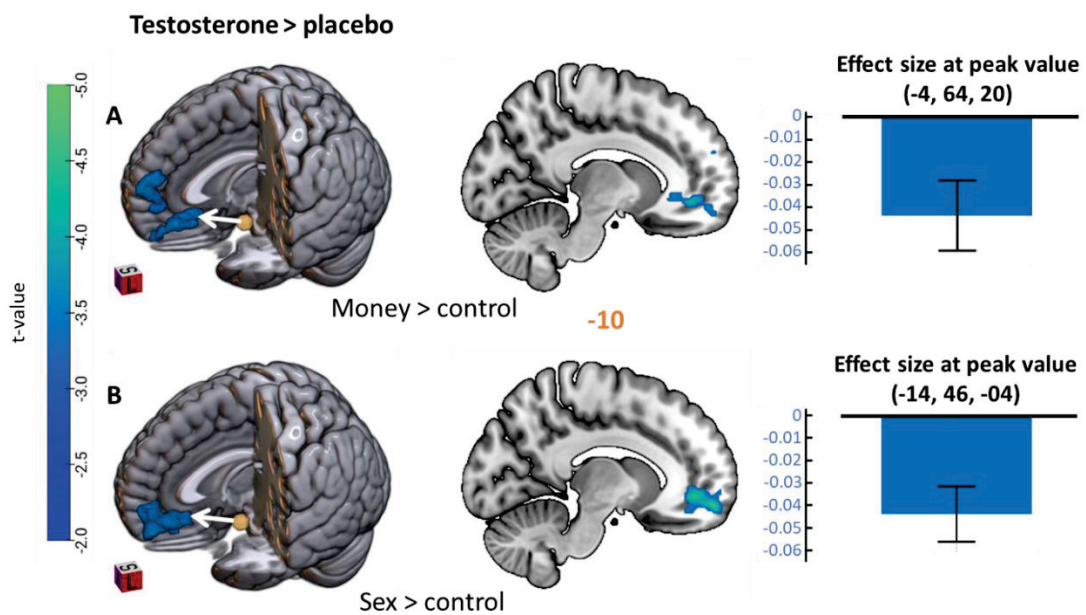
value for erotic rewards explains a significant proportion of the variance of signal change variance in the bilateral SV among testosterone participants.

### III.5. Influence of testosterone on functional connectivity

Although we did not find any effect of testosterone on specific regions, we examined the functional connectivity changes between left amygdala and the ventromedial portion induced by testosterone. We conducted a task-dependent analysis of the connectivity with a group contrast gPPI analysis (testosterone > placebo). We were interested in the 4 different contrasts (E - C and M - C in anticipation and reward phase).

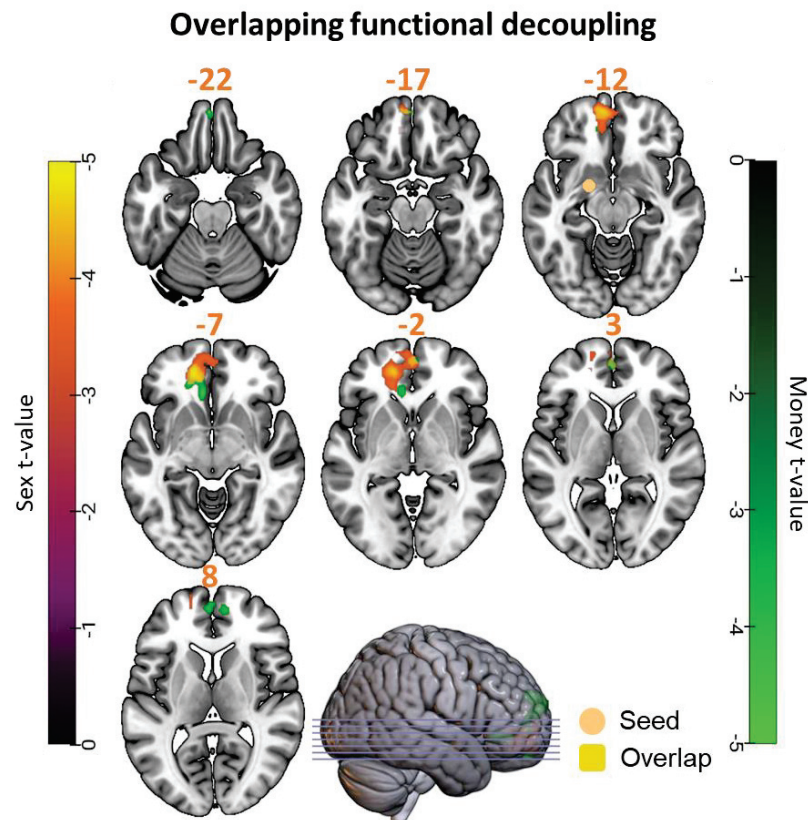
In the context of anticipating erotic reward, gPPI (IE-IC) analyses showed that activity in the first left amygdala ROI (-22, -8, -18) was not followed by task-dependent functional interaction with VMPFC. Nonetheless, the second left amygdala ROI (-18, -2, -12) was accompanied by task-dependent functional interaction with a large cluster of voxels including VMPFC [peak: (-6, 66, 18),  $k=103$ ,  $t(38)=-4.65$ ] in testosterone participants compared with placebos (Figure 7A). In the context of anticipating monetary reward, gPPI (IM-IC) analyses showed that the first left amygdala ROI (-22, -8, -18) was not followed by task-dependent functional interaction with VMPFC in testosterone participants, while the second left amygdala ROI (-18, 2, -12) was (VMPFC:  $k=156$ ,  $t=-4.28$ ) (Figure 7B). The reported values survived voxelwise  $p<.001$  uncorrected and clusterwise  $p<.05$  FDR-corrected. They show that testosterone reduced functional coupling between the left amygdala and the VMPFC in anticipation of reward for both types of stimuli in a lateralized portion of amygdala.





**Figure 7: Task-dependent interactions of VMPFC with left amygdala (-18, -2, -12) in erotic and monetary conditions for the contrast testosterone > placebo.** A. Task-dependent functional decoupling in the VMPFC in interaction with left amygdala in erotic condition (contrast IE - IC) and B. Task-dependent functional decoupling in the VMPFC in interaction with left amygdala in monetary condition (contrast IM - IC). All voxelwise  $p < .001$  uncorrected and clusterwise  $p < .05$  FDR-corrected. Copper-tinted spheres represent the left amygdala seed. White arrows illustrate task-dependent interactions between left amygdala and VMPFC.

A common region analysis of both contrasts with xjView indicates that, in testosterone participants, a left amygdala ROI (-18, -2, -12) consistently interacts regardless of reward-type condition with VMPFC [(-12, 66, -14),  $k=78$ ,  $z=1.04$ ,  $p < 0.001$  not corrected] in testosterone participants relative to placebos (Figure 8). The effect of testosterone on the reduction of functional coupling concerns the same regions in anticipation of erotic and monetary rewards. Note that FDR-correction does not apply to xjView's common region analysis.



**Figure 8: Task-dependent functional decoupling of VMPFC with left amygdala (-18, -2, -12) overlapping between erotic and monetary conditions in testosterone compared to placebo subjects. Overlapping functional decoupling shown in MricroGL software, voxelwise  $p < .001$  uncorrected clusterwise  $p < .05$  FDR-corrected.**

## IV. Discussion

### IV.1. Behavioral effects

The high percentage of accurate responses (above 90% in all conditions) indicates the participants were motivated to answer correctly and attempting to win the rewards. However, performance among the testosterone-treated group decreased with probability to obtain rewards. Their performance rate was lower than that for low, mid or high probability rewards in the placebo group and the performance drop was steeper for the lowest probability. Since there was no reduction in reaction time in the testosterone group associated with probability, the drop in performances for the lowest probability does not appear to be the result of increased impulsivity. Indeed, reaction times in this group were apparently at an optimally low level

throughout the task. One possible explanation of this observation is that testosterone subjects were willing to get faster to a trial with a cue indicating higher probability to obtain a reward. Looking at the control group, reaction times were not subject to variations until maximum probability was reached. At that point, participants were as fast as testosterone participants. Since this was not associated with any loss in accuracy we can only assume that highly motivated participants are able to perform with optimal accuracy at this reaction time. Results suggest there is a dissociation in behavior between both groups in regard to reward probability. This may reflect distinct consequences of reduced motivation or allocation of central nervous system resources in testosterone and placebo treated participants and this is consistent with known effects of testosterone in enhancing reactions to challenging situations.

There was no difference in the hedonic ratings between our groups suggesting that both groups gave the same value to the erotic and monetary rewards. Very interestingly, all the subjects gave higher hedonic ratings and had faster reaction times for erotic than for monetary rewards, while in the previous studies using this protocol, erotic and monetary rewards always had similar subjective values in the hedonic ratings and similar reaction times in healthy participants (Sescousse, Barbalat, Domenech, & Dreher, 2013; Sescousse et al., 2010). Exogenous testosterone had no effect on hedonic ratings, but there might be a placebo effect driving all the subjects towards a preference for the erotic more than the monetary stimuli and potentially hiding any testosterone effect on hedonic ratings.

#### IV.2. Reward networks

The brain activity dedicated to the processing of both reward indices was consistent with one of our previous studies (Sescousse et al., 2014). Whereas we did not find evidence of differences within this network attributable to direct effect of testosterone, examining the impact of the value of relative motivation (RMV) yielded VS activity differences between groups. This observation indicates that the activation of the VS in testosterone participants was mediated by the intrinsic motivation for rewards. This result, may be explained by the VS position in the mesolimbic pathway and helps to explain the optimal results of the testosterone participants during the behavioral task. It indicates as well that testosterone participants responded more to cues for erotic reward relative to placebo participants, which might explain the faster responses for erotic cues. The representation of subjective values is known to take place in the same area of VS for each subject, regardless of the reward type (Sescousse, Li, et al., 2013; Sescousse et al., 2010). The BOLD activity of the VS is also known to correlate

positively with the value and magnitude of rewards when making decisions or delivering primary and monetary rewards (Bartra et al., 2013; Kable & Glimcher, 2007; Knutson, Rick, Wimmer, Prelec, & Loewenstein, 2007b). VMPFC on the other hand is known to integrate reward values through several dimensions and reward types (Levy & Glimcher, 2012; Metereau & Dreher, 2013). Additionally, VS results add evidence that erotic rewards are more prominent than monetary rewards in the brain, possibly due to the more powerful biological role of sex in relation to money.

Although we were unable to find any effect of testosterone on the regions of the network activated during the reward anticipation, we found a treatment effect on the functional connectivity of both amygdala and the ventromedial portion of OFC. This change results in a negative correlation between the two regions. This correlation concerns the same location of the left amygdala (-18, -2, -10) and the VMPFC (-12, 66, 14) regardless of the type of reward. Studies on testosterone report a decoupling between the left amygdala and the OFC (P. A. Bos, Hermans, et al., 2012; Spielberg et al., 2013; G. van Wingen et al., 2010), extending over a large region of the mPFC (respectively: [-16, 32, -14] ; [-24, 58, -2] and [34, 48, -16]). As testosterone participants responded as quickly as possible for all rewarded trials, involving 25% chance of winning trials, it is possible that deliberation processes leading to anything else than impulsive response are suppressed, resulting in the implementation of efforts leading to victory. Amygdala receives numerous projections of cortical areas, including VS and OFC (Carmichael & Price, 1995; Haber & Knutson, 2010; E. A. Murray, 2007). Its position in the reward network could consist in implementing decisions and directly transforming value signals into choices that guide actions (Grabenhorst & Rolls, 2011), while VS role could consist in updating prediction errors (Oldham et al., 2018).

### IV.3. Conclusion

We see many limitations in this study's parameters that may have participated in diminishing the protocol's power. Intergroup differences may have intervened as our testosterone group had a slightly lower score than placebos on the SADI questionnaire assessing the overall sexual arousal status of participants. A between-subjects design alternating random injections of placebo and testosterone could yield more robust results. In this sense, the simple fact of passing a test may have raised testosterone levels, as has been shown in men (Carré et al., 2011), masking some results. In hamsters, intracerebral-ventricular testosterone injection induces greater basal activity in the ATV as indicated by Fos staining (Dimeo & Wood, 2006;

Nagypál & Wood, 2007). As in experiments with testosterone self-administration in rodents (Johnson & Wood, 2001; Wood, Johnson, Chu, Schad, & Self, 2004), experiments with conditioned place preference (Frye et al., 2002; Packard et al., 1998; Teresa Arnedo et al., 2002) or studies on ASA abuse in humans, this suggests that testosterone is rewarding in itself. It is possible that in addition to being a cause of attraction to sexual rewards, testosterone itself may have acted as a positive enhancer. Its increase in anticipation (Graham & Desjardins, 1980) and/or during and/or after a sexual encounter (Dabbs & Mohammed, 1992) would induce a rewarding increase in activity in the mesolimbic pathway. From this point of view, the perception of decreased sexual motivation in hypogonadic patients may be related to the fact that, without testosterone, sexual thoughts and actions are no longer rewarding. It would be particularly interesting to compare placebo and testosterone participants during a rest state, since the baseline used for fMRI analyses would already show some differences between participants, particularly in the mesolimbic pathway. Finally, many researchers believe that a low concentration of testosterone is already sufficient to ensure its physiological effects on motivation and reward. A meta-analysis revealed that testosterone-induced improvement in libido would eventually be restricted to hypogonadal and eugonadal men with a total testosterone concentration, baseline at 9am, below 12nmol/L (Isidori et al., 2005), while our participants were already above this critical value before the injection. The physiological testosterone concentrations of healthy young men in our study may already be sufficient to ensure optimal motivation and reward experience for erotic stimuli. It would then be interesting to conduct the same study in older men with lower initial testosterone levels.

However, we found an interesting effect of testosterone on the activity of the anticipating reward network. We found a negative effect of testosterone on the coupling between the amygdala and the prefrontal cortex, consistent with the literature, when anticipating rewards. This activity is also consistent with testosterone-induced reward dependence. The value of rewards would be updated upon receipt of the reward, creating a dependency upon presentation of the reward indices. This dependence would result in an impulsive response to these indices, as shown by behavioral results. It is possible testosterone decreases the regulation of amygdala activity by affecting prefrontal activity in a dynamic event-wise way, depending on stimuli presentation. It would be interesting to look at the testosterone concentration levels during the task and the presence of dynamic variations. It has been shown that changes in testosterone concentration levels, rather than the baseline, are responsible for modulating behaviours such as aggression or competitiveness (Carré, Campbell, Lozoya, Goetz, & Welker,

2013). Finally, stress over-activates the salience network and induces an increase in amygdala responses (Oei et al., 2012; van Marle, Hermans, Qin, & Fernández, 2010) while and suppressing responses from the executive control network (Ossewaarde et al., 2011). Testosterone could similarly induce a preparatory effect for a reflex response when anticipating a reward.

To our knowledge, this is the first study that directly addresses the effect of testosterone on the common reward network. Although the results are encouraging, we believe it is necessary to continue to investigate the question of a dynamic change in the mode of action of testosterone and its effect on reward circuits.



# General conclusion

Overall this thesis work represents a contribution in the understanding of the processing of information. In the first chapter, we investigated the mechanisms of information-seeking. We designed a new behavioral economics experiment to test the relationships between the evaluation of news, the confidence in this evaluation and information-seeking. Our findings indicate that the uncertainty surrounding news impact the way individuals estimate the truthfulness of news. More precisely, the more the content of the news was imprecise, the more they declared the news to be false. The results also showed that participants' metacognitive abilities were not accurate in the task. The confidence they had in their estimation of the truthfulness was not predictive of their ability to discern truth from falsehood. This indicates that participants had an inaccurate representation of their knowledge. In spite of that, they relied upon their metacognition to choose which extra information about news they wanted to receive. The less they were confident in their truthfulness estimation, the more they chose to receive extra-information acquisition. These findings demonstrate a key role of metacognitive monitoring in fake news evaluation and characterize an important role of the estimation of truthfulness in information-seeking.

In the second chapter we investigated the mechanisms and neurocomputations underlying the inference of others' preferences for seeking information. Participants in a fMRI machine evaluated the same news as those in the first chapter. They played the role of Senders and had to predict, for each news, whether the participants from the first chapter (Receivers) chose to receive or not extra information about the news. This study showed that when Senders were provided no information about Receivers' beliefs, they chose to send the more information when they were the less confident in their truthfulness estimation. However, they overestimated Receivers' preferences. When they were provided with information about Receivers' beliefs, they improved their ability to predict Receivers' preferences. The results suggest that participants performed better when they shifted their weight in favour of their beliefs about Receivers' beliefs and stopped using their beliefs about information. Furthermore, the study revealed that the valuation network was more engaged in the decisions to send extra information when they based their decision on their truthfulness estimation. This valuation network was composed of the midbrain, the striatum, the ventromedial prefrontal cortex (vmPFC) and the dorsolateral prefrontal cortex (dlPFC). Results show that the activity within the striatum, the vmPFC and the dlPFC correlated with beliefs about Receivers' preferences. When participants



based their decision on their beliefs about Receivers' beliefs, a social processing network composed of the dorsomedial prefrontal cortex (dmPFC), the superior temporal sulcus (STS) and the temporo-parietal junction (TPJ) was more engaged. The activity within the TPJ correlated with beliefs about Receivers' beliefs. This chapter highlights the brain systems engaged in the processing of information and in the updating of beliefs about others' preferences for information.

In the third chapter, we studied the activity within the reward valuation network at the anticipation of primary and secondary rewards. More specifically, we investigated the effect of testosterone on the brain activity and connectivity within the valuation network. Participants in a fMRI machine were shown cues about upcoming rewards that varied in probability and magnitude. After a discrimination task, participants received the outcome. We found that anticipating monetary and secondary rewards elicits brain activity within the ventral striatum, the vmPFC and the amygdala. In particular, the ventral striatum reacted more to the anticipation of erotic rewards. Most importantly, we found that the testosterone reduced the connectivity between the amygdala and the vmPFC. In testosterone participants, at the anticipation of both rewards, the activity within the vmPFC decorrelated with the increase of activity within the amygdala. This chapter highlight the brain network engaged when anticipating rewards. It reveals the similitude within the brain between processing primary rewards, secondary rewards and information. The vmPFC is known for encoding the subjective value of rewards in a common neural currency while the evaluation of information elicits activity within the vmPFC. Testosterone is known for its effect on rewards processing. Given the effect we found on the connectivity between the amygdala and the vmPFC, we can hypothesize that these hormones impact the processing of information, hence impact how we form beliefs about the state of the world.

In summary, by combining behavioral economics and neuroimaging (model-based fMRI), we investigated the mechanisms and brain systems engaged in the decisions to seek information and the processes of inferring others' preferences for information. We manipulated true and false non-ego relevant information of cognitive utility and followed the guidelines developed in previous studies on individuals' behaviors toward false information. Through the incentive structure of our experimental designs, we have shown the importance of participants' beliefs about their own knowledge. Thus, we contribute to the literature on explanatory factors of information seeking by going beyond the framework of motivated beliefs, heuristics and

biases or depth of reasoning. We also contribute to the literature on the processes underlying preference inference by extending it to the domain of information and more specifically to media information. These results have implications for understanding how we exchange information and for potential solutions to fight misinformation. For example, they illustrate the importance of one's own perceived knowledge in the pursuit of information and a potential solution by helping individuals recalibrate these perceptions. Furthermore, the experimental designs constructed for these studies can be easily used to investigate other information seeking and sharing behaviors. For instance, the incentive structure can be modified to study dishonest misinformation behavior. The nature of the incentives can also be modified to apply our models to information whose utility may be instrumental to the receivers in a second phase of experimentation. There are, however, limitations to our contribution. In particular, the models of preference inference that we have tested do not represent the full range of existing models in theory of mind that offer alternative solutions to the inference problem. Nevertheless, the results we report raise broader questions about the similarity between reward processing and the processing of information, such as news, that has non-instrumental utility. It also raises questions regarding the neuropsychopharmacology of information processing.

# Conclusion générale

Cette thèse contribue à la compréhension du traitement de l'information. Dans le premier chapitre, nous avons étudié les mécanismes de la recherche d'information. Nous avons conçu une nouvelle expérience d'économie comportementale pour tester les relations entre l'évaluation de la véracité de nouvelles, la confiance dans cette évaluation et la recherche d'information supplémentaire susceptible de réduire l'incertitude. Nos résultats indiquent que l'incertitude entourant les brèves informations a un impact sur la façon dont les individus estiment la véracité des informations. Plus le contenu des brèves était imprécis, plus les participants la déclaraient fausse. Les résultats ont également montré que les capacités métacognitives des participants étaient imprécises dans cette tâche. La confiance qu'ils avaient dans leur estimation de la véracité n'était pas prédictive de leur capacité à discerner les vraies brèves des fausses. Cela suggère que les participants avaient une représentation inexacte de leurs connaissances. Malgré cela, ils se sont appuyés sur leur métacognition pour choisir les informations supplémentaires sur les brèves qu'ils souhaitaient recevoir. Moins ils étaient confiants dans leur estimation de la véracité, plus ils choisissaient de recevoir des informations supplémentaires. Ces résultats démontrent un rôle clé de la métacognition dans l'évaluation des vraies et fausses informations et caractérisent un rôle important de l'estimation de la véracité dans la recherche d'informations.

Dans le deuxième chapitre nous nous sommes intéressés aux mécanismes et calculs cérébraux qui sous-tendent l'inférence des préférences des autres pour la recherche d'informations. Des participants placés dans un IRMf ont évalué les mêmes brèves que celles du premier chapitre. Ils jouaient le rôle d'émetteurs et devaient prédire, pour chaque brève, si les participants du premier chapitre (récepteurs) avaient choisi de recevoir ou non des informations supplémentaires sur cette brève. Cette étude a montré que lorsque les émetteurs ne recevaient aucune information sur les croyances des récepteurs, ils choisissaient d'envoyer plus d'informations lorsqu'ils étaient moins confiants dans leur estimation de la véracité. Cependant, les préférences des récepteurs pour recevoir l'information supplémentaire étaient surestimées. Lorsqu'il leur a été fourni des informations sur les croyances des récepteurs, ils ont amélioré leur capacité à prédire les préférences des destinataires. Les résultats suggèrent que les participants ont obtenu de meilleurs résultats lorsqu'ils ont attribué plus de poids à leurs croyances de second ordre (sur les croyances des récepteurs) et ont cessé d'utiliser leurs croyances sur les informations. En outre, l'étude a révélé que le réseau d'évaluation était plus engagé dans les décisions d'envoyer des informations supplémentaires lorsqu'ils fondaient leur

décision sur leur estimation de la véracité. Ce réseau d'évaluation était composé du mésencéphale, du striatum, du cortex préfrontal ventromédial (vmPFC) et du cortex préfrontal dorsolatéral (dlPFC). Les résultats montrent que l'activité dans le striatum, le vmPFC et le dlPFC était corrélée aux croyances sur les préférences des récepteurs. Lorsque les participants fondaient leur décision sur leurs croyances concernant les croyances des récepteurs, un réseau impliqué dans les processus sociaux composé du cortex préfrontal dorsomédial (dmPFC), du sillon temporal supérieur (STS) et de la jonction temporo-pariétale (TPJ) était plus engagé. L'activité au sein de la jonction temporo-pariétale était en corrélation avec les croyances sur les croyances des récepteurs. Ce chapitre met en évidence les systèmes cérébraux impliqués dans le traitement de l'information et dans la mise à jour des croyances concernant les préférences des autres pour l'information.

Dans le troisième chapitre, nous avons étudié l'activité au sein du réseau d'évaluation des récompenses lors de l'anticipation de récompenses primaires et secondaires. Plus précisément, nous avons étudié l'effet de la testostérone sur l'activité cérébrale et la connectivité au sein de ce réseau. Des participants placés dans un IRMf ont reçu des indices sur des récompenses à venir dont la probabilité et la magnitude variaient. Après une tâche de discrimination, les participants ont pu recevoir une récompense suivant la probabilité de réception. Nous avons constaté que l'anticipation de récompenses monétaires et secondaires a suscité une activité cérébrale dans le striatum ventral, le vmPFC et l'amygdale. En particulier, le striatum ventral a réagi davantage à l'anticipation de récompenses érotiques. Plus important encore, nous avons constaté que la testostérone réduisait la connectivité entre l'amygdale et le vmPFC. Chez les participants sous testostérone, lors de l'anticipation des deux récompenses, l'activité dans le vmPFC était décorrélée de l'augmentation de l'activité dans l'amygdale. Ce chapitre met en évidence le réseau cérébral engagé lors de l'anticipation d'une récompense. Il révèle la similitude dans le cerveau entre le traitement des récompenses primaires, des récompenses secondaires et des informations. Le vmPFC est connu pour encoder la valeur subjective des récompenses dans une monnaie neuronale commune tandis que l'évaluation d'informations suscite l'activation du vmPFC. La testostérone est connue pour son effet sur le traitement des récompenses. Compte tenu de l'effet que nous avons constaté sur la connectivité entre l'amygdale et le vmPFC, nous pouvons faire l'hypothèse que ces hormones ont un impact sur le traitement de l'information, et donc sur la façon dont nous formons des croyances sur l'état du monde.

En résumé, en combinant l'économie comportementale et la neuro-imagerie (IRMf basée sur des modèles), nous avons étudié les mécanismes et systèmes cérébraux engagés dans les décisions de recherche d'information et dans les processus d'inférence des préférences des autres pour l'information. Nous avons manipulé des informations vraies et fausses d'utilité cognitive et non relatives à l'ego en respectant les lignes directrices développées dans de précédentes études sur les comportements des individus face aux fausses informations. Par la structure incitative de nos designs expérimentaux, nous avons montré l'importance, dans ces comportements, des croyances des participants sur leurs propres connaissances. De ce fait, nous contribuons à la littérature sur les facteurs explicatifs de la recherche d'informations en sortant du cadre des croyances motivées, des heuristiques et biais et de la profondeur de raisonnement. Nous contribuons également à la littérature sur les processus sous-jacents à l'inférence des préférences en l'étendant au domaine des informations et plus particulièrement celles issues de médias. Ces résultats ont des implications sur la compréhension de nos échanges d'informations et de potentielles solutions pour lutter contre la désinformation. Par exemple, ils illustrent l'importance que joue la perception de ses propres connaissances sur la poursuite d'informations et une potentielle solution en aidant les individus à recalibrer ces perceptions. De plus, les designs expérimentaux construits pour ces études peuvent être aisément utilisés pour investiguer d'autres comportements de recherche et de partage d'informations. Par exemple la structure d'incitations peut être modifiée pour étudier les comportements malhonnêtes de désinformation. La nature des incitations peut également être modifiée pour appliquer nos modèles à des informations dont l'utilité peut s'avérer instrumentale pour les récepteurs dans une seconde phase d'expérimentation. Il existe cependant des limites à notre contribution. Particulièrement, les modèles d'inférence des préférences que nous avons testés ne représentent pas l'ensemble de la gamme de modèles existants en théorie de l'esprit et offrant des solutions alternatives au problème de l'inférence. Néanmoins, les résultats que nous rapportons soulèvent plus largement des questions concernant la similitude entre le traitement des récompenses et le traitement d'informations à utilité non instrumentale. Cela soulève également des questions concernant la neuropsychopharmacologie du traitement de l'information.

# Chapter I. Supplementary Materials

## I. Preliminary experiment: stimuli rating instructions

Welcome to this session.

Please, turn off your phone.

You will be participating in a session consisting of 42 periods.

This session will last approximately 40 minutes.

In each period, you will see a short piece of information. For each brief news, you will be asked to answer five questions. The first four questions will be, for each brief, mandatory.

The brief news will be different in each period.

Some of the brief news are from the French written press from the years 2017-2020. Others have been fabricated.

Unless specified otherwise, their content is recent, topical and concerns the French territory.

Please read each brief and each question carefully, then take the time you need to answer.

## II. Main experiment: task instructions

Welcome to this session.

Please, turn off your phone.

Then, please read the following instructions carefully. They will give you all the information you need to participate in this session.

**Please note: The Testable.org platform will expire suspicious activity (window change, inactivity, excessively long duration). Once an entry has expired, it is no longer possible to re-enter.**

**Please click OK to continue.**

You will receive **\$9** for participating in the session.

Depending on your decisions, you may earn Experimental Currency Units (ECU). These ECU will be converted into US Dollars at a ratio of 100 ECU = \$2.

**At the end of the session, you will receive the monetary amount equivalent to the conversion of the ECU you earned plus the \$9 for your participation.**

This session is composed of **two parts**. **Your answers during these two parts will be anonymous.**

They may be used in a future experiment. Nothing will identify you.

The instructions you will receive now are those for the first part.

You will receive the instructions for the second part at the end of the first part.

All the instructions will be displayed on your screen.

### 1. Part 1

The first part of this experiment involves evaluating 12 different organizations using six criteria. You will evaluate each criterion using the information you are given. You can also use your personal knowledge.

You will first learn about the organization to be evaluated and a summary of their purpose from their website, using the original wording, as in the example below:

#### **La Pétanque Carryenne**

Depuis plus de soixante ans, la Pétanque Carryenne existe sur la commune de Carry-le-Rouet. Autrefois, les concours se faisaient sur la place du marcé et au Family (aujourd'hui disparu). Pas moins de 10 présidents se sont transmis le flambeau afin que perdure la pratique des jeux de boules ; tant pétanque que jeu provençal.

Le terrain actuel de la Pétanque Carryenne a été construit dans les années 70 et le nombre de licenciés n'a cessé d'augmenter. Ce club a eu l'honneur d'accueillir en son sein par trois fois des équipes championnes de France à pétanque en catégorie vétérans et des vainqueurs du prestigieux concours Mondial La Marseillaise à Pétanque.

Actuellement, le club compte près de trois cents membres et depuis trois ans, une école de pétanque permettant de transmettre aux nouvelles générations la passion des sports de boules. Et comme le dit la devise du club « Ici, la pétanque est plus qu'un jeu, c'est une religion ! » C'est dans un cadre exceptionnel, au cœur de la ville et à deux pas de la grande bleue que les Carryens vous accueilleront chaque jour pour des parties passionnantes et conviviales.

#### **Organizations rating**

Then, you will evaluate this organization by answering the following 6 questions using a 7-point response scale:



- How familiar is this organization to you?
- How close do you feel the values of this organization are to your own?
- How much do you like this organization?
- How familiar do you think this organization is to those closest to you?
- How close do you think your family and friends feel that the values of this organization are to their own?
- How much do you think your loved ones would appreciate this organization?

For example, below, 1 means that the organization is 'Not familiar at all' and 7 means that the organization is 'Very familiar':



These evaluations are done privately.

Some of them may be shared with other participants in a future experiment, but they will not be associated with any element that could identify you.

### The first part will now begin

Throughout the session, use your computer mouse to respond.

When you are ready, please click OK.

### First part completed

The first part is now complete. The instructions for the second part will begin.

Please read them carefully.

They will outline the rest of the session and explain how you will be compensated based on your performance.

To view them, please press "OK".

## 2. Part 2

At the beginning of each period, you will see a brief news. These briefs will each be different and come from different media. For each brief, there is a supplementary information.

**The supplementary information consists of an investigative file.** Their content is related to the content of the brief and can be related to other information around the content of the brief.

**Please note: briefs can be true or false. Some of the stories you will be exposed to and play with will be false information.**

False information has been fabricated. The non-fabricated news briefs are from the French print media over the period 2017 - 1st quarter 2020. Unless specified otherwise, their content is recent, topical and concerns the French territory.

Each period consists of two steps: the evaluation of the truthfulness of the news item and the decision to receive or not more information.

### First step: news evaluation

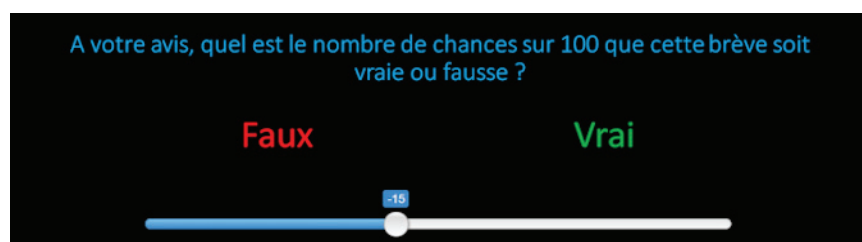
After the brief information is displayed, you will be asked to rate whether its content is true or false as follows:

"In your opinion, what is the number of chances out of 100 that this brief is true or false?".

To answer, you will drag the slider below the question, as in the example below:

- To answer that the content of the brief is false, you will drag the slider between the 0 and -100 bounds.

- To answer that the content of the brief is true, you will drag the cursor between the 0 and +100 bounds.



### First step: robots to help you answer

To help you answer this question, there will be "bots" available during this step.

There are 100 different robots. Each robot has a certain accuracy level.

This accuracy level is the number of chances out of 100 that the robot correctly evaluates the brief news.

This number is an integer between 1 and 100. Each robot has a different accuracy level than the other robots.

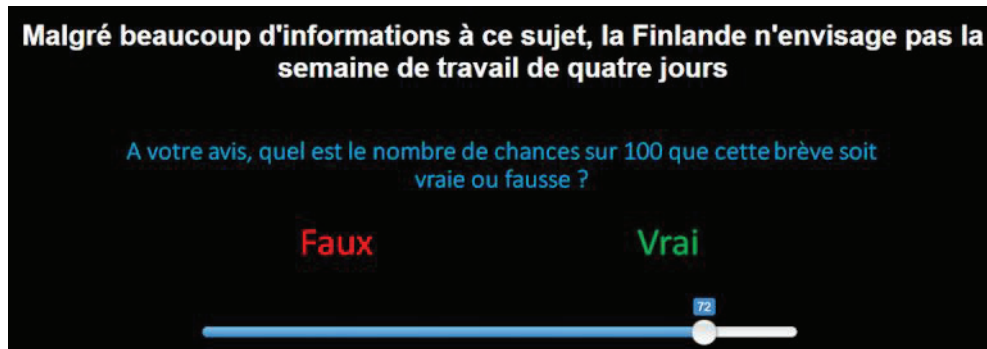
That is, there is one robot that has a 1 in 100 chance of answering correctly, there is one robot that has a 2 in 100 chance of answering correctly, and so on until the 100th robot that has a 100 in 100 chance of answering correctly.

A robot that has 75 chances out of 100 to answer correctly will give a correct answer 75% of the time and will give a wrong answer 25% of the time.

At each period, the computer will draw one robot from the 100 robots. All robots have the same chance of being drawn. You will not know which one has been drawn and the robot will change randomly each period.

When providing your answer about the truthfulness of the brief, you will have to specify which bots you would let answer for you.

**To do this, you will first move the cursor between the bounds corresponding to your assessment of truthfulness ('False'/'True'). Then, you decide how confident you are in your answer by choosing an accuracy threshold with the slider, as in the following example:**



The accuracy threshold determines the threshold at which you would prefer the program to consider a robot's response rather than yours. The program will take the robot's answer into **account if and only if** its accuracy level is higher than the threshold you have chosen.

Thus, you will choose your accuracy threshold so that, for any robot with an **accuracy level less than or equal to your threshold**, you would prefer it to be **your answer** that the program takes into account.

For example:

- If you choose 75 as your threshold and the randomly selected robot has an accuracy level of 90, the program will consider that robot's answer. The robot will have a 90 out of 100 chance of giving the right answer.

- If you choose 75 as the accuracy threshold and the randomly selected robot has an accuracy level of 20, the program will consider your answer.

Thus, it is in your best interest to truly state how correct you think your answer is.

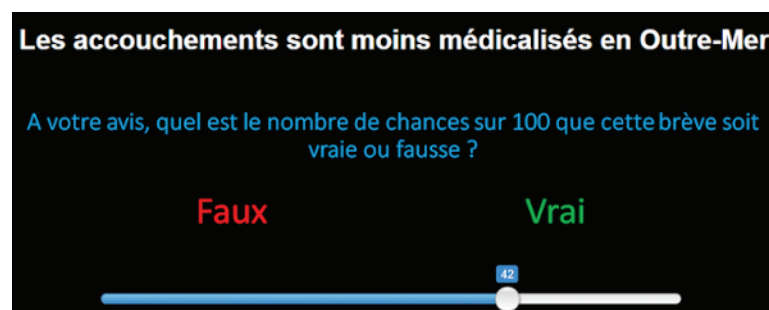
The less certain you are of your answer, the better it is to choose a low threshold.

That is, if you think there is a 25 in 100 chance that your answer is correct, it is better to give a threshold of 25.

In the following example, there are two decisions:

1) The participant answers that the content of the brief is true. He/she does this by positioning the cursor between the bounds 0 and 100. These bounds correspond to the answer 'True' to the question about truthfulness.

2) He/she would prefer that the program considers the robot's answer if and only if the robot drawn has more than 42 chances out of 100 to answer correctly. Therefore, he/she thinks that there is a 42 out of 100 chance that the brief is true.



**Step 1: Compensation based on your performance.**

**Eight** of your assessments will be **randomly** selected at the end of the session.

**Your compensation will be determined based only on the randomly selected assessments.**

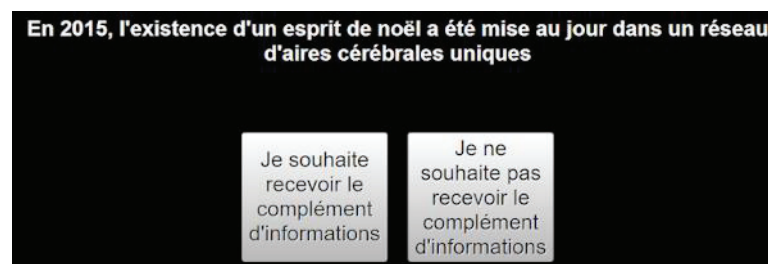
Each correct randomly selected assessment will earn you 50 ECU.

**Please note: 0 is not a valid answer. Any time you answer with 0 will automatically be considered a failure.**

### Step 2: Receiving more information

In each period, once you have evaluated the brief, you will be asked **whether you would like to receive or not more information** about the content of the brief. You will do this by positioning the cursor on the desired answer, as in the example below.

**Eight** periods will be drawn at the end of the session, each with the same number of chances. Your choices in these eight periods will help determine your earnings.



### Step 2: Willingness to receive or avoid more information

Once you have chosen, you will be asked how many ECU you would be willing to deduct from your initial endowment to have your decision implemented.

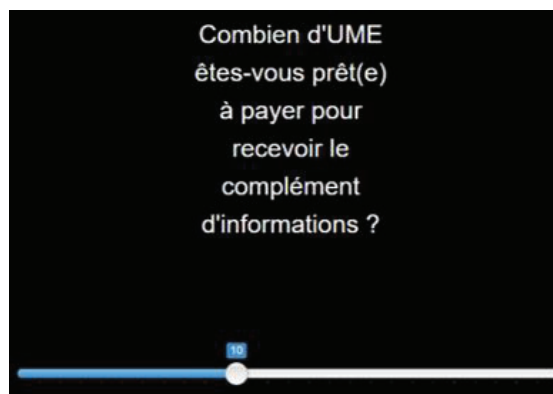
**You start the session with 200 ECU.**

If you have indicated that you would like to receive more information, you will be asked how many ECU you would be willing to deduct from your initial endowment to **receive** the additional information.

**You will be asked to choose how many ECU, between 0 and 25 ECU, you would be willing to deduct from your initial endowment.**

To indicate the desired value, you will move the cursor between the values 0 and 25, as in the example below.

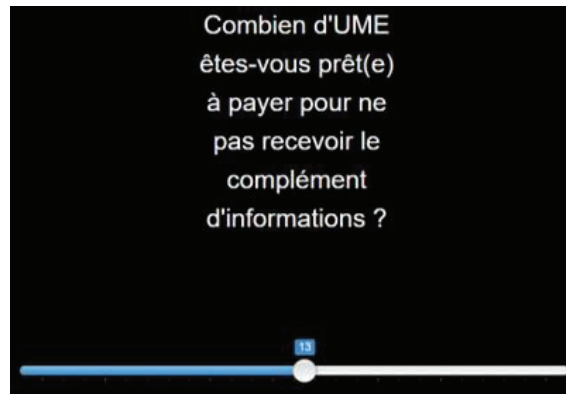
Attention: the position of the buttons "I wish to receive the additional information" and "I do not wish to receive the additional information" will be modified during the second half of the session.



If you have indicated that you would not like to receive more information, you will be asked how many ECU you would be willing to deduct from your initial endowment to not receive the top-up.

**You will be asked to choose how many ECU between 0 and 25 ECU you would be willing to deduct from your initial endowment.**

To indicate the desired value, you will move the cursor between the values 0 and 25, as in the example below:



### Step 2: Receiving more information and payment based on your decisions

**Eight** of your decisions, separate from the eight briefs evaluation periods, will be **randomly** selected at the end of the session.

They will determine whether or not you receive more information at the end of the session.

For each decision period selected at random, we will draw a number  $Y$  between 0 and 25. Your choice will be implemented if the drawn number of ECU  $Y$  is **less than or equal** to the number  $X$  of ECU you are willing to pay. **Your initial allocation will then decrease by  $Y$  ECU, the number drawn at random.**

Choosing 15 ECU means that you are willing to pay up to 15 ECU for your decision to be implemented.

### Step 2: Implementing your decisions

**Depending on your choices during the randomly selected periods at the end of the session, you will actually receive additional information in the days following your participation.**

That is to say, there is an information supplement for each brief. The supplementary information consists of an investigation file. Their content is related to the content of the news item and can also be related to other information around the content of the news item.

**For each randomly selected period, your choice of reception will be retained and implemented. The implementation consists in deducting from your initial endowment the amount of ECU and in sending you by email the files associated with the selected information.**

The implementation of your choice will depend on your choice (to receive or not to receive) and the amount of ECU you are willing to pay.

**Your decisions about whether or not to receive more information will therefore impact your reward at the end of the task in terms of information and ECUs received.**

### Calculating your earnings

**A. Information evaluation:**

At the end of the experiment, eight periods will be drawn from the 48 periods in Part 2. Each period will have an equal chance of being drawn.

A correct answer, either yours or that of the robot, will earn you 50 ECU. An incorrect answer will earn you 0 ECU.

**B. Receiving information:**

At the end of the session, eight more periods will be drawn from the 48 periods in Part 2. Each period will have the same chance to be drawn.

For each period where the number Y drawn is less than or equal to X, the number of ECU you are willing to pay, Y ECU will be deducted from your initial endowment.

**C. Final Compensation:**

ECU will be converted into U.S. Dollars at a ratio of 100 ECU = \$2.

The compensation for your performance and the compensation for your response to the questionnaires will be added to your initial \$9.

Before you begin the second part, you will complete the following comprehension quiz. It will begin with questions about evaluating the news.

Once you are ready, please click OK to continue.

### III. Organizations summaries

#### 1. Democracy-related organizations

##### **France FREXIT**

France FREXIT is a private and independent initiative, created in March 2018 and aiming at informing, gathering, exchanging, proposing on the theme of Frexit. In all legality, without any violence, and in the respect of the Institutions. Taking advantage of the historical opportunity of the Frexit, FRANCE FREXIT proposes the complete reform of the Republic and the French State by a great FRENCH NATIONAL COORDINATION, based on a new formula of Power, including a truly democratic organization taking the best of the Republic and leaving the least good, and using modern means of communication: votes, electronic votes, draws, as well as some monarchical aspects, mainly in terms of spirituality and transmission of universal and traditional values of France. FRANCE FREXIT is politically opposed to the euro-extremist and euro-identitarian parties such as LaREM, LR, MODEM, EELV, PS, UDI, Parti Radical. FRANCE FREXIT does not support the euro-alternative parties and other decoy parties such as the RN, the LFI or DLF, the NPA, LO and some other euro-compatible parties.

##### **Parti Libertarien**

Our observation is the same as the majority of French people: the weight of the state and its scope of action are constantly increasing, hindering our freedoms more and more, with the catastrophic results that everyone can see. We are also facing an extremely worrying legislative inflation, the accumulation of standards, regulations and laws make the system incomprehensible for the majority of French people and impracticable for entrepreneurs. Only large groups and multinationals benefit from this complexity and can expand without real competition. These regulations paralyze any personal initiative, block the social elevator and increase inequalities. Beyond the dramatic consequences of such an intrusion of the state in the life of individuals, the latter, by wanting to regulate every aspect of our daily life, goes beyond its prerogatives and violates our fundamental right to manage our life as we see fit. We are libertarians and we consider that it is up to free individuals to write their own history.

##### **Le Mouvement Européen – France**

The European Movement - France has been mobilizing since 1950, across all generations, to bring to life a pluralist public debate on Europe. It deploys its activities around pedagogy, the organization of debate between citizens and the formulation of proposals to build Europe. It gathers thousands of volunteers gathered in more than 50 local sections, about twenty member organizations as well as a college of qualified personalities. Heir to the spirit of the founding fathers of Europe, the Movement is the first actor of civil society in France on European issues. It also mobilizes through its youth branch, the Young Europeans - France, which has 26 local groups throughout the territory. At the European level, the European Movement is also a member of the European Movement - International, a network of 39 organizations that make our commitment resonate throughout the continent. The European Movement-France is recognized as an association of general interest since July 22, 2016 and is also approved as an "educational association complementary to public education".

##### **Fondation Robert Schuman**

Created in 1991 and recognized as a public utility, the Robert Schuman Foundation works in favour of European construction. As a reference research center the Foundation develops studies on the European Union and its policies and promotes their content in France, Europe and abroad. The Foundation is an open and multinational network. Its main mission is to keep alive the spirit and inspiration of one of the "Fathers of Europe", Robert Schuman, and to promote European values and ideals both within and beyond the borders of the Union. The Foundation produces numerous studies on European policies which constitute a valuable source of information for all those who want to understand European issues and challenges. Its independence allows it to deal with all current issues in an in-depth and objective manner. Its studies and analyses provide European decision-makers with information,



arguments and food for thought that are appreciated for their usefulness and scientific quality. It multiplies initiatives in the field to advance the European democratic model.

## 2. Ecology-related organizations

### **Greenpeace**

Since its creation some 50 years ago, Greenpeace has acted on land and sea according to the principles of non-violence to protect the environment and promote peace. Today, we remain faithful to this mission, as well as to our total financial and ideological independence. Climate change, growing inequality, social injustice, migration and armed conflict... All the major challenges of our time, to which we must urgently respond, are intimately linked - as are the power structures that create them and the mentalities that accommodate them. This is why it is necessary to transform them together. Greenpeace is present in 55 countries, on all continents and oceans, through its 28 national and regional offices and its three boats. It has more than three million members and over 36,000 volunteers worldwide. We place citizen power at the heart of our campaigns by giving resonance to the work of all those who share our vision, our hopes and our conviction that we need profound transformations in our societies.

### **World Wild Fund for Nature**

WWF is one of the world's leading independent environmental organizations. WWF works to stop the degradation of the planet's natural environment and to build a future in which humans live in harmony with nature, conserving the world's biological diversity, ensuring the sustainable use of renewable natural resources, and promoting the reduction of pollution and waste. Since 1973, WWF France has been carrying out concrete actions to safeguard natural environments and their species, promote sustainable lifestyles, train decision-makers, support companies in reducing their ecological footprint, and educate young people. WWF France, a public utility foundation, works for a living planet from Paris, Marseille, the Alps, Guyana and New Caledonia. WWF is committed to action based on dialogue and respect for others, and adopts a global approach that takes into account the interdependence between the state of the planet and human development.

### **Groupe d'experts non-gouvernemental sur l'évolution du climat (NIPCC)**

The Non-Governmental Panel on Climate Change (NIPCC) is an international group of non-governmental scientists and academics who have come together to present a comprehensive, reliable and realistic assessment of the science and economics of global warming. Because it is not a government agency, and because its members are not predisposed to believe that climate change is caused by human greenhouse gas emissions, the NIPCC is able to offer an independent "second opinion" to the evidence reviewed - or not reviewed - by the Intergovernmental Panel on Climate Change (IPCC) on the issue of global warming. Since its founding in 2008, the NIPCC has been producing publications and reports for public policy. These reports aim, for example, to show that the impact of human-induced global warming is benign and could be beneficial to humanity and the natural world; that the evidence for rising sea levels is unreliable; or that there is no scientific consensus in the climate change debate.

### **Association des climato-réalistes**

Appeared in France in 2015, climatorealism sees the climate as an object of science and not ideology. Climate change is multiple and poorly understood, so there is no evidence that our way of life would cause "climate disruption." To say so is not selfishness, denial or anti-environmentalism, but realism. We need to think about how best to use our resources, and put our efforts where they really matter. The purpose of the association of climato-realists is to promote an open and free debate on the evolution of the climate and the societal and environmental issues related to it, by encouraging the expression of rigorous and well-founded opinions in all its forms. The association aims to make citizens aware of the stakes of climate and energy policies conducted in the name of the fight against global warming. The association is apolitical and totally free in the expression of its ideas. It strives to disseminate reliable information gathered from serious sources.

### 3. Social justice – related organizations

#### **SOS Méditerranée**

SOS Méditerranée is a European civil sea rescue association, independent of any political party and any religion, created in 2015 and made up of citizens mobilized to face the humanitarian emergency in the Mediterranean. SOS Méditerranée is based on the respect of man and his dignity, whatever his nationality, origin, social, religious, political or ethnic affiliation. The association's vocation is to provide assistance to any person in distress at sea who is within the scope of its action, without any discrimination. The persons concerned are men, women or children, migrants or refugees, who are in danger of death when crossing the Mediterranean Sea. The association also aims to ensure the protection of the survivors until their arrival in a safe port and to bear witness to the situation in the central Mediterranean. The association is financed by private donations and public grants. The funds collected are allocated to the rental of the boat, daily maintenance and rescue costs.

#### **FEMEN**

FEMEN is an international movement of feminist political activists with bare torsos, painted with slogans, and heads crowned with flowers. Our slogans are short and punchy; our chests are our banners. From the militant necessity is born the accomplishment of powerful and provocative but always non-violent actions. The movement was born in 2008 in Kiev, Ukraine. Since 2010, the activists are politicizing and using their breasts as a support for their demands. With our provocative and resounding actions, we target the multiple manifestations of the patriarchal order: dictatorships, sex industry and religions. We are a female revenge against the sclerotic patriarchal culture. We are an expression of freedom and pluralism. Our ambition is to change mentalities and the public image of women by exposing our strength, our courage and our convictions. It is by developing our political action that we will succeed in changing even our most intimate reality.

#### **Génération Identitaire**

Génération Identitaire is a political youth movement that brings together boys and girls across Europe. It was founded in September 2012. We call on young people to raise their heads: in the face of scum, in the face of those who want to control our lives and our thoughts, in the face of the standardization of peoples and cultures, in the face of the tidal wave of massive immigration, in the face of a school that hides the history of our people from us to prevent us from loving it, in the face of a so-called living together that turns into a nightmare... Génération Identitaire is the front line of resistance. Aware of the challenges we face, we do not refuse any battle. Proud of our heritage and confident in our destiny, we have only one watchword: we will not back down! We are the sacrificed generation, but not the lost generation. For we are going to war against all those who want to tear us away from our roots and make us forget who we are. Our ideal is reconquest, and we will carry it out to the end. Génération Identitaire is the barricade on which the youth in struggle for its identity stands.

#### **La Manif Pour Tous**

La Manif Pour Tous defends marriage and filiation in coherence with the sexual reality of humanity, whose consequence is both the difference and the complementarity of the sexes, which is essential to conceive a child and to assume the difference between father and mother, fatherhood and motherhood. Our goal is the respect of the superior interest and the elementary needs of the child, today threatened by the societal reforms inspired by the gender ideology. La Manif Pour Tous takes a pragmatic approach whose objective is to promote the well-being and the future of the child, the adult and society as a whole, what is commonly called the general interest. It acts for present and future generations. Finally, La Manif Pour Tous reminds us that the family is the crossroads of the difference between the sexes and the difference between generations. Only this context meets the essential needs of the child to come into the world, to know its personal identity, to enter little by little in relations with the others, to insert itself in the group then in the society, to contribute to the social peace.

## IV. Post-task questionnaires

### 1. Epistemic Curiosity

Below are several statements that people use to describe themselves. Please read each statement and then select the appropriate response, using the scale below to indicate how you feel about these statements. There are no right or wrong answers. Please do not spend too much time on each statement but give the answer that describes how you generally feel.

**1 = Almost Never 2 = Sometimes 3 = Often 4 = Almost Always**

1. I enjoy exploring new ideas.
2. Difficult conceptual problems can keep me awake all night thinking about solutions.
3. I enjoy learning about subjects that are unfamiliar to me.
4. I can spend hours on a single problem because I just can't rest without knowing the answer.
5. I find it fascinating to learn new information.
6. I feel frustrated if I can't figure out the solution to a problem, so I work even harder to solve it.
7. When I learn something new, I would like to find out more about it.
8. I brood for a long time in an attempt to solve some fundamental problem.
9. I enjoy discussing abstract concepts.
10. I work like a fiend at problems that I feel must be solved.

### 2. Exposition to information

Now please read each statement below and select the appropriate answer(s). There are no right or wrong answers.

What information platforms do you use? Please check one box or more.

- Television
- Radio
- Newspapers and magazines
- Internet
- Social networks
- Your friends

How often do you consult information sources?

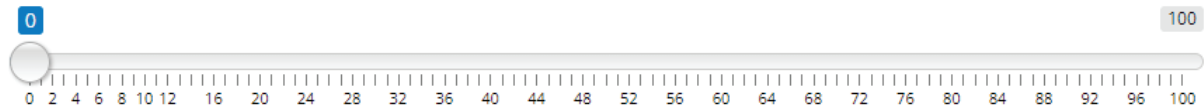
- Less than 1 time per week
- 1 time per week
- 2-3 times per week
- 4-5 times per week
- 1 time per day
- 2-3 times per day
- 4-5 times per day

- More than 4-5 times per day

How many different news sources do you consult regularly?

### 3. Perceived percentage of fake news

Please answer each question below by dragging the slider (in the case of a default answer, please move the slider to activate it):



On social networks, in your opinion, what is the percentage of fake news when the information comes from a journalist?

On social networks, in your opinion, what is the percentage of fake news when the information comes from a politician?

On social networks, in your opinion, what is the percentage of fake news when the information comes from a doctor?

On social networks, in your opinion, what is the percentage of fake news when the information comes from a researcher?

On social networks, in your opinion, what is the percentage of fake news when the information comes from a social justice actor?

On social networks, in your opinion, what is the percentage of fake news when the information comes from an ecological actor?

On the internet, in general, in your opinion, what is the percentage of fake news?

### 4. Manip check

Regarding the experiment:

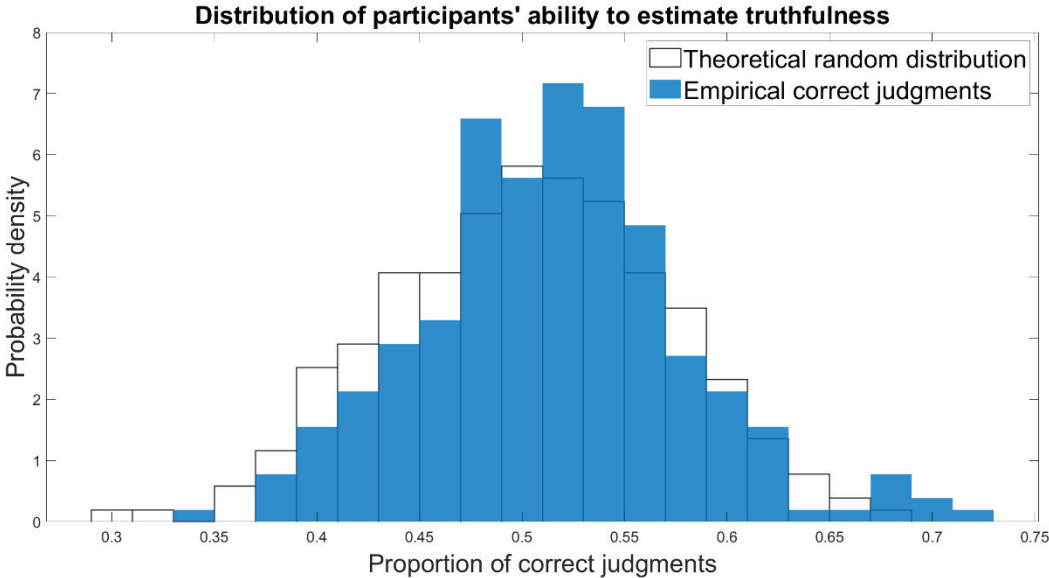
What was your strategy for choosing the number of chances out of 100 that the news was true or false?

What was your strategy for choosing what additional information to receive?

Did you think you would actually receive the additional information at the end of the experiment?

V. Behavioral analysis

1. Truthfulness Estimation



**Figure S1:** Distribution of the proportion of participants’ correct judgments against a theoretical random distribution (n=258, p=.05)

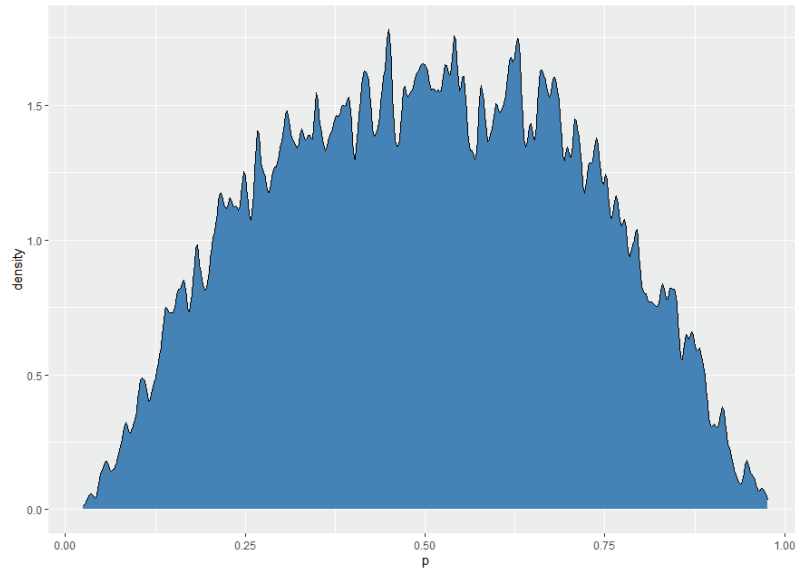
**Table S1:** Summary of success and judgment MLMs.

<i>Predictors</i>	success			success			judgment		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.98	0.85 – 1.11	0.717	0.47	0.40 – 0.55	< <b>0.001</b>	1.49	1.27 – 1.75	< <b>0.001</b>
judgment	0.97	0.83 – 1.14	0.725						
confidence	1.00	1.00 – 1.00	<b>0.041</b>	1.00	1.00 – 1.00	<b>0.004</b>	1.00	1.00 – 1.00	<b>0.018</b>
judgment * confidence	1.00	1.00 – 1.00	0.877						
truthfulness				3.70	3.05 – 4.49	< <b>0.001</b>	1.12	0.92 – 1.36	0.243
theme				1.11	1.04 – 1.18	<b>0.001</b>	0.93	0.87 – 0.99	<b>0.016</b>
truthfulness * theme				0.87	0.79 – 0.95	<b>0.002</b>	1.00	0.91 – 1.09	0.994
<b>Random Effects</b>									
$\sigma^2$	3.29			3.29			3.29		
$\tau_{00}$	0.00	subject		0.00	subject		0.11	subject	
	0.00	order		0.00	order		0.00	order	
	0.00	year		0.00	year		0.00	year	
	0.00	group		0.00	group		0.00	group	
ICC							0.03		
N	258	subject		258	subject		258	subject	
	2	group		2	group		2	group	
	48	order		48	order		48	order	
	2	year		2	year		2	year	

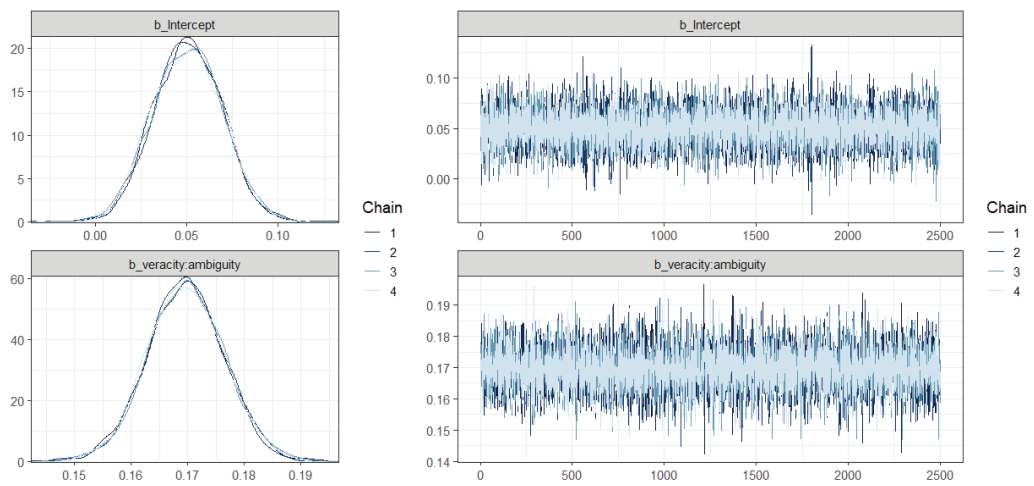
**Table S2:** Summary of the content-precision model, with success and judgment as response.

<i>Predictors</i>	<b>success</b>			<b>confidence</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.25	0.19 – 0.33	<0.001	51.86	48.18 – 55.53	<0.001
confidence	1.00	1.00 – 1.00	<b>0.001</b>			
truthfulness	13.02	9.00 – 18.84	<0.001	3.27	1.12 – 5.42	<b>0.003</b>
theme	1.09	1.03 – 1.16	<b>0.006</b>	1.60	0.90 – 2.31	<0.001
imprecision	1.13	1.08 – 1.18	<0.001	-0.21	-0.72 – -0.29	0.404
truthfulness * theme	0.89	0.81 – 0.97	<b>0.008</b>	-1.43	-2.42 – -0.44	<b>0.005</b>
truthfulness * imprecision	0.78	0.74 – 0.83	<0.001			
judgment				7.09	3.40 – 10.78	<0.001
judgment * imprecision				-1.22	-1.88 – -0.55	<0.001
<b>Random Effects</b>						
$\sigma^2$	3.29			527.30		
$\tau_{00}$	0.00	subject		254.47	subject	
	0.00	order		0.04	order	
	0.00	year		0.00	year	
	0.00	group		0.00	group	
ICC				0.33		
N	258	subject		258	subject	
	2	group		2	group	
	48	order		48	order	
	2	year		2	year	

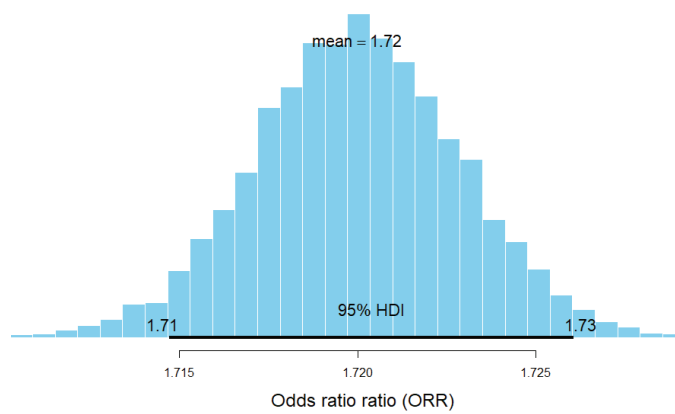




**Figure S2:** Weakly-informative prior distribution for Bayesian MLM of content precision.



**Figure S3:** Convergence plot for Bayesian MLM of content precision.



**Figure 4:** Plot of interaction effect of content precision model for success as response.  
For a unit difference in content precision, the ratio of the odds ratios is 1.72

## 2. Information-seeking

**Table S2:** Models of the decision to demand or avoid additional information and WTP.

<i>Predictors</i>	<b>reception</b>			<b>WTP</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.27	0.36 – 4.55	0.711	7.26	1.31 – 13.21	<b>0.017</b>
judgment	0.87	0.70 – 1.07	0.175			
confidence	0.98	0.98 – 0.99	<b>&lt;0.001</b>	0.02	0.02 – 0.02	<b>&lt;0.001</b>
year	0.87	0.42 – 1.79	0.709	-1.61	-5.30 – 2.09	0.395
judgment * confidence	1.01	1.00 – 1.01	<b>&lt;0.001</b>			
reception				1.86	1.55 – 2.17	<b>&lt;0.001</b>
reception * confidence				-0.02	-0.02 – -0.01	<b>&lt;0.001</b>
<b>Random Effects</b>						
$\sigma^2$	3.29			12.79		
$\tau_{00}$	7.04	subject		27.84	subject	
	0.00	order		0.00	order	
	0.00	year		1.52	year	
	0.00	group		0.00	group	
N	258	subject		258	subject	
	2	group		2	group	
	48	order		48	order	
	2	year		2	year	

**Table S3:** Mediation model predicting the demand for additional information

Effect		Estimated Individual performance	CI [2.5%, 97,5%]	SE
direct effect				
$\beta_a$	CP -> Reception	-.017	[.046, .012]	.014
$\beta_c$	<b>CP -&gt; Confidence</b>	<b>-.84 ***</b>	<b>[-1.23, -.444]</b>	<b>.202</b>
$\beta_b$	<b>Confidence -&gt; Reception</b>	<b>-.009 *</b>	<b>[-.01, -.008]</b>	<b>.001</b>
Indirect effect				
	<b>CP -&gt; Confidence -&gt; Reception</b>	<b>-.0018 ***</b>	<b>[-.0009, .00]</b>	
Total effect				
$\beta_c + \beta_{ab}$		-0.0041	[-0.011, .00]	

Note. CI= confidence interval. CP = Content precision. \* $p < 0.05$ . \*\* $p < 0.01$ . \*\*\* $p < 0.001$

# Chapter II. Supplementary Materials

## I. Stimuli rating instructions

Welcome to this session.

We ask that you turn off your phone.

You will be participating in a session consisting of 42 periods.

This session will last approximately 40 minutes.

In each period, you will see a short piece of information. For each brief you will be asked to answer 5 questions. The first 4 questions will be, for each brief, mandatory.

The brief information will be different for each period.

Some of the news briefs are from the French written press for the period 2017-2020. Others have been fabricated.

Unless specified, their content is recent, topical and concerns the French territory.

Please read each brief and each question carefully, then take the time you need to answer.

## II. Participants instructions

### **Instructions of the experiment**

Welcome to this experiment.

First, we ask you to turn off your phone.

Please read these instructions carefully. They give you all the information you need to participate in this experiment. If you don't understand something, don't hesitate to raise your hand. We will come and answer you.

**At the end of this experiment, depending on your decisions, you can earn Experimental Currency Units (ECU).** These ECUs will be converted into euros at a ratio of 100 ECUs = 1€.

**At the end of the experiment, you will receive the monetary amount equivalent to the conversion of the ECUs you have earned plus the 60 euros initially provided for your participation.**

This session is composed of **2 parts**.

You will first receive instructions for the first part, then instructions for the second part. At the end of the task, we will ask you to fill in several questionnaires.

**Your answers throughout the experiment will be and remain anonymous.**

### **II.1. Part 1 of the experiment**

The first part of this experiment consists of evaluating 12 different organizations based on 6 criteria. You will first see the organization to be evaluated and a summary of its purpose from its website, as in the example below:

## LA PÉTANQUE CARYENNE

Depuis plus de soixante ans, la Pétanque Carryenne existe sur la commune de Carry-le-Rouet. Autrefois, les concours se faisaient sur la place du marché et au Family (aujourd'hui disparu). Pas moins de 10 présidents se sont transmis le flambeau afin que perdure la pratique des jeux de boules ; tant pétanque que jeu provençal.

Le terrain actuel de la Pétanque Carryenne a été construit dans les années 70 et le nombre de licenciés n'a cessé d'augmenter. Ce club a eu l'honneur d'accueillir en son sein par trois fois des équipes championnes de France à pétanque en catégorie vétérans et des vainqueurs du prestigieux concours Mondial La Marseillaise à Pétanque.

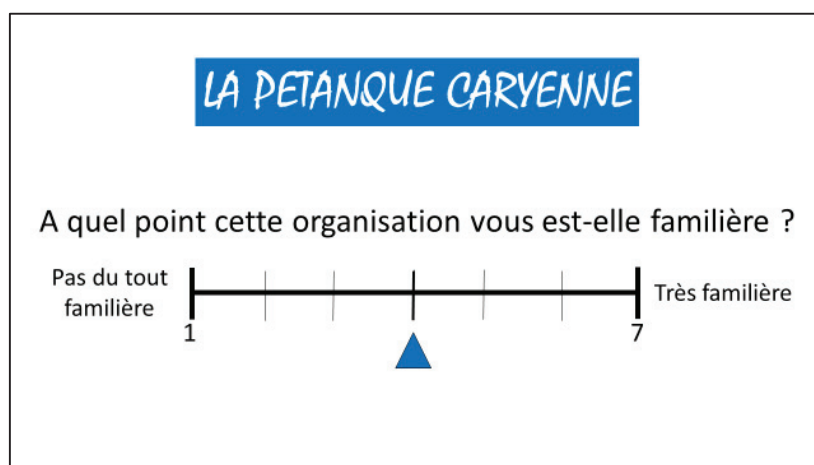
Actuellement, le club compte près de trois cents membres et depuis trois ans, une école de pétanque permettant de transmettre aux nouvelles générations la passion des sports de boules.

Et comme le dit la devise du club : « Ici, la pétanque est plus qu'un jeu, c'est une religion ! » C'est dans un cadre exceptionnel, au cœur de la ville et à deux pas de la grande bleue que les Carryens vous accueilleront chaque jour pour des parties passionnantes et conviviales.

Then you will rate it on the following 6 criteria using a 7-point scale:

- How familiar is this organization to you?
- How close do you feel to this organization?
- How much do you like this organization?
- How familiar is this organization to those closest to you?
- How close do or would your family feel to this organization?
- How much do or would your relatives like this organization?

For example, below, 1 means that the organization is 'Not at all familiar' to you and 7 means that the organization is 'Very familiar' to you:



These evaluations will be conducted in private. They will not be visible to any other participant or experimenter. They will not be associated with any element that could identify you.

## II.2. Part 2 of the experiment

There are two types of participants in the second part of the experiment: Participant A and Participant B. **You are a Participant A.** You will keep this role throughout this part. Participant B has completed their task in a **previous experiment**.

You will play with answers from these B participants.

**There are 20 B participants. Participants B are all-comers.** There is no way for you to recognize these B participants.

The answers you will play with have been drawn from all their answers. There is no way for you to know which participant B the answers you will play with come from.

There are 96 periods. **Each period, you will be matched with a different B Participant.** You are likely to be paired with the same person several times. However, it is very unlikely that you will be associated with the same person several periods in a row. You will not be able to identify which participants you are associated with.

### A. Evaluation of the brief news:

At the beginning of each period, you will see a short piece of information.

Each brief will be different and will come from different media. **For each news brief, there is a supplement.** The supplementary information consists of an investigative file. Their content is related to the content of the brief and can be related to other information around the content of the brief.

Warning: news items may be true or false. Some of the stories you will be exposed to and manipulate will be false information.

**False information has been fabricated.**

**The non-fabricated news items are taken from the French written press over the period 2017 - 1st quarter 2020. The non-fabricated news items are all true.**

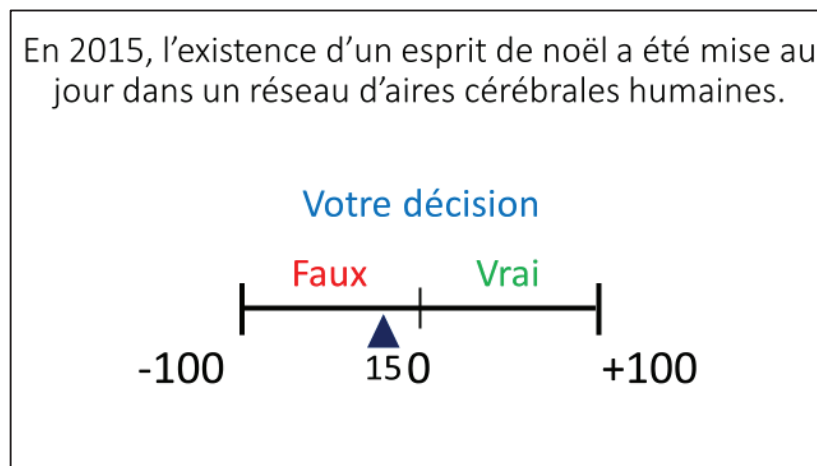
**Unless specified, their content is therefore recent, current and concerns the French territory at the time of the 1st quarter 2020.**



At the end of the experiment, you will be debriefed on the information you have been exposed to.

After the brief information has been displayed, you will be asked to evaluate whether its content is true or false by specifying, “in your opinion, what is the number of chances out of 100 that this brief is true or false? “

To answer that a news item is true or false, you will drag the slider below the question, as in the following example:



- To answer that the content of the brief is false, you will drag the cursor between the **0 and -100 bounds: these bounds delimit the 'False' answer.**

- To answer that the content of the brief is true, you will drag the cursor between the **0 and +100 bounds: these bounds delimit the answer 'True'.**

To help you answer this question, "robots" will be available during this step: There are 100 different robots. **Each robot has a different accuracy level** than the other robots. This accuracy level is the **number of chances out of 100 that the robot evaluates the short information correctly.**

This number is an integer Y between 1 and 100.

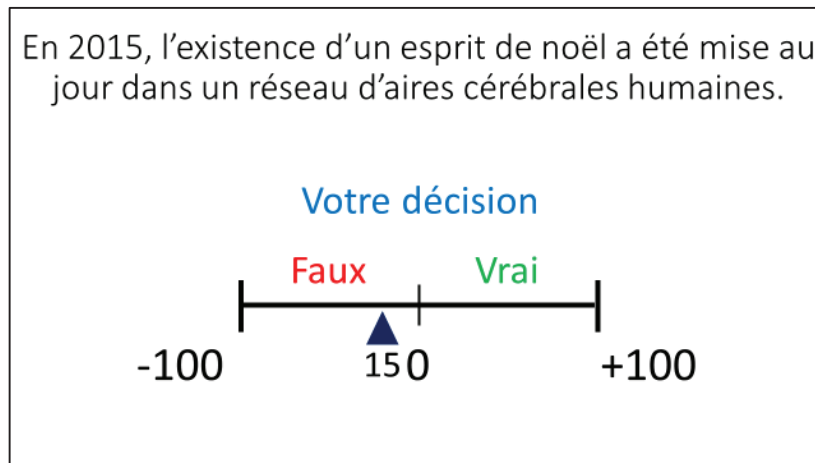
That is, there is one robot that has a 1 in 100 chance of answering correctly, there is one robot that has a 2 in 100 chance of answering correctly, and so on until the 100th robot that has a 100 in 100 chance of answering correctly. A robot that has 75 chances out of 100 to answer correctly will give a correct answer 75% of the time and will give a wrong answer 25% of the

time.

At each period, the computer will draw one robot from the 100 robots. All robots have the same chance of being drawn. You will not know which one was drawn and the robot will change randomly each period.

You will have to choose which robots you would let answer the truthfulness question for you. To do this, you will decide how confident you are in your answer by choosing a threshold (a number  $X$  between 1 and 100) using the slider.

If the robot's  $Y$  number is **higher** than your  $X$  threshold, the robot will answer for you. If the robot's  $Y$  number is **lower than or equal** to your  $X$  threshold, **your answer will be submitted**.



Here, the participant answers that the content of the short information is wrong.

If the drawn robot has more than a 15% chance of answering correctly, the participant would prefer the submitted answer to be the robot's answer rather than their own.

The confidence threshold will be such that, for any robot with an accuracy level **higher than your threshold**, you would prefer the program to consider the **robot's answer** rather than yours.

For any robot with an accuracy level **lower than or equal to your threshold**, you would prefer the program to take **your answer** into account.

Thus, for example:

- If you choose 75 as your confidence threshold and the randomly selected robot has an accuracy level of 90, the program will consider that robot's answer. **The robot will have a 90 out of 100 chance of giving the right answer.**

- If you choose 75 as your confidence level and the randomly selected robot has an accuracy level of 20, the program will consider your answer.

**You are therefore encouraged to say how correct you really think your answer is.** The less sure you are of your answer, the more you are likely to choose a low threshold (i.e., if you think there is a 25 in 100 chance that your answer is correct, it is better to give a threshold of 25).

The program is indifferent to the sign of the confidence level. Choosing the numbers '75' or '- 75' will mean to the program that you think there is a 75 out of 100 chance that your assessment of the brief is correct. The direction of your response (positive or negative) is used to indicate your assessment of the brief ('False' or 'True').


**Sixteen** of your ratings out of 96 will be **randomly** selected at the end of the task. They will reward your performance. Your reward will be determined only on the randomly selected ratings. Each correct randomly selected assessment will earn you **50 ECUs**.

**Participant B will have evaluated the same brief as you did, at the same time of the task as you did, and in the same way as you did.**

**B. Sending Information:**

Each period, once you have evaluated the brief, you will be asked to **choose** between **sending** Participant B **additional information** about the content of the brief and **not sending anything**, as in the example below:

En 2015, l'existence d'un esprit de Noël a été mise au jour dans un réseau d'aires cérébrales humaines.

 • Envoyer plus d'informations

• Ne pas envoyer plus d'informations

B Participants will have chosen between **receiving additional information** about the content of the brief and **not receiving anything**.

**Sixteen** of your sending choices out of 96 will be **randomly** selected at the end of the task. They will reward your performance in the following way:

For each of these sending choices,

- If your choice **matches** Participant B's choice, this period will **reward** you with 50 Experimental Currency Units (ECU).

- If your choice **does not match** the Participant B choice, this period will **not reward you** with ECUs.

B-Participants were asked to declare their desirability for each piece of information in the following manner:

B-Participants were asked in **each period** to declare

**1) their choice of reception** (to receive or not to receive more information)

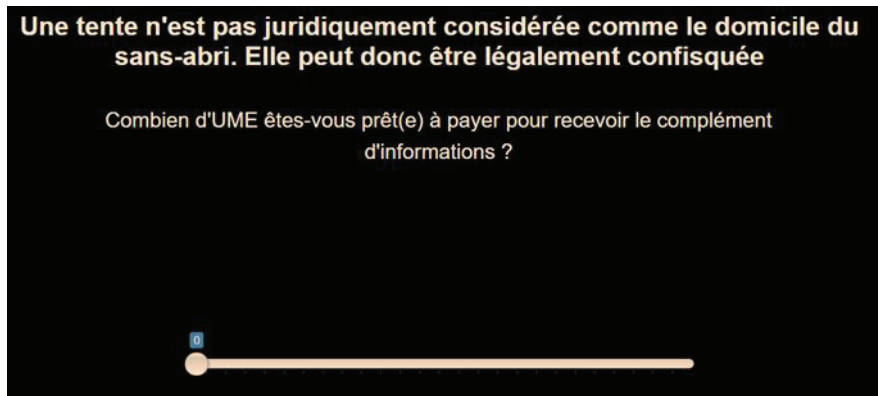
**2) their willingness to see their choice fulfilled** by indicating what fraction of their monetary allocation they were willing to spend to see their choice fulfilled.

B Participants thus had **monetary incentives** to reveal their **true preference**.

For example:



- 1) B Participants had to declare at each period their choice of reception (to receive or not to receive more information)



- 2) B Participants were then asked to declare their willingness to have their choice fulfilled by indicating what fraction of their monetary allowance they were willing to spend.

At the end of their task, a limited number of periods were randomly selected to make their choice. Their willingness to have their choice made then determined the chances that their choice would **actually happen**.

Their task was thus designed to lead them to choose, for each brief, whether they actually wanted to receive additional information.

**C. End of Period:**

You will know at the end of the period if your choice matches Participant B's choice:

Votre choix ne correspond pas au choix du Participant B

#### D. Information on Participant B :

There are 2 types of periods:

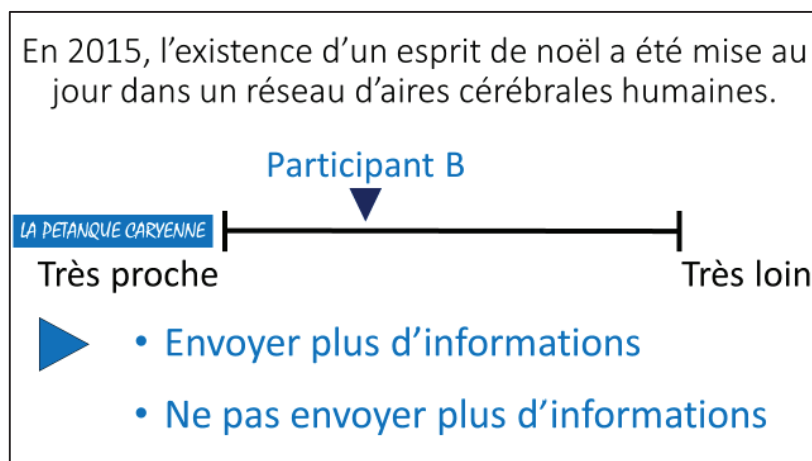
- The first half of the periods (48 periods) is characterized by the absence of information on Participant B.
- The second half of the periods (48 periods) is characterized by the display, in each period and at the time of the sending choice, of information about Participant B.

The information about Participant B will be displayed **during** your choice to send more information or not. It will take the form of the **distance** Participant B rated between him/her and a randomly selected organization.

This organization will be **one of the ones** that you yourself have evaluated in the first part of this experiment. It will be related to the brief you have to choose to send or not to send.

The distance assessed corresponds to a projection on a scale of 0 to 100 of Participant B's answers to the **6 organization assessment questions**.


For example, here Participant B tends to feel close to the organization 'La Pétanque Carryenne':



Periods without information on Participant B will indicate this lack of information. They will take the following form:

En 2015, l'existence d'un esprit de Noël a été mise au jour dans un réseau d'aires cérébrales humaines.

Pas d'indice



▶ • Envoyer plus d'informations  
• Ne pas envoyer plus d'informations

**E. End of Part 2:**

Once this Part 2 is complete, we will ask you to complete several questionnaires. One of these will earn you several ECUs. You will be given instructions on these quizzes in due course.

Finally, you will be debriefed on the information and receive the verified version of the information you were exposed to.

**3. Calculation of Final Compensation**

**A. Information Evaluation:**

At the end of the experience, sixteen periods will be drawn from the 96 periods in Part 2. They will reward your judgment of truthfulness.

You will be rewarded, for each period drawn, as follows:

- If your answer is selected for review,  
If your answer is correct, that period will be rewarded.  
In case your answer is incorrect, that period will not be rewarded.
  
- If the drawn robot has a probability of answering correctly that is higher than your confidence level, the **probability** that this period will **reward** you will be the **probability Y** that the robot **answers correctly**.



A correct answer, either yours or the robot's, will get you 50 ECUs. An incorrect answer will get you 0 ECUs.

B. Sending information:

At the end of the experiment, sixteen more periods will be drawn from the 96 periods in Part 2. These will reward your performance in making a choice that matches Participant B's choice.

A choice that matches Participant B's choice will earn you 50 ECUs.

A choice that does not match Participant B's choice will earn you 0 ECUs.

C. Questionnaires:

At the end of the experiment, one of your decisions on one of the questionnaires will be drawn.

You will win the amount of ECUS that you have allocated for this decision.

D. Final Remuneration:

The ECUs will be converted into Euros at a ratio of 100 ECUs = 1€. The compensation for your performance and the compensation for answering the questionnaires will be added to your initial sum of 60 euros.

#### 4. Experimentation

You will start the experiment with a training phase. Please ask your questions at this point in the task.

The information, themes and clues you see during the training phase will not be present during the experimentation phase. The answers you play with during the practice phase **will not be taken from** the answers of the 20 B participants. They will all have been chosen **at random** to allow you to familiarize yourself with the task.

Attention: **The task is constrained by a limited time to answer.** This constraint exists during the training phase and during the experimentation phase. You will become familiar with this time during the training phase. Throughout the task, you will have less than 10 seconds to

respond to each decision. Each trial that you do not answer in time will come back at the end of the task.

### III. Organizations summaries

#### III.1. Democracy related organizations

##### **France FREXIT:**

France FREXIT is a private and independent initiative, created in March 2018 and aiming at informing, gathering, exchanging, proposing on the theme of Frexit. In all legality, without any violence, and in the respect of the Institutions. Taking advantage of the historical opportunity of the Frexit, FRANCE FREXIT proposes the complete reform of the Republic and the French State by a great FRENCH NATIONAL COORDINATION, based on a new formula of Power, including a truly democratic organization taking the best of the Republic and leaving the least good, and using modern means of communication: votes, electronic votes, draws, as well as some monarchical aspects, mainly in terms of spirituality and transmission of universal and traditional values of France. FRANCE FREXIT is politically opposed to the euro-extremist and euro-identitarian parties such as LaREM, LR, MODEM, EELV, PS, UDI, Parti Radical. FRANCE FREXIT does not support the euro-alternative parties and other decoy parties such as the RN, the LFI or DLF, the NPA, LO and some other euro-compatible parties.

##### **Parti Libertarien:**

Our observation is the same as the majority of French people: the weight of the state and its scope of action are constantly increasing, hindering our freedoms more and more, with the catastrophic results that everyone can see. We are also facing an extremely worrying legislative inflation, the accumulation of standards, regulations and laws make the system incomprehensible for the majority of French people and impracticable for entrepreneurs. Only large groups and multinationals benefit from this complexity and can expand without real competition. These regulations paralyze any personal initiative, block the social elevator and increase inequalities. Beyond the dramatic consequences of such an intrusion of the state in the life of individuals, the latter, by wanting to regulate every aspect of our daily life, goes beyond

its prerogatives and violates our fundamental right to manage our life as we see fit. We are libertarians and we consider that it is up to free individuals to write their own history.

**Le Mouvement Européen – France:**

The European Movement - France has been mobilizing since 1950, across all generations, to bring to life a pluralist public debate on Europe. It deploys its activities around pedagogy, the organization of debate between citizens and the formulation of proposals to build Europe. It gathers thousands of volunteers gathered in more than 50 local sections, about twenty member organizations as well as a college of qualified personalities. Heir to the spirit of the founding fathers of Europe, the Movement is the first actor of civil society in France on European issues. It also mobilizes through its youth branch, the Young Europeans - France, which has 26 local groups throughout the territory. At the European level, the European Movement is also a member of the European Movement - International, a network of 39 organizations that make our commitment resonate throughout the continent. The European Movement-France is recognized as an association of general interest since July 22, 2016 and is also approved as an "educational association complementary to public education".

**Fondation Robert Schuman:**

Created in 1991 and recognised as a public utility, the Robert Schuman Foundation works in favour of European construction. As a reference research center the Foundation develops studies on the European Union and its policies and promotes their content in France, Europe and abroad. The Foundation is an open and multinational network. Its main mission is to keep alive the spirit and inspiration of one of the "Fathers of Europe", Robert Schuman, and to promote European values and ideals both within and beyond the borders of the Union. The Foundation produces numerous studies on European policies which constitute a valuable source of information for all those who want to understand European issues and challenges. Its independence allows it to deal with all current issues in an in-depth and objective manner. Its studies and analyses provide European decision-makers with information, arguments and food for thought that are appreciated for their usefulness and scientific quality. It multiplies initiatives in the field to advance the European democratic model.

**III.2. Ecology-related organizations**

**Greenpeace:**

Since its creation some 50 years ago, Greenpeace has acted on land and sea according to the principles of non-violence to protect the environment and promote peace. Today, we remain faithful to this mission, as well as to our total financial and ideological independence. Climate change, growing inequality, social injustice, migration and armed conflict... All the major challenges of our time, to which we must urgently respond, are intimately linked - as are the power structures that create them and the mentalities that accommodate them. This is why it is necessary to transform them together. Greenpeace is present in 55 countries, on all continents and oceans, through its 28 national and regional offices and its three boats. It has more than three million members and over 36,000 volunteers worldwide. We place citizen power at the heart of our campaigns by giving resonance to the work of all those who share our vision, our hopes and our conviction that we need profound transformations in our societies.

**WWF:**

WWF is one of the world's leading independent environmental organizations. WWF works to stop the degradation of the planet's natural environment and to build a future in which humans live in harmony with nature, conserving the world's biological diversity, ensuring the sustainable use of renewable natural resources, and promoting the reduction of pollution and waste. Since 1973, WWF France has been carrying out concrete actions to safeguard natural environments and their species, promote sustainable lifestyles, train decision-makers, support companies in reducing their ecological footprint, and educate young people. WWF France, a public utility foundation, works for a living planet from Paris, Marseille, the Alps, Guyana and New Caledonia. WWF is committed to action based on dialogue and respect for others, and adopts a global approach that takes into account the interdependence between the state of the planet and human development.

**Le Groupe d'experts non-gouvernemental sur l'évolution du climat (NIPCC):**

The Non-Governmental Panel on Climate Change (NIPCC) is an international group of non-governmental scientists and academics who have come together to present a comprehensive, reliable and realistic assessment of the science and economics of global warming. Because it is not a government agency, and because its members are not predisposed to believe that climate change is caused by human greenhouse gas emissions, the NIPCC is able to offer an independent "second opinion" to the evidence reviewed - or not reviewed - by the Intergovernmental Panel on Climate Change (IPCC) on the issue of global warming. Since its founding in 2008, the NIPCC has been producing publications and reports for public policy.

These reports aim, for example, to show that the impact of human-induced global warming is benign and could be beneficial to humanity and the natural world; that the evidence for rising sea levels is unreliable; or that there is no scientific consensus in the climate change debate.

#### **L'association des climato-réalistes:**

Appeared in France in 2015, climatorealism sees the climate as an object of science and not ideology. Climate change is multiple and poorly understood, so there is no evidence that our way of life would cause "climate disruption." To say so is not selfishness, denial or anti-environmentalism, but realism. We need to think about how best to use our resources, and put our efforts where they really matter. The purpose of the association of climato-realists is to promote an open and free debate on the evolution of the climate and the societal and environmental issues related to it, by encouraging the expression of rigorous and well-founded opinions in all its forms. The association aims to make citizens aware of the stakes of climate and energy policies conducted in the name of the fight against global warming. The association is apolitical and totally free in the expression of its ideas. It strives to disseminate reliable information gathered from serious sources.

### **III.3. Social justice – related organizations**

#### **SOS Méditerranée:**

SOS Méditerranée is a European civil sea rescue association, independent of any political party and any religion, created in 2015 and made up of citizens mobilized to face the humanitarian emergency in the Mediterranean. SOS Méditerranée is based on the respect of man and his dignity, whatever his nationality, origin, social, religious, political or ethnic affiliation. The association's vocation is to provide assistance to any person in distress at sea who is within the scope of its action, without any discrimination. The persons concerned are men, women or children, migrants or refugees, who are in danger of death when crossing the Mediterranean Sea. The association also aims to ensure the protection of the survivors until their arrival in a safe port and to bear witness to the situation in the central Mediterranean. The association is financed by private donations and public grants. The funds collected are allocated to the rental of the boat, daily maintenance and rescue costs.

#### **FEMEN:**

FEMEN is an international movement of feminist political activists with bare torsos, painted with slogans, and heads crowned with flowers. Our slogans are short and punchy; our chests are our banners. From the militant necessity is born the accomplishment of powerful and provocative but always non-violent actions. The movement was born in 2008 in Kiev, Ukraine. Since 2010, the activists are politicizing and using their breasts as a support for their demands. With our provocative and resounding actions, we target the multiple manifestations of the patriarchal order: dictatorships, sex industry and religions. We are a female revenge against the sclerotic patriarchal culture. We are an expression of freedom and pluralism. Our ambition is to change mentalities and the public image of women by exposing our strength, our courage and our convictions. It is by developing our political action that we will succeed in changing even our most intimate reality.

### **Génération Identitaire:**

Génération Identitaire is a political youth movement that brings together boys and girls across Europe. It was founded in September 2012. We call on young people to raise their heads: in the face of scum, in the face of those who want to control our lives and our thoughts, in the face of the standardization of peoples and cultures, in the face of the tidal wave of massive immigration, in the face of a school that hides the history of our people from us to prevent us from loving it, in the face of a so-called living together that turns into a nightmare... Génération Identitaire is the front line of resistance. Aware of the challenges we face, we do not refuse any battle. Proud of our heritage and confident in our destiny, we have only one watchword: we will not back down! We are the sacrificed generation, but not the lost generation. For we are going to war against all those who want to tear us away from our roots and make us forget who we are. Our ideal is reconquest, and we will carry it out to the end. Génération Identitaire is the barricade on which the youth in struggle for its identity stands.

### **La Manif Pour Tous:**

La Manif Pour Tous defends marriage and filiation in coherence with the sexual reality of humanity, whose consequence is both the difference and the complementarity of the sexes, which is essential to conceive a child and to assume the difference between father and mother, fatherhood and motherhood. Our goal is the respect of the superior interest and the elementary needs of the child, today threatened by the societal reforms inspired by the gender ideology. La Manif Pour Tous takes a pragmatic approach whose objective is to promote the well-being and the future of the child, the adult and society as a whole, what is commonly called the general

interest. It acts for present and future generations. Finally, La Manif Pour Tous reminds us that the family is the crossroads of the difference between the sexes and the difference between generations. Only this context meets the essential needs of the child to come into the world, to know its personal identity, to enter little by little in relations with the others, to insert itself in the group then in the society, to contribute to the social peace.

#### IV. Task comprehension questionnaire

*Please answer the following comprehension questions. Please call the experimenter if you have any questions.*

##### 1. General Questions

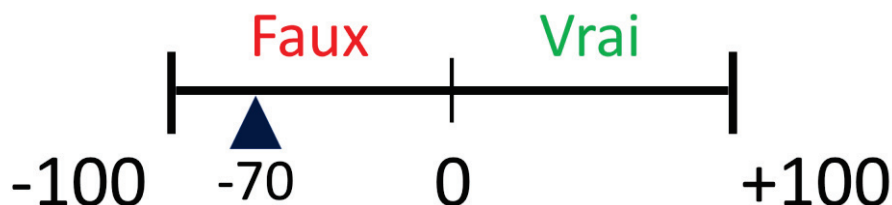
What is your goal during the information evaluation step?

What is your goal during the sending of additional information step?

How many B Participants are there? How were they chosen?

##### 2. Information evaluation case n°1

Consider the case where you answer that a piece of information is false and you drag the slider to the value -70. If the randomly drawn robot has a 50% chance of answering correctly:



What evaluation will be submitted?

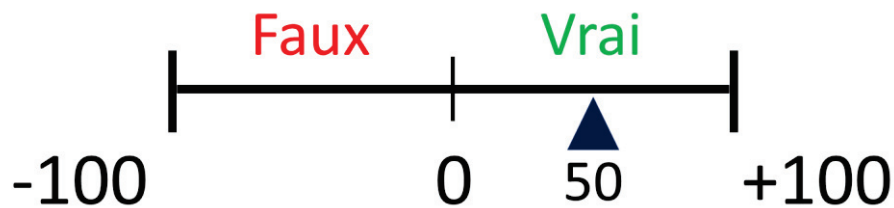
- A. The submitted evaluation will be yours
- B. The submitted evaluation will be the robot's



Which answer will be submitted between True and False?

3. Information evaluation case n°2

Consider the case where you answer that a piece of information is true and you drag the slider to the value 50. If the randomly drawn robot has a 75% chance of answering correctly:



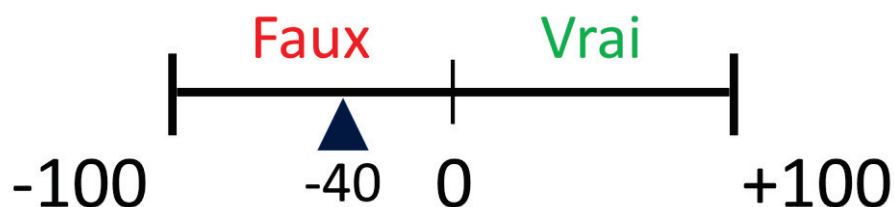
What evaluation will be submitted?

- A. The evaluation submitted will be yours
- B. The submitted evaluation will be the robot's

What will be the submitted answer if the submitted evaluation is the robot's?

4. Information evaluation case n°3

Consider the case where you answer that a piece of information is false and you drag the slider to the value -40. If the randomly drawn robot has a 60% chance of answering correctly:



What evaluation will be submitted?

- A. The evaluation submitted will be yours
- B. The submitted evaluation will be the robot's

What will be the submitted answer if the submitted evaluation is the robot's?

5. Information transmission case n°1

Consider the case where you choose not to send more information and Participant B has chosen not to receive more information.

Is the period a failure?

How many MEUs is this period worth?

6. Information transmission case n°2

Consider the case where you choose not to send more information and Participant B has chosen to receive more information.

Is the period a failure?

How many MEUs is this period worth?

7. Information transmission case n°3

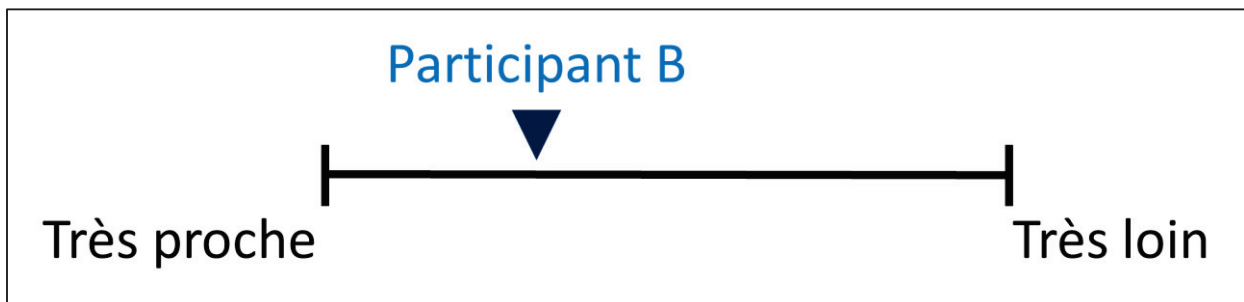
Consider the case where you choose to send more information and Participant B has chosen not to receive more information:

Is the period a failure?

How many MEUs is this period worth?

8. Information about Participant B

Let's consider the case where the following information about Participant B is provided to you before the choice to send information:



Participant B sees himself or herself as rather:

- A. Very close to the organization
- B. Moderately close to the organization
- C. Neither close nor far from the organization
- D. Somewhat distant from the organization
- E. Very far from the organization

## V. Post-task questionnaires

### V.1. 11-20 game

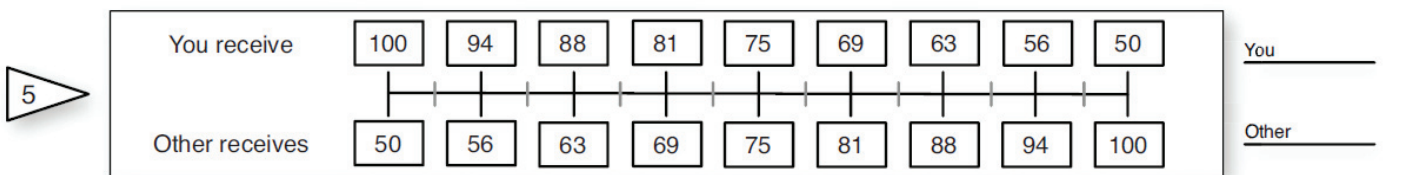
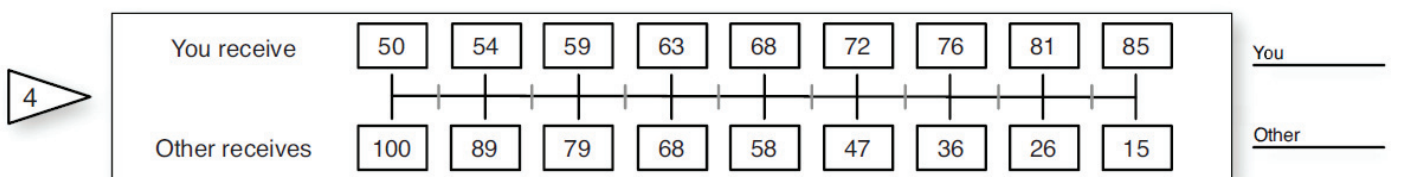
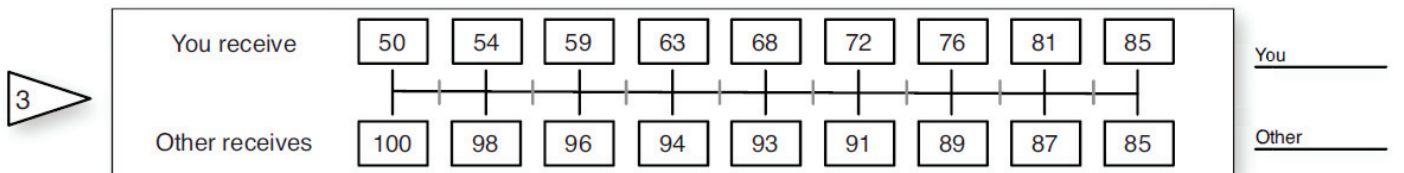
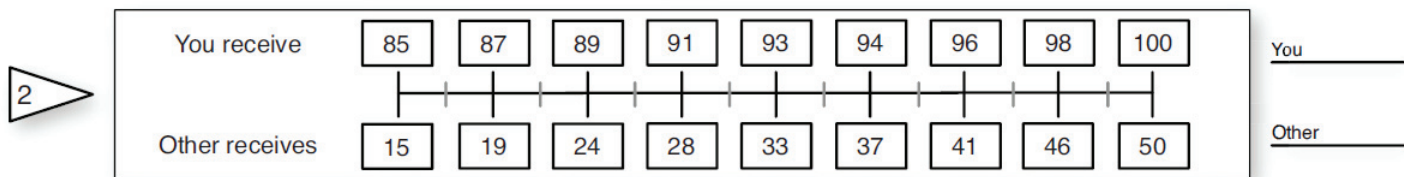
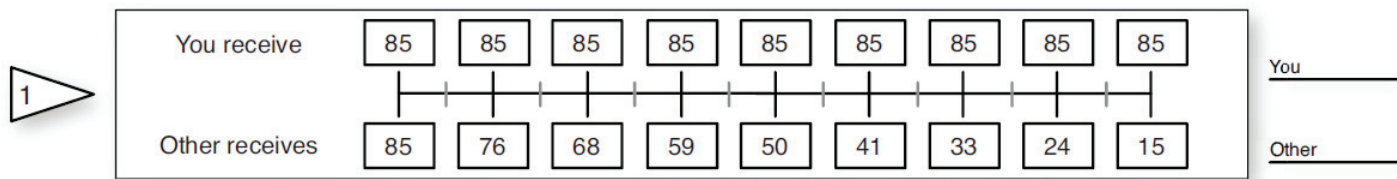
You and another person will participate in a game in which each player requests an amount of MEU. The amount must be a whole number between 11 and 20. The other person will be selected from the participants in this study. Each player will receive the amount he/she requests. A player will receive an additional 20 MEU if he/she requests exactly 1 MEU less than the other player. The total amount will be added to the amount of MEUs earned so far. You will only participate in this game once and with only one other person.

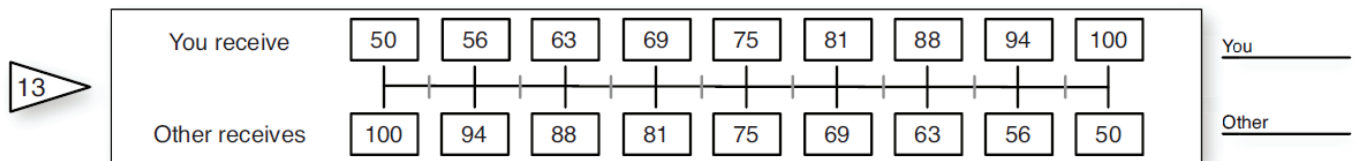
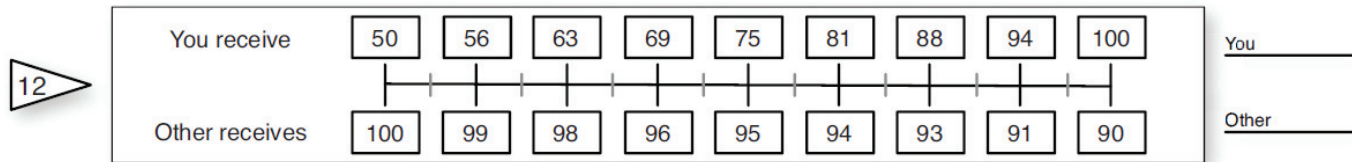
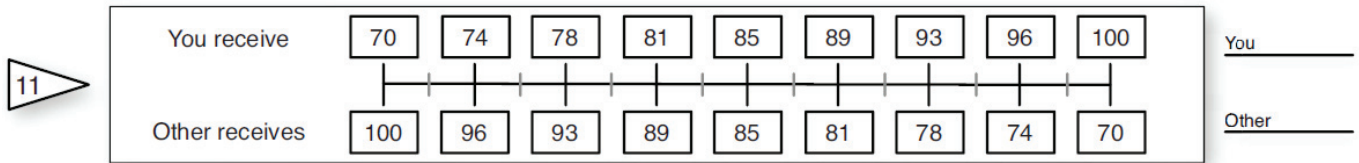
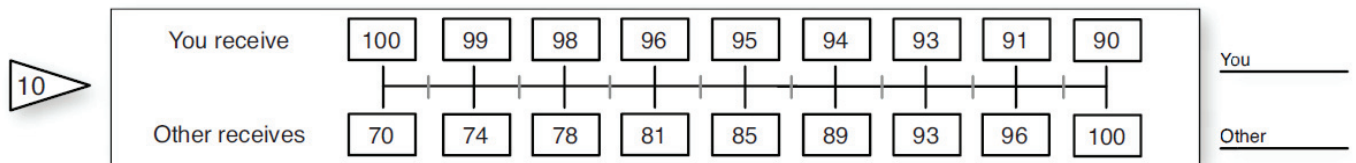
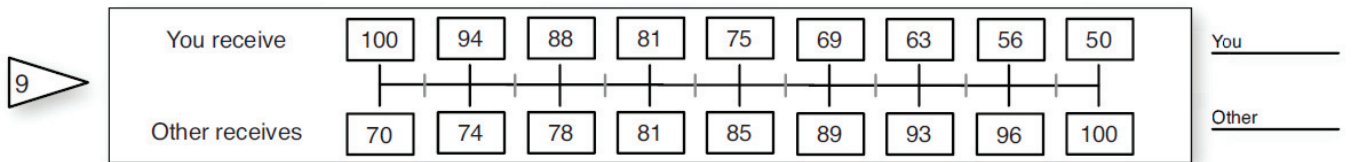
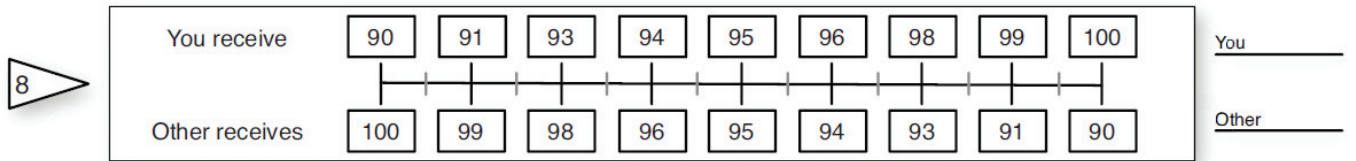
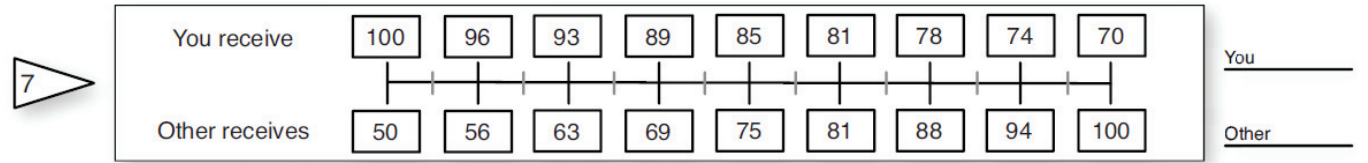
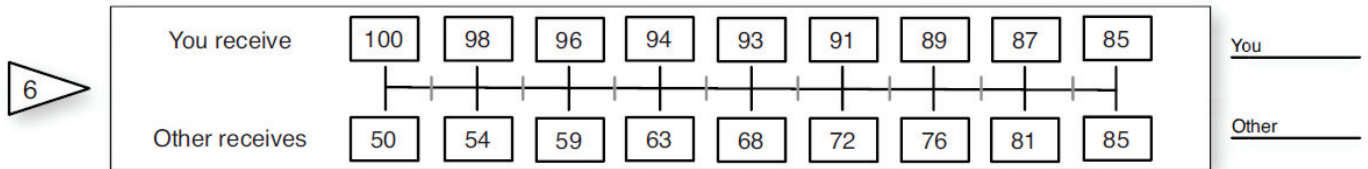
What amount are you requesting? Please answer as soon as possible.

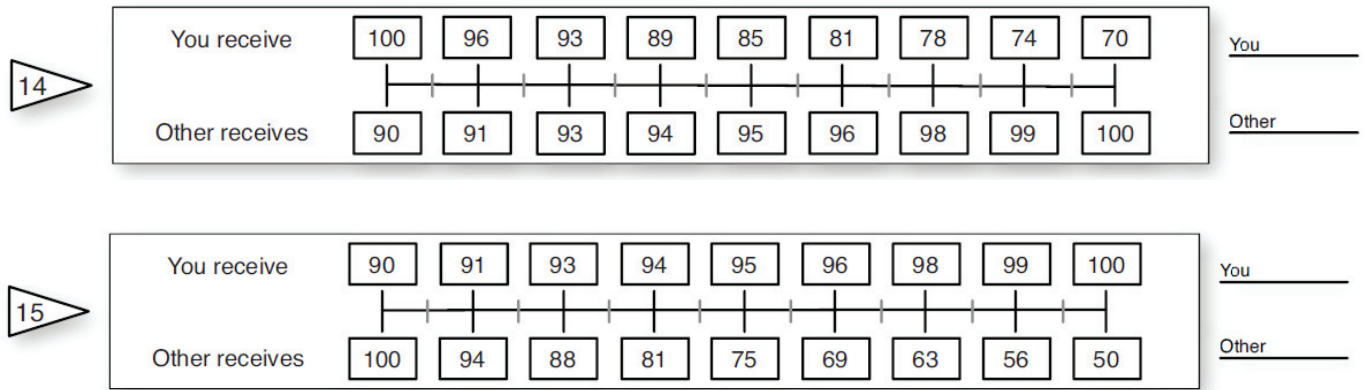
### V.2. Murphy's slider (SVO)

In this task you have been randomly paired with another person, whom we will refer to as the other. This other person is someone you do not know and will remain mutually anonymous. All of your choices are completely confidential. You will be making a series of decisions about allocating resources between you and this other person. For each of the following questions, please indicate the distribution you prefer most by marking the

respective position along the midline. You can only make one mark for each question. Your decisions will yield money for both yourself and the other person. In the example below, a person has chosen to distribute money so that he/she receives 50 dollars, while the anonymous other person receives 40 dollars. There are no right or wrong answers, this is all about personal preferences. After you have made your decision, write the resulting distribution of money on the spaces on the right. As you can see, your choices will influence both the amount of money you receive as well as the amount of money the other receives.







### V.3. Manip check questionnaire

1. What was your strategy for choosing the number of chances out of 100 that the information was true or false?
2. What was your strategy for choosing what additional information to send?
3. How did you guess whether the receiver wanted or did not want to receive more information?
4. Did you understand that you were playing each round with a different receiver's response, taken from all 20 receivers?

## VI. Behavioral analyses

### VI.1. Truthfulness estimation

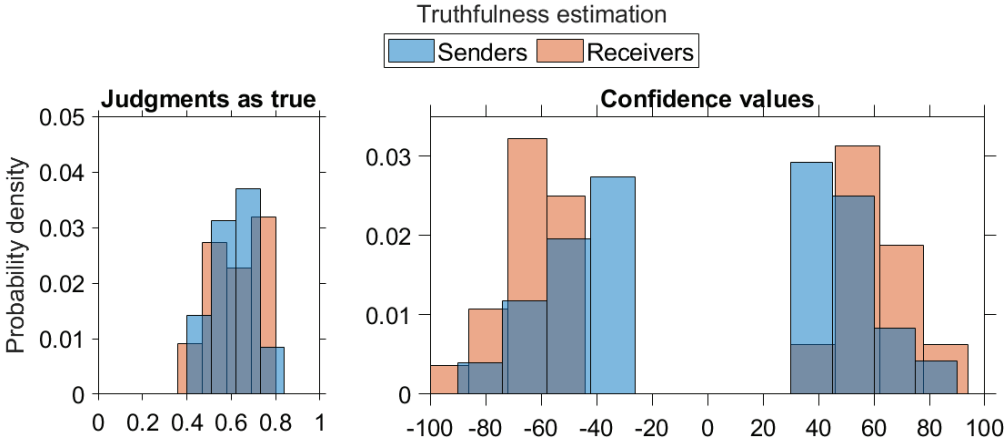
Average response time (RT) to estimate truthfulness was  $8.72 \pm 2.02$  seconds (control RT=8.97, cue RT=8.47). Participants were not better than chance at this estimation, with performances centered around 50.98% success (sd = 4.18) and average success rate higher in control condition ( $53.19 \pm 05.45\%$ ) than in cue condition ( $48.76 \pm 07.11$ ). They estimated truthfulness with a bias:  $61.2 \pm 09.82\%$  of information was declared as true (control condition:  $62.04 \pm 09.45\%$ ; cue condition:  $60.35 \pm 12.63\%$ ). We looked at the impact of truthfulness judgment on confidence degree. Average confidence value was  $49.9 \pm 14.35$  and neither main nor interaction effects of truthfulness with theme (all  $p > .1$ ) were significant in explaining confidence degree. Truthfulness judgment was close to being statistically significant (estimate=1.48, se=.77,  $p=.054$ ) (Table S1). We looked at differences between conditions posterior samples with  $\mu \sim N(50,10)$ . Delta of confidence values posterior samples between control condition and cue condition for judgments as true was equal to .004 (95% Credible Interval [-.869, .874]). Delta of confidence values posterior samples between conditions for judgments as false was equal to -.004 (95% Credible Interval [-.871, .879]). Therefore, despite a bias towards the true, participants' confidence in information truthfulness estimation didn't differ between judgments nor conditions.

**Table S1:** Summary of judgment and confidence MLMs.



<i>Predictors</i>	<b>judgment</b>			<b>confidence</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.72	1.22 – 2.44	<b>0.002</b>	48.25	42.55 – 53.94	<b>&lt;0.001</b>
theme	0.86	0.76 – 0.98	<b>0.024</b>	0.49	-0.75 – 1.74	0.438
truthfulness	1.02	0.68 – 1.51	0.941	-0.51	-4.33 – 3.31	0.793
confidence	1.00	1.00 – 1.01	<b>0.028</b>			
theme * truthfulness	1.02	0.85 – 1.23	0.797	0.01	-1.75 – 1.78	0.989
judgment				1.48	-0.03 – 2.98	0.054
<b>Random Effects</b>						
$\sigma^2$	3.29			411.21		
$\tau_{00}$	0.00	time		0.21	time	
	0.12	subject		200.80	subject	
	0.00	bloc		0.00	bloc	
N	32	subject		32	subject	
	2	bloc		2	bloc	
	96	time		96	time	
Observations	3072			3072		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.007 / NA			0.002 / NA		

After having established that participants behavior were stable across judgments and conditions, we compared Bayesian beta-binomial posterior probabilities of judgment between senders and receivers from which reception data was extracted. We found a delta of .03 (95% Credible Interval [.006, .064]) with higher judgments as true for receivers (.62 ± .02%). Average confidence value was 61.11 ± 12.28 for receivers, which is 11.2 points higher than senders'. We compared senders' Bayesian normal posterior distributions of confidence values with those of receivers', regardless of judgments. For a prior  $\mu \sim N(50,10)$  and common variance, we found a delta of -.115 (95% Credible Interval [-.068, .051], Deviance Information Criterion = 432.2), with a Cohen's d = -.008. Testing with another non-informative half-Cauchy prior for  $\mu$  where we set the two moments at 0 and 10, we found a delta of -11.18 (95% Credible Interval [-18.908, -3.254], Deviance Information Criterion = 423.2), with a Cohen's d = -.824. A lower deviance in Information Criterion reflecting a better fit, it is more likely that receivers from which reception data was extracted and senders were very different in truthfulness judgment. Given the low number of  $n$  in both senders' and receivers' groups, priors are assigned a great weight. In the case we consider both come from the same population, we find no differences between receivers and senders confidence values (Figure S1).



**Figure S1: Plots of senders’ proportion of judgments as true and confidence values against receivers.**

## VI.2. Preferences estimation

Table S2: Summary of send MLMs.

<i>Predictors</i>	send			send			send			send		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.08	0.80 – 1.44	0.616	0.49	0.38 – 0.64	<0.001	0.61	0.47 – 0.80	<0.001	0.69	0.58 – 0.84	<0.001
judgment	1.01	0.87 – 1.18	0.885	1.33	1.06 – 1.68	0.015	1.24	0.99 – 1.56	0.057	1.02	0.87 – 1.19	0.815
confidence	0.51	0.40 – 0.66	<0.001	0.93	0.82 – 1.06	0.274	0.96	0.85 – 1.09	0.537	0.80	0.73 – 0.87	<0.001
condition	0.75	0.64 – 0.87	<0.001									
confidence * condition	1.34	1.16 – 1.57	<0.001									
cue				0.62	0.55 – 0.69	<0.001						
dist btwn sender and receiver							1.29	1.14 – 1.48	<0.001			
condition * order										0.94	0.89 – 0.99	0.015
<b>Random Effects</b>												
$\sigma^2$	3.29			3.29			3.29			3.29		
$\tau_{00}$	0.01	time		0.04	time		0.05	time		0.02	time	
	0.16	subject		0.29	subject		0.25	subject		0.16	subject	
	0.00	condition										
ICC				0.09			0.08			0.05		
N	32	subject		32	subject		32	subject		32	subject	
	2	condition		48	time		48	time		96	time	
	96	time										
Observations	3072			1536			1536			3072		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.027 / NA			0.063 / 0.149			0.017 / 0.097			0.017 / 0.067		

## VI.3. Success in estimating others' preferences

MLMs shown that relying on their truthfulness estimation proved detrimental to senders' success in estimating receivers' preferences. The cue provided about receivers' distance to organizations proved useful for senders as it increased rates of success. These results indicate subjects in control condition may have overweighed their own beliefs about information in the preferences inference process.

**Table S3:** Summary of send success MLMs.

<i>Predictors</i>	<b>send_success</b>			<b>control_send_success</b>			<b>cue_send_success</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.71	0.54 – 0.94	<b>0.016</b>	1.01	0.89 – 1.14	0.902	1.38	1.21 – 1.57	<b>&lt;0.001</b>
judgment	1.06	0.91 – 1.23	0.472						
confidence	0.95	0.84 – 1.08	0.442						
condition	1.37	1.16 – 1.61	<b>&lt;0.001</b>						
cue	1.38	1.24 – 1.54	<b>&lt;0.001</b>				1.38	1.24 – 1.54	<b>&lt;0.001</b>
order	1.04	0.96 – 1.13	0.301						
judgment * confidence	1.15	0.99 – 1.34	0.064	1.12	0.99 – 1.28	0.081	1.09	0.95 – 1.25	0.244
<b>Random Effects</b>									
$\sigma^2$	3.29			3.29			3.29		
$\tau_{00}$	0.03	time		0.04	time		0.02	time	
	0.03	subject		0.02	subject		0.04	subject	
	0.00	condition							
ICC				0.02			0.02		
N	32	subject		32	subject		32	subject	
	2	condition		48	time		48	time	
	96	time							
Observations	3072			1536			1536		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.025 / NA			0.003 / 0.020			0.031 / 0.047		

## VI.4. Modelling estimation of others' preferences

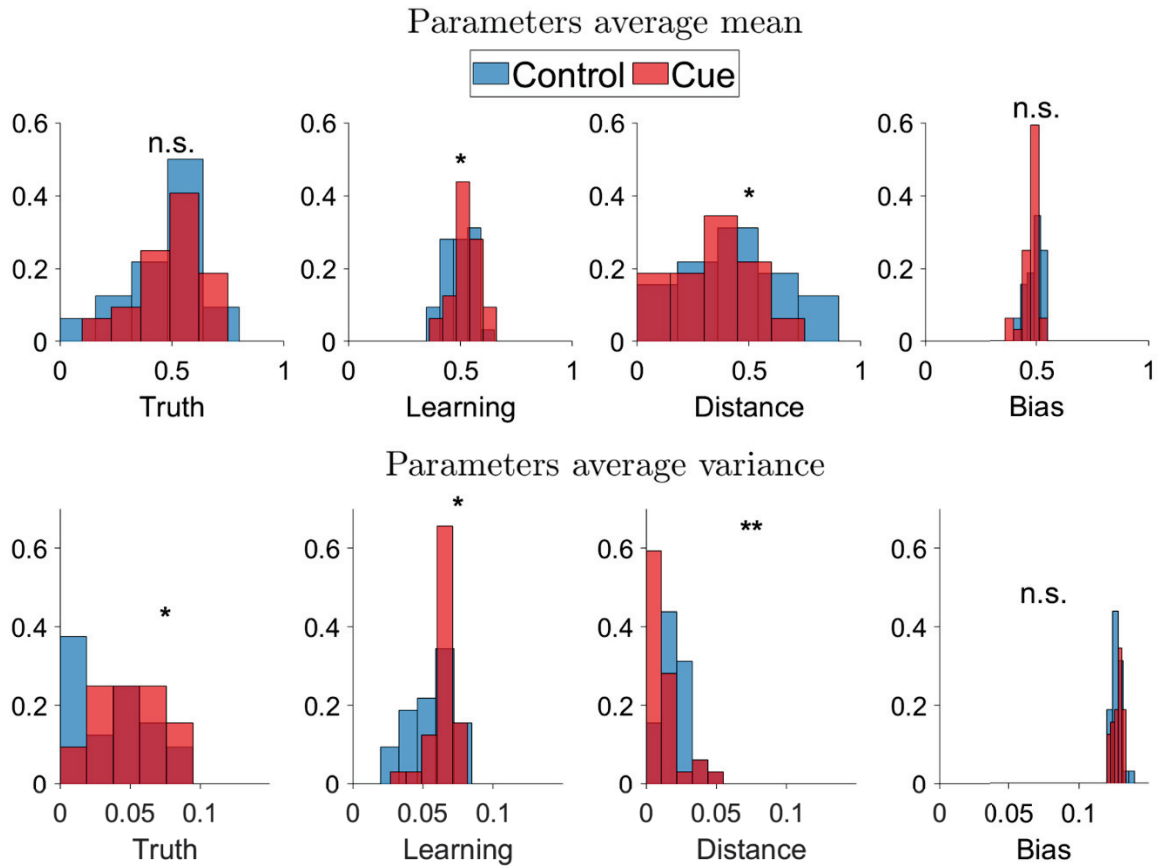
**Table S4:** MLM of send probability explained by modes of the Bayesian model parameters.

<i>Predictors</i>	<b>send</b>			<b>send</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.16	0.10 – 0.25	< <b>0.001</b>	0.23	0.13 – 0.41	< <b>0.001</b>
Mode truth	3.22	2.37 – 4.38	< <b>0.001</b>	1.48	0.75 – 2.95	0.259
Mode distance	6.53	4.27 – 10.00	< <b>0.001</b>	4.29	1.66 – 11.08	<b>0.003</b>
Mode learning	1.02	1.00 – 1.05	0.096	0.98	0.91 – 1.06	0.578
send RT	1.09	1.01 – 1.17	<b>0.020</b>	1.07	0.99 – 1.15	0.106
theme	1.06	0.93 – 1.20	0.416	1.07	0.94 – 1.22	0.273
judgment	0.86	0.70 – 1.07	0.188	1.25	1.00 – 1.57	<b>0.048</b>
<b>Random Effects</b>						
$\sigma^2$	3.29			3.29		
$\tau_{00}$	0.00 <sub>subject</sub>			0.21 <sub>subject</sub>		
ICC				0.06		
N	32 <sub>subject</sub>			32 <sub>subject</sub>		
Observations	1536			1536		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.208 / NA			0.028 / 0.086		

**Table S5:** MLM of send probability explained by modes of the Bayesian model parameters after introducing the mode of  $\theta$  beta distribution.

<i>Predictors</i>	<b>Send_control</b>			<b>Send_cue</b>		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.07	0.04 – 0.12	< <b>0.001</b>	0.03	0.02 – 0.06	< <b>0.001</b>
Mode truth	0.99	0.66 – 1.48	0.961	1.04	0.67 – 1.61	0.875
Mode distance	0.88	0.48 – 1.63	0.688	0.83	0.43 – 1.60	0.575
Mode learning	1.01	0.99 – 1.03	0.398	0.96	0.88 – 1.04	0.275
Mode final	127.24	43.00 – 376.50	< <b>0.001</b>	295.64	147.46 – 592.70	< <b>0.001</b>
send RT	1.08	1.00 – 1.17	<b>0.044</b>	1.05	0.97 – 1.13	0.247
theme	1.07	0.94 – 1.23	0.317	1.19	1.03 – 1.36	<b>0.017</b>
judgment	0.86	0.69 – 1.08	0.201	1.46	1.14 – 1.86	<b>0.002</b>
<b>Random Effects</b>						
$\sigma^2$	3.29			3.29		
$\tau_{00}$	0.00 <sub>subject</sub>			0.00 <sub>subject</sub>		
N	32 <sub>subject</sub>			32 <sub>subject</sub>		
Observations	1536			1536		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.200 / NA			0.293 / NA		

## VII. Modelling



**Figure S2: Participants update their beliefs and the confidence about their beliefs in the cue condition.** The mean is a property of beta distributions central tendency. In the context of our study, it represents participants' beliefs about the receivers' preferences. The variance represents the dispersion around the mean, representing participants' confidence in their beliefs. In the cue condition, average beliefs about receivers' preferences given the learning increase towards action  $a = to\ receive$ . However, participants' confidence in their beliefs decrease. In the contrary; average beliefs given the distance increase away from action  $a = to\ receive$ , with higher confidence, compared to control condition.



## VIII. fMRI analyses

**Table S6:** Summary of significant activations for fMRI GLM with behavioural variables. Voxel-level threshold:  $p < .001$ , uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. DLPFC: dorsolateral prefrontal cortex. VMPFC: ventromedial prefrontal cortex. DMPFC: dorsomedial prefrontal cortex. STS: superior temporal sulcus.

contrast: Send > Not send	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	p (cluster uncorrected)	peak voxel MNI coordinates		
<i>Increasing confidence, control condition</i>	midbrain	104	4.62	.035	.006	6	27	-6
	striatum	765	5.25	.000	.000	-12	15	-3 LR
	DLPFC	68	4.02	.118	.022	-42	42	15 L
	DLPFC	149	4.5	.009	.002	39	36	12 R
	VMPFC	55	4.04	.188	.036	-9	39	0
<i>Increasing distance, cue condition</i>	DMPFC	368	4.61	.000	.000	-15	48	36
	STS	569	4.31	.000	.000	-51	-9	-6 L
			4.5	.000	.000	-42	33	-12 L
	STS	93	4.15	.032	.005	-51	-42	-6 L
	STS	63	4.19	.105	.016	-39	-57	18 L
	STS	55	4.23	.147	.023	54	3	-15 R
	STS	57	4.16	.135	.021	63	-15	-12 R

**Table S7:** Summary of significant activations for fMRI GLM with Bayesian parameters. Voxel-level threshold:  $p < .001$ , uncorrected. FWE: familywise error. MNI: Montreal Neurological Institute. L: left. R: right. VLPFC: ventrolateral prefrontal cortex. VMPFC: ventromedial prefrontal cortex. DMPFC: dorsomedial prefrontal cortex. TPJ: temporo-parietal junction.

contrast: Decisions not to send	label	voxels at $p < .001$	peak z-score	p (cluster FWE corrected)	p (cluster uncorrected)	peak voxel MNI coordinates		
<i>Decreasing <math>\theta</math>(truth), control condition</i>	striatum	73	3.97	.101	.016	-9	15	-6 L
	VMPFC	52	3.49	.218	.037	-3	30	-9
	DLPFC	53	3.91	.210	.035	39	36	12 R
<i>Decreasing <math>\theta</math>(distance), cue condition</i>	TPJ	324	3.8	.000	.000	-39	-69	18 L
	TPJ	228	4.76	.001	.000	45	-63	15 R
contrast: Time of sending decision								
<i>Decreasing <math>\theta</math>, both conditions</i>	VLPFC	44	3.63	.161	.017	24	51	0 R

# Chapter III. Supplementary

## Materials

### I. Questionnaires

#### Beck Depression Inventory(BDI)

The Beck Depression Inventory (BDI, BDI-II), created by Dr. A.T. Beck, is a 21-question multiple-choice self-report inventory, one of the most widely used instruments for measuring the severity of depression. In its current version the questionnaire is designed for individuals aged 13 and over, and is composed of items relating to symptoms of depression such as hopelessness and irritability, cognitions such as guilt or feelings of being punished, as well as physical symptoms such as fatigue, weight loss, and lack of interest in sex.

#### Machiavellian Personality Test (MACH-IV)

Machiavellianism is a term that some social and personality psychologist use to describe a person's tendency to deceive and manipulate others for personal gain. In the 1960s, Richard Christie and Florence L. Geis developed a test for measuring a person's level of Machiavellianism. This eventually became the MACH-IV test, a twenty-statement personality survey that is now the standard self-assessment tool of Machiavellianism. People scoring above 60 out of 100 on the MACH-IV are considered high Machs. People scoring below 60 out of 100 on the MACH-IV are considered low Machs.

#### Beck Anxiety Inventory(BAI)

The Beck Anxiety Inventory (BAI) is a 21-question multiple-choice self-report inventory that is used for measuring the severity of an individual's anxiety. These questions concern how the subject has been feeling in the last week, expressed as common symptoms of anxiety (such as numbness, hot and cold sweats, or feelings of dread).

#### Barratt Impulsiveness Scale

The Barratt Impulsiveness Scale (BIS) is one of the oldest and most widely used measures of impulsive personality traits. The BIS-11 is a 30-item self-report questionnaire, that is scored to yield a total score, three second-order factors (Attentional, Motor, Nonplanning), and six first-order factors (Attention, Motor, Self-Control, Cognitive Complexity, Perseverance, Cognitive Instability)

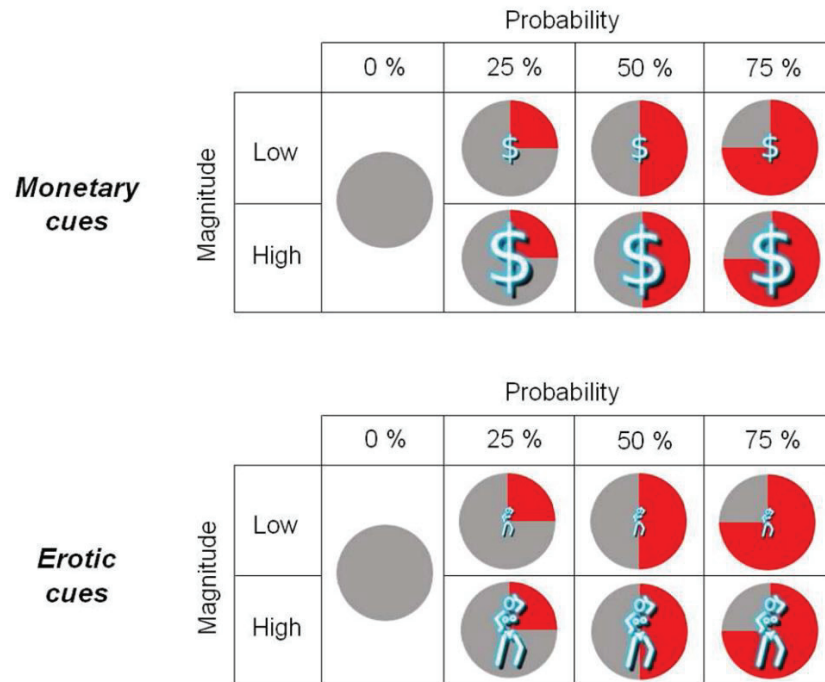
International Personality Item Pool(IPIP) (<http://ipip.ori.org/>)

- 1) Dominance
- 2) Anger
- 3) Emotional Stability
- 4) Leadership
- 5) Empathy
- 6) Machiavellianism
- 7) Risk Taking
- 8) Anxiety
- 9) Conformity(Cooperativeness)
- 10) Sociability
- 11) Social Confidence
- 12) Behavioural Inhibition System(BIS): Anxiety
- 13) Behavioural Approaching/Activation System(BAS)/Fun-Seeking
- 14) BAS/Drive
- 15) BAS/Reward-Responsiveness

Post-hoc survey questions after scanning were:

- Do you think that you were injected with placebo (inactive substance) or testosterone yesterday?
- How do you think that testosterone administration would modify any given person's behavior?
- Would you expect testosterone administration to affect your behavior if you were to receive it?

## II. Results



**Figure S1: Reward cues.** The pictogram indicated the type of reward (Up: monetary, Down: erotic), its magnitude (Low vs High) and its probability (0%: control trials, 25%, 50% and 75%).

<i>Predictors</i>	response			response			response		
	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>	<i>Odds Ratios</i>	<i>CI</i>	<i>p</i>
(Intercept)	10.88	5.38 – 21.97	<0.001	20.60	9.12 – 46.53	<0.001	20.40	8.54 – 48.74	<0.001
treatment	0.66	0.40 – 1.08	0.095	0.23	0.11 – 0.51	<0.001	0.24	0.09 – 0.60	0.002
proba	1.01	1.01 – 1.02	<0.001	1.00	0.99 – 1.01	0.793	1.01	1.01 – 1.02	<0.001
type	1.21	0.94 – 1.56	0.143	1.21	0.94 – 1.56	0.142	1.21	0.94 – 1.56	0.143
intensity	1.19	0.92 – 1.53	0.183	1.19	0.92 – 1.53	0.182	0.78	0.52 – 1.18	0.239
treatment * proba				1.02	1.01 – 1.04	0.001			
treatment * intensity							1.99	1.18 – 3.36	0.010
<b>Random Effects</b>									
$\sigma^2$	3.29			3.29			3.29		
$\tau_{00}$	0.42	subject		0.42	subject		0.42	subject	
ICC	0.11			0.11			0.11		
N	39	subject		39	subject		39	subject	
Observations	5616			5616			5616		
Marginal $R^2$ / Conditional $R^2$	0.032 / 0.141			0.037 / 0.147			0.037 / 0.147		

**Table S1: Summary of hit rates MLMs.**

<i>Predictors</i>	response_time			response_time			response_time		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	666.15	631.98 – 700.31	<0.001	653.48	618.20 – 688.76	<0.001	673.91	638.97 – 708.84	<0.001
treatment	-36.30	-70.08 – -2.51	0.035	-12.01	-49.83 – 25.81	0.534	-51.22	-87.79 – -14.66	0.006
type	-19.19	-24.86 – -13.52	<0.001	-10.77	-18.96 – -2.58	0.010	-19.19	-24.85 – -13.52	<0.001
proba	-0.14	-0.28 – -0.00	0.044	-0.14	-0.28 – -0.00	0.044	-0.30	-0.50 – -0.10	0.004
intensity	-17.46	-23.13 – -11.79	<0.001	-17.46	-23.13 – -11.80	<0.001	-17.45	-23.11 – -11.78	<0.001
miss	3.05	0.19 – 5.90	0.037	3.05	0.19 – 5.90	0.036	3.03	0.17 – 5.89	0.038
outliers	-14.29	-23.93 – -4.65	0.004	-14.30	-23.93 – -4.66	0.004	-14.29	-23.94 – -4.64	0.004
treatment * type				-16.16	-27.49 – -4.82	0.005			
treatment * proba							0.30	0.02 – 0.58	0.036
<b>Random Effects</b>									
$\sigma^2$	11397.35			11381.07			11388.00		
$\tau_{00}$	2628.14 <sub>subject</sub>			2624.60 <sub>subject</sub>			2631.00 <sub>subject</sub>		
ICC	0.19			0.19			0.19		
N	40 <sub>subject</sub>			40 <sub>subject</sub>			40 <sub>subject</sub>		
Observations	5456			5456			5456		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.086 / 0.257			0.087 / 0.258			0.087 / 0.258		

Table S2: Summary of response times MLMs.

<i>Predictors</i>	rating			rating			rating		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.15	-0.35 – 0.65	0.557	-0.12	-0.65 – 0.42	0.665	0.08	-0.45 – 0.62	0.758
treatment	0.17	-0.39 – 0.73	0.546	0.68	0.03 – 1.34	0.041	0.30	-0.35 – 0.96	0.368
type	0.93	0.81 – 1.04	<0.001	1.10	0.94 – 1.27	<0.001	0.93	0.81 – 1.04	<0.001
proba	-0.00	-0.00 – 0.00	0.428	-0.00	-0.00 – 0.00	0.440	-0.00	-0.00 – 0.00	0.427
intensity	2.52	2.41 – 2.64	<0.001	2.52	2.41 – 2.64	<0.001	2.57	2.41 – 2.73	<0.001
treatment * type				-0.34	-0.56 – -0.11	0.003			
treatment * intensity							-0.09	-0.31 – 0.14	0.457
<b>Random Effects</b>									
$\sigma^2$	2.29			2.28			2.29		
$\tau_{00}$	0.78 <sub>subject</sub>			0.78 <sub>subject</sub>			0.78 <sub>subject</sub>		
ICC	0.25			0.26			0.25		
N	40 <sub>subject</sub>			40 <sub>subject</sub>			40 <sub>subject</sub>		
Observations	2765			2765			2765		
Marginal R <sup>2</sup> / Conditional R <sup>2</sup>	0.371 / 0.531			0.373 / 0.533			0.371 / 0.531		

Table S: Summary of ratings MLMs.

# References

- Aarts, H., & Van Honk, J. (2009). Testosterone and unconscious positive priming increase human motivation separately. *NeuroReport*, 20(14), 1300–1303. <https://doi.org/10.1097/WNR.0b013e3283308cdd>
- Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Computational Biology*, 11(10), 1–23. <https://doi.org/10.1371/journal.pcbi.1004519>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>
- Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32, 201–271.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111(2), 256–274. <https://doi.org/10.1037/0033-2909.111.2.256>
- Anderson, R. A., Bancroft, J., & Wu, F. C. W. (1992). The Effects of Exogenous Testosterone on Sexuality and Mood of Normal Men. *Endocrinology And Metabolism*, 75(6), 1503–1507.
- Apicella, C. L., Carré, J. M., & Dreber, A. (2015). Testosterone and Economic Risk Taking: A Review. *Adaptive Human Behavior and Physiology*, 1(3), 358–385. <https://doi.org/10.1007/s40750-014-0020-2>
- Apps, M. A. J., Lockwood, P. L., & Balsters, J. H. (2013). The role of the midcingulate cortex in monitoring others' decisions. *Frontiers in Neuroscience*, 7(7 DEC), 1–7. <https://doi.org/10.3389/fnins.2013.00251>
- Apps, M. A. J., Rushworth, M. F. S., & Chang, S. W. C. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, 90(4), 692–707. <https://doi.org/10.1016/j.neuron.2016.04.018>
- Arad, A., & Rubinstein, A. (2012). The 11 – 20 Money Request Game : *American Economic Review*, 102(7), 3561–3573.

- Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience and Biobehavioral Reviews*, 30(3), 319–345. <https://doi.org/10.1016/j.neubiorev.2004.12.007>
- Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2021). Echo Chambers, Filter Bubbles, and Polarisation: a Literature Review. *The Royal Society*, 1–42.
- Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110(3), 486–498. <https://doi.org/10.1037/0033-2909.110.3.486>
- Armor, D. A., & Taylor, S. E. (2012). When Predictions Fail: The Dilemma of Unrealistic Optimism. *Heuristics and Biases*, 334–347. <https://doi.org/10.1017/cbo9780511808098.021>
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1), 95–113. <https://doi.org/10.1016/j.neuroimage.2007.07.007>
- Ashburner, J., & Friston, K. J. (2009). Computing average shaped tissue probability templates. *NeuroImage*, 45(2), 333–341. <https://doi.org/10.1016/j.neuroimage.2008.12.008>
- Aubele, T., & Kritzer, M. F. (2012). Androgen influence on prefrontal dopamine systems in adult male rats: Localization of cognate intracellular receptors in medial prefrontal projections to the ventral tegmental area and effects of gonadectomy and hormone replacement on glutamate-stimulated. *Cerebral Cortex*, 22(8), 1799–1812. <https://doi.org/10.1093/cercor/bhr258>
- Badre, D., & D’Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9), 659–669. <https://doi.org/10.1038/nrn2667>
- Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, 22(2), 170–188. <https://doi.org/10.1016/j.tics.2017.11.005>
- Baek, E. C., Scholz, C., O’Donnell, M. B., & Falk, E. B. (2017). The Value of Sharing Information: A Neural Account of Information Transmission. *Psychological Science*, 28(7), 851–861. <https://doi.org/10.1177/0956797617695073>
- Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8), 1608–1613. <https://doi.org/10.1037/xge0000729>



- Bail, C. A., Guay, B., Maloney, E., Combs, A., Sunshine Hillygus, D., Merhout, F., ... Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(1), 243–250. <https://doi.org/10.1073/pnas.1906420116>
- Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., ... West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-022-01388-6>
- Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *ELife*, *9*, 1–25. <https://doi.org/10.7554/eLife.56477>
- Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087. <https://doi.org/10.1073/pnas.1800795115>
- Bang, D., Moran, R., Daw, N. D., & Fleming, S. M. (2021). Neurocomputational mechanisms of confidence in self and others. *BioRxiv*, 2021.03.05.434065. Retrieved from <https://www.biorxiv.org/content/10.1101/2021.03.05.434065v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.03.05.434065v1.abstract>
- Barasch, A. (2020). The consequences of sharing. *Current Opinion in Psychology*, *31*, 61–66. <https://doi.org/10.1016/j.copsyc.2019.06.027>
- Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size affects what people share. *Journal of Marketing Research*, *51*(3), 286–299. <https://doi.org/10.1509/jmr.13.0238>
- Baron, J., & Jost, J. T. (2019). False Equivalence: Are Liberals and Conservatives in the United States Equally Biased? *Perspectives on Psychological Science*, *14*(2), 292–303. <https://doi.org/10.1177/1745691618788876>
- Barreto, M. D. S., Caram, C. D. S., Santos, J. L. G. Dos, de Souza, R. R., Goes, H. L. D. F., & Marcon, S. S. (2021). Fake news about the COVID-19 pandemic: perception of health professionals and their families. *Revista Da Escola de Enfermagem*, *55*, 1–9. <https://doi.org/10.1590/1980-220X-REEUSP-2021-0007>
- Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based

- meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage*, 76, 412–427. <https://doi.org/10.1016/j.neuroimage.2013.02.063>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Beck, A. T., & Beck, R. W. (1972). Screening depressed patients in family practice. A rapid technic. *Postgraduate Medicine*, 52(6), 81–85. <https://doi.org/10.1080/00325481.1972.11713319>
- Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science*, 324(5931), 1160–1164. <https://doi.org/10.1126/science.1169694>
- Behzadi, Y., Restom, K., Liou, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1), 90–101. <https://doi.org/10.1016/j.neuroimage.2007.04.042>
- Belke, B., Leder, H., Strobach, T., & Carbon, C. C. (2010). Cognitive Fluency: High-Level Processing Dynamics in Art Appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, 4(4), 214–222. <https://doi.org/10.1037/a0019648>
- Belot, M., & van de Ven, J. (2017). How private is private information? The ability to spot deception in an economic game. *Experimental Economics*, 20(1), 19–43. <https://doi.org/10.1007/s10683-015-9474-8>
- Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'Economie Politique*, 125(5), 665–685. <https://doi.org/10.3917/redp.255.0665>
- Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–164. <https://doi.org/10.1257/jep.30.3.141>
- Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, 24(4), 586–607.
- Berns, G. S., Chappelow, J., Cekic, M., Zink, C. F., Pagnoni, G., & Martin-Skurski, M. E. (2006). Neurobiological substrates of dread. *Science*, 312(5774), 754–758. <https://doi.org/10.1126/science.1123721>

- Bialy, M., & Sachs, B. D. (2002). Androgen implants in medial amygdala briefly maintain noncontact erection in castrated male rats. *Hormones and Behavior*, *42*(3), 345–355. <https://doi.org/10.1006/hbeh.2002.1821>
- Birrell, J., Meares, K., Wilkinson, A., & Freeston, M. (2011). Toward a definition of intolerance of uncertainty: A review of factor analytical studies of the Intolerance of Uncertainty Scale. *Clinical Psychology Review*, *31*(7), 1198–1208. <https://doi.org/10.1016/j.cpr.2011.07.009>
- Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron*, *85*(3), 602–614. <https://doi.org/10.1016/j.neuron.2014.12.050>
- Blank, G. (2017). The Digital Divide Among Twitter Users and Its Implications for Social Research. *Social Science Computer Review*, *35*(6), 679–697. <https://doi.org/10.1177/0894439316671698>
- Blank, G., & Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, *61*(7), 741–756. <https://doi.org/10.1177/0002764217717559>
- Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11818–11823. <https://doi.org/10.1073/pnas.191355898>
- Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1), 1–12. <https://doi.org/10.1093/nc/niz004>
- Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484. <https://doi.org/10.1523/JNEUROSCI.0797-14.2015>
- Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior*, *36*(2), 235–262. <https://doi.org/10.1007/s11109-013-9238-0>
- Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and*

- Social Psychology Review*, 10(3), 214–234.  
[https://doi.org/10.1207/s15327957pspr1003\\_2](https://doi.org/10.1207/s15327957pspr1003_2)
- Bond, C. F., & DePaulo, B. M. (2008). Individual Differences in Judging Deception: Accuracy and Bias. *Psychological Bulletin*, 134(4), 477–492. <https://doi.org/10.1037/0033-2909.134.4.477>
- Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., & Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*, 62(5), 733–743.  
<https://doi.org/10.1016/j.neuron.2009.05.014>
- Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual choice and learning in a Neural Network centered on human lateral frontopolar cortex. *PLoS Biology*, 9(6).  
<https://doi.org/10.1371/journal.pbio.1001093>
- Bos, L., Kruikemeier, S., & De Vreese, C. (2016). Nation binding: How public service broadcasting mitigates political selective exposure. *PLoS ONE*, 11(5), 1–11.  
<https://doi.org/10.1371/journal.pone.0155112>
- Bos, P. A., Hermans, E. J., Ramsey, N. F., & Van Honk, J. (2012). The neural mechanisms by which testosterone acts on interpersonal trust. *NeuroImage*, 61(3), 730–737.  
<https://doi.org/10.1016/j.neuroimage.2012.04.002>
- Bos, P. A., Panksepp, J., Bluthé, R. M., & Honk, J. van. (2012). Acute effects of steroid hormones and neuropeptides on human social-emotional behavior: A review of single administration studies. *Frontiers in Neuroendocrinology*, 33(1), 17–35.  
<https://doi.org/10.1016/j.yfrne.2011.01.002>
- Boschin, E. A., Piekema, C., & Buckley, M. J. (2015). Essential functions of primate frontopolar cortex in cognition. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9), E1020–E1027.  
<https://doi.org/10.1073/pnas.1419649112>
- Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online. *Perspectives on Psychological Science*, 15(4), 978–1010.  
<https://doi.org/10.1177/1745691620917336>

- Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using an SPM toolbox. *NeuroImage*, *16*, 497.
- Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30. <https://doi.org/10.1037/1076-898X.12.1.11>
- Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, *63*(1), 119–126. <https://doi.org/10.1016/j.neuron.2009.06.009>
- Bromberg-Martin, E. S., & Monosov, I. E. (2020). Neural circuitry of information seeking. *Current Opinion in Behavioral Sciences*, *35*, 62–70. <https://doi.org/10.1016/j.cobeha.2020.07.006>
- Bromberg-Martin, E. S., & Sharot, T. (2020). The Value of Beliefs. *Neuron*, *106*(4), 561–565. <https://doi.org/10.1016/j.neuron.2020.05.001>
- Brunnermeier, M. K., Parker, J. A., Abel, A., Bénabou, R., Bernheim, D., Caplin, A., ... Veldkamp, L. (2003). Optimal Expectations. *NATIONAL BUREAU OF ECONOMIC RESEARCH WORKING PAPER SERIES*. Retrieved from <http://www.nber.org/papers/w10707>
- Brydevall, M., Bennett, D., Murawski, C., & Bode, S. (2018). The neural encoding of information prediction errors during non-instrumental information seeking. *Scientific Reports*, *8*(1), 1–11. <https://doi.org/10.1038/s41598-018-24566-x>
- Budescu, D. V., Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, *113*(2), 181–200. <https://doi.org/10.1007/s10584-011-0330-3>
- Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal*, *10*(1), 395–411. <https://doi.org/10.32614/rj-2018-017>
- Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5). <https://doi.org/10.18637/JSS.V100.I05>

- Camerer, C. F., & Weber, M. (1991). Recent developments in modelling preferences: Uncertainty and ambiguity. *Manuskripte Aus Den Instituten Für Betriebswirtschaftslehre Der Universität Kiel*, (275).
- Carmichael, S. T., & Price, J. L. (1995). Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. *Journal of Comparative Neurology*, 363(4), 642–664. <https://doi.org/10.1002/cne.903630409>
- Carré, J. M., Campbell, J. A., Lozoya, E., Goetz, S. M. M., & Welker, K. M. (2013). Changes in testosterone mediate the effect of winning on subsequent aggressive behaviour. *Psychoneuroendocrinology*, 38(10), 2034–2041. <https://doi.org/10.1016/j.psyneuen.2013.03.008>
- Carré, J. M., McCormick, C. M., & Hariri, A. R. (2011). The social neuroendocrinology of human aggression. *Psychoneuroendocrinology*, 36(7), 935–944. <https://doi.org/10.1016/j.psyneuen.2011.02.001>
- Carver, C. S., & White, T. L. (1994). Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, 67(2), 319–333. <https://doi.org/10.1037/0022-3514.67.2.319>
- Cauda, F., D'Agata, F., Sacco, K., Duca, S., Geminiani, G., & Vercelli, A. (2011). Functional connectivity of the insula in the resting brain. *NeuroImage*, 55(1), 8–23. <https://doi.org/10.1016/j.neuroimage.2010.11.049>
- Cauda, F., Geminiani, G. C., & Vercelli, A. (2014). Evolutionary appearance of von Economo's neurons in the mammalian cerebral cortex. *Frontiers in Human Neuroscience*, 8(MAR), 1–11. <https://doi.org/10.3389/fnhum.2014.00104>
- Chang, S. W. C., Gariépy, J. F., & Platt, M. L. (2013). Neuronal reference frames for social decisions in primate frontal cortex. *Nature Neuroscience*, 16(2), 243–250. <https://doi.org/10.1038/nn.3287>
- Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, 115(31), E7255–E7264. <https://doi.org/10.1073/pnas.1800547115>

- Chiao, J. Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., ... Iidaka, T. (2009). Neural basis of individualistic and collectivistic views of self. *Human Brain Mapping, 30*(9), 2813–2820. <https://doi.org/10.1002/hbm.20707>
- Chiao, J. Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., ... Iidaka, T. (2010). Dynamic cultural influences on neural representations of the self. *Journal of Cognitive Neuroscience, 22*(1), 1–11. <https://doi.org/10.1162/jocn.2009.21192>
- Christie, R., & Geis, F. L. (1970). *Studies in machiavellianism*.
- Cloutman, L. L., Binney, R. J., Drakesmith, M., Parker, G. J. M., & Lambon Ralph, M. A. (2012). The variation of function across the human insula mirrors its patterns of structural connectivity: Evidence from in vivo probabilistic tractography. *NeuroImage, 59*(4), 3514–3521. <https://doi.org/10.1016/j.neuroimage.2011.11.016>
- Coggan, D. D., Baker, D. H., & Andrews, T. J. (2016). The role of visual and semantic properties in the emergence of category-specific patterns of neural response in the human brain. *ENeuro, 3*(4), 821–825. <https://doi.org/10.1523/ENEURO.0158-16.2016>
- Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2022). Independent and Interacting Value Systems for Reward and Information in the Human Brain. *ELife, 11*, 1–22. <https://doi.org/10.7554/eLife.66358>
- Cogliati Dezza, I., Maher, C., & Sharot, T. (2022). People adaptively use information to improve their internal states and external outcomes. *Cognition, 228*(July), 105224. <https://doi.org/10.1016/j.cognition.2022.105224>
- Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 933–942. <https://doi.org/10.1098/rstb.2007.2098>
- Cosme, D., Scholz, C., Chan, H. Y., Doré, B. P., Pandey, P., Carreras-Tartak, J., ... Falk, E. B. (2022). Message Self and Social Relevance Increases Intentions to Share Content: Correlational and Causal Evidence From Six Studies. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001270>
- Czekalla, N., Stierand, J., Stolz, D. S., Mayer, A. V., Voges, J. F., Rademacher, L., ... Müller-Pinzler, L. (2021). Self-beneficial belief updating as a coping mechanism for stress-



- induced negative affect. *Scientific Reports*, *11*(1), 1–13. <https://doi.org/10.1038/s41598-021-96264-0>
- Dabbs, J. M., & Mohammed, S. (1992). Male and female salivary testosterone concentrations before and after sexual activity. *Physiology and Behavior*, *52*(1), 195–197. [https://doi.org/10.1016/0031-9384\(92\)90453-9](https://doi.org/10.1016/0031-9384(92)90453-9)
- Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. <https://doi.org/10.1007/s00199-006-0153-z>
- Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. <https://doi.org/10.1038/nature04766>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. <https://doi.org/10.1038/nn.3279>
- Deaner, R. O., Khera, A. V., & Platt, M. L. (2005). Monkeys pay per view: Adaptive valuation of social images by rhesus macaques. *Current Biology*, *15*(6), 543–548. <https://doi.org/10.1016/j.cub.2005.01.044>
- Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science*, *29*(5), 761–778. <https://doi.org/10.1177/0956797617744771>
- Dieckmann, N. F., Gregory, R., Peters, E., & Hartman, R. (2017). Seeing What You Want to See: How Imprecise Uncertainty Ranges Enhance Motivated Reasoning. *Risk Analysis*, *37*(3), 471–486. <https://doi.org/10.1111/risa.12639>
- Diekhof, E. K., Kaps, L., Falkai, P., & Gruber, O. (2012). The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude - An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. *Neuropsychologia*, *50*(7), 1252–1266. <https://doi.org/10.1016/j.neuropsychologia.2012.02.007>
- Dimeo, A. N., & Wood, R. I. (2006). Self-administration of estrogen and dihydrotestosterone in male hamsters. *Hormones and Behavior*, *49*(4), 519–526. <https://doi.org/10.1016/j.yhbeh.2005.11.003>

- Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., ... Zinger, J. F. (2019). At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291. <https://doi.org/10.1177/1745691617746796>
- Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486. <https://doi.org/10.1126/science.1252254>
- Dreher, J. C., Koechlin, E., Tierney, M., & Grafman, J. (2008). Damage to the fronto-polar cortex is associated with impaired multitasking. *PLoS ONE*, *3*(9). <https://doi.org/10.1371/journal.pone.0003227>
- Dwenger, N., & Lohse, T. (2019). Do individuals successfully cover up their lies? Evidence from a compliance experiment. *Journal of Economic Psychology*, *71*(August), 74–87. <https://doi.org/10.1016/j.joep.2018.08.007>
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. <https://doi.org/10.1038/s44159-021-00006-y>
- Eil, D., & Rao, J. M. (2011). American Economic Association The Good News-Bad News Effect: Asymmetric Processing of Objective Information about. *Journal: Microeconomics*, *3*(2), 114–138. Retrieved from <https://www.jstor.org/stable/pdf/41237187.pdf?refreqid=excelsior%3A313f5ea98f2b7fa0193d8ac934546e9c>
- Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*, 0–12. <https://doi.org/10.37016/mr-2020-71>
- Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *Review of Economic Studies*, *83*(2), 587–628. <https://doi.org/10.1093/restud/rdv051>
- Exley, C. L., & Kessler, J. B. (2019). Motivated errors. *National Bureau of Economic Research*.
- Exley, C. L., & Kessler, J. B. (2021). Information Avoidance and Image Concerns. *National*

- Bureau of Economic Research*. Retrieved from <http://www.nber.org/data-appendix/w28376>
- Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, *6*(1), 21–29. [https://doi.org/10.1016/0191-8869\(85\)90026-1](https://doi.org/10.1016/0191-8869(85)90026-1)
- Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin and Review*, *26*(5), 1705–1710. <https://doi.org/10.3758/s13423-019-01651-4>
- Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences*, *11*(10), 419–427. <https://doi.org/10.1016/j.tics.2007.09.002>
- FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. <https://doi.org/10.1038/s41562-019-0590-x>
- Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(Specialissue1), 298–320. <https://doi.org/10.1093/poq/nfw006>
- Flechsig, P. (1901). Developmental (Myelogenetic) Localisation of the Cerebral Cortex in the Human Subject. *The Lancet*, *158*(4077), 1027–1030. [https://doi.org/10.1016/S0140-6736\(01\)01429-5](https://doi.org/10.1016/S0140-6736(01)01429-5)
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. <https://doi.org/10.1037/rev0000045>
- Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117–6125. <https://doi.org/10.1523/JNEUROSCI.6489-11.2012>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(JULY), 1–9. <https://doi.org/10.3389/fnhum.2014.00443>
- Fletcher, R., Newman, N., & Schulz, A. (2020). A Mile Wide, an Inch Deep: Online News and Media Use in the 2019 UK General Election. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3582441>

- Fletcher, R., Robertson, C. T., & Nielsen, R. K. (2021). How Many People Live in Politically Partisan Online News Echo Chambers in Different Countries? *Journal of Quantitative Description: Digital Media*, 1, 1–56. <https://doi.org/10.51685/jqd.2021.020>
- Foerster, M., & van der Weele, J. J. (2018). Denial and Alarmism in Collective Action Problems. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3135783>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- Freddi, E. (2021). Do People Avoid Morally Relevant Information? Evidence From the Refugee Crisis. *Review of Economics and Statistics*, 103(4), 605–620. [https://doi.org/10.1162/rest\\_a\\_00934](https://doi.org/10.1162/rest_a_00934)
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O’Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, 68, 862–879. <https://doi.org/10.1016/j.neubiorev.2016.06.022>
- Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4), 187–224. <https://doi.org/10.1080/17588928.2015.1020053>
- Frye, C. A., Rhodes, M. E., Rosellini, R., & Svare, B. (2002). The nucleus accumbens as a site of action for rewarding properties of testosterone and its 5 $\alpha$ -reduced metabolites. *Pharmacology Biochemistry and Behavior*, 74(1), 119–127. [https://doi.org/10.1016/S0091-3057\(02\)00968-1](https://doi.org/10.1016/S0091-3057(02)00968-1)
- Galvan, A., Hare, T. A., Davidson, M., Spicer, J., Glover, G., & Casey, B. J. (2005). The role of ventral frontostriatal circuitry in reward-based learning in humans. *Journal of Neuroscience*, 25(38), 8650–8656. <https://doi.org/10.1523/JNEUROSCI.2431-05.2005>
- Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication*, 14(2), 265–285. <https://doi.org/10.1111/j.1083-6101.2009.01440.x>
- Ghaziri, J., Tucholka, A., Girard, G., Houde, J. C., Boucher, O., Gilbert, G., ... Nguyen, D. K. (2017). The Corticocortical Structural Connectivity of the Human Insula. *Cerebral Cortex (New York, N.Y. : 1991)*, 27(2), 1216–1228. <https://doi.org/10.1093/cercor/bhv308>

- Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *ELife*, 7, 1–28. <https://doi.org/10.7554/eLife.38293>
- Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, 1(1), 107–143. <https://doi.org/10.1111/j.1756-8765.2008.01006.x>
- Globig, A. L. K., Holtz, N., & Sharot, T. (2022). *Changing the Incentive Structure of Social Media Platforms to Halt the Spread of Misinformation*. 1–26.
- Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., ... Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21), 8174–8179. <https://doi.org/10.1073/pnas.0402680101>
- Gold J, & Shadlen M. (2001). Neural computations that underlie decisions about sensory stimuli. *TRENDS in Cognitive Sciences*, 5(1), 10–16.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, Vol. 7, pp. 7–28.
- Goldberg, Lewis R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84–96. <https://doi.org/10.1016/j.jrp.2005.08.007>
- Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1), 96–135. <https://doi.org/10.1257/jel.20151245>
- Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, 5(3), 143–164. <https://doi.org/10.1037/dec0000068>
- Golman, R., Loewenstein, G., Molnar, A., & Saccardo, S. (2021). The Demand for, and Avoidance of, Information. *Management Science*. <https://doi.org/10.1287/mnsc.2021.4244>
- Gomes, C. M., & Boesch, C. (2009). Wild chimpanzees exchange meat for sex on a long-term

- basis. *PLoS ONE*, 4(4), 16–18. <https://doi.org/10.1371/journal.pone.0005116>
- Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, 15(2), 56–67. <https://doi.org/10.1016/j.tics.2010.12.004>
- Graham, J. M., & Desjardins, C. (1980). Classical conditioning: Induction of luteinizing hormone and testosterone secretion in anticipation of sexual activity. *Science*, 210(4473), 1039–1041. <https://doi.org/10.1126/science.7434016>
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425), 374–378. <https://doi.org/10.1126/science.aau2706>
- Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, 60(11), 2659–2665. <https://doi.org/10.1287/mnsc.2014.1989>
- Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1), 173–217. <https://doi.org/10.1093/jeea/jvw001>
- Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. *Neuron*, 84(2), 486–496. <https://doi.org/10.1016/j.neuron.2014.08.060>
- Guess, A. M., Barberá, P., Munzert, S., & Yang, J. H. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences of the United States of America*, 118(14), 1–8. <https://doi.org/10.1073/pnas.2013464118>
- Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35(1), 4–26. <https://doi.org/10.1038/npp.2009.129>
- Hamann, S., Herman, R. A., Nolan, C. L., & Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. *Nature Neuroscience*, 7(4), 411–416. <https://doi.org/10.1038/nn1208>
- Hamann, S., Stevens, J., Vick, J. H., Bryk, K., Quigley, C. A., Berenbaum, S. A., & Wallen, K. (2014). Brain responses to sexual images in 46,XY women with complete androgen insensitivity syndrome are female-typical. *Hormones and Behavior*, 66(5), 724–730. <https://doi.org/10.1016/j.yhbeh.2014.09.013>



- Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., ... Duncan, J. S. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping, 19*(4), 224–247. <https://doi.org/10.1002/hbm.10123>
- Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *Annals of the American Academy of Political and Social Science, 659*(1), 63–76. <https://doi.org/10.1177/0002716215570866>
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin, 135*(4), 555–588. <https://doi.org/10.1037/a0015701>
- Hayden, B. Y., Parikh, P. C., Deaner, R. O., & Platt, M. L. (2007). Economic principles motivating social attention in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*(1619), 1751–1756. <https://doi.org/10.1098/rspb.2007.0368>
- Hermans, E. J., Bos, P. A., Ossewaarde, L., Ramsey, N. F., Fernández, G., & van Honk, J. (2010a). Effects of exogenous testosterone on the ventral striatal BOLD response during reward anticipation in healthy women. *NeuroImage, 52*(1), 277–283. <https://doi.org/10.1016/j.neuroimage.2010.04.019>
- Hermans, E. J., Bos, P. A., Ossewaarde, L., Ramsey, N. F., Fernández, G., & van Honk, J. (2010b). Effects of exogenous testosterone on the ventral striatal BOLD response during reward anticipation in healthy women. *NeuroImage, 52*(1), 277–283. <https://doi.org/10.1016/j.neuroimage.2010.04.019>
- Hermans, E. J., Ramsey, N. F., & van Honk, J. (2008). Exogenous Testosterone Enhances Responsiveness to Social Threat in the Neural Circuitry of Social Aggression in Humans. *Biological Psychiatry, 63*(3), 263–270. <https://doi.org/10.1016/j.biopsych.2007.05.013>
- Hertwig, R., & Engel, C. (2016). Homo Ignorans: Deliberately Choosing Not to Know. *Perspectives on Psychological Science, 11*(3), 359–372. <https://doi.org/10.1177/1745691616635594>
- Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O’doherly, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience, 20*(8), 1142–1149. <https://doi.org/10.1038/nn.4602>



- Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Bateson, M., ... Wolfe, J. W. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. <https://doi.org/10.1016/j.tics.2014.10.004>
- Huber, R. E., Klucharev, V., & Rieskamp, J. (2013). Neural correlates of informational cascades: Brain mechanisms of social influence on belief updating. *Social Cognitive and Affective Neuroscience*, *10*(4), 589–597. <https://doi.org/10.1093/scan/nsu090>
- Hume, D. (2003). *A treatise of human nature*. Courier Corporation.
- Isidori, A. M., Giannetta, E., Gianfrilli, D., Greco, E. A., Bonifacio, V., Aversa, A., ... Lenzi, A. (2005). Effects of testosterone on sexual function in men: Results of a meta-analysis. *Clinical Endocrinology*, *63*(4), 381–394. <https://doi.org/10.1111/j.1365-2265.2005.02350.x>
- Jackson, M. O., Malladi, S., & McAdams, D. (2022). Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(34). <https://doi.org/10.1073/pnas.2205549119>
- Jacobs, B., Driscoll, L., & Schall, M. (1997). Life-span dendritic and spine changes in areas 10 and 18 of human cortex: A quantitative golgi study. *Journal of Comparative Neurology*, *386*(4), 661–680. [https://doi.org/10.1002/\(SICI\)1096-9861\(19971006\)386:4<661::AID-CNE11>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1096-9861(19971006)386:4<661::AID-CNE11>3.0.CO;2-N)
- Jacobs, B., Schall, M., Prather, M., Kapler, E., Driscoll, L., Baca, S., ... Treml, M. (2001). Regional dendritic and spine variation in human cerebral cortex: A quantitative golgi study. *Cerebral Cortex*, *11*(6), 558–571. <https://doi.org/10.1093/cercor/11.6.558>
- Jamieson, K. H., & Cappella, J. J. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.
- Jayles, B., Kim, H. rin, Escobedo, R., Cezera, S., Blanchet, A., Kameda, T., ... Theraulaz, G. (2017). How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(47), 12620–12625. <https://doi.org/10.1073/pnas.1703695114>
- Jezzini, A., Bromberg-Martin, E. S., Trambaiolli, L. R., Haber, S. N., & Monosov, I. E. (2021). A prefrontal network integrates preferences for advance information about uncertain

- rewards and punishments. *Neuron*, 109(14), 2339-2352.e5. <https://doi.org/10.1016/j.neuron.2021.05.013>
- Johnson, L. R., & Wood, R. I. (2001). Oral testosterone self-administration in male hamsters. *Neuroendocrinology*, 73(4), 285–292. <https://doi.org/10.1159/000054645>
- Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *Npj Science of Learning*, 2(1), 1–8. <https://doi.org/10.1038/s41539-017-0009-2>
- Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12), 1625–1633. <https://doi.org/10.1038/nn2007>
- Kahan, D. M. (2018). *Misconceptions , Misinformation , and the Logic of Identity-protective Cognition I . Introduction : Whence misconceptions and misinformation ?* (164).
- Kalla, J. L., & Broockman, D. E. (2018). The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *American Political Science Review*, 112(1), 148–166. <https://doi.org/10.1017/S0003055417000363>
- Kang, P., Lee, J., Sul, S., & Kim, H. (2013). Dorsomedial prefrontal cortex activity predicts the accuracy in estimating others' preferences. *Frontiers in Human Neuroscience*, 7(NOV), 1–11. <https://doi.org/10.3389/fnhum.2013.00686>
- Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25), 11163–11170. <https://doi.org/10.1073/pnas.1005062107>
- Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, 23(1), 130–137. <https://doi.org/10.1038/s41593-019-0549-2>
- Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, 77(2), 603–606. <https://doi.org/10.3982/ecta7833>
- Kelly, C. A., & Sharot, T. (2021). Individual differences in information-seeking. *Nature Communications*, 12(1), 1–13. <https://doi.org/10.1038/s41467-021-27046-5>
- Kelly, F., Bronstein, M., Cerf, V., Edwards, L., McAuley, D., Neff, G., ... Terras, M. (2022).

- The online information environment: Understanding how the internet shapes people's engagement with scientific information.* London.
- Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2020). Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research, 47*(1), 104–124. <https://doi.org/10.1177/0093650217719596>
- Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of Increasing Monetary Reward Selectively Recruits Nucleus Accumbens Brain. *21*, 1–5. <https://doi.org/20015472> [pii]
- Knutson, B., & Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience, 27*(31), 8174–8177. <https://doi.org/10.1523/JNEUROSCI.1564-07.2007>
- Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007a). Neural Predictors of Purchases. *Neuron, 53*(1), 147–156. <https://doi.org/10.1016/j.neuron.2006.11.010>
- Knutson, B., Rick, S., Wimmer, G. E., Prelec, D., & Loewenstein, G. (2007b). Neural Predictors of Purchases. *Neuron, 53*(1), 147–156. <https://doi.org/10.1016/j.neuron.2006.11.010>
- Kobayashi, K., & Hsu, M. (2019). Common neural code for reward and information value. *Proceedings of the National Academy of Sciences of the United States of America, 116*(26), 13061–13066. <https://doi.org/10.1073/pnas.1820145116>
- Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature, 399*(6732), 148–151. <https://doi.org/10.1038/20178>
- Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11*(6), 229–235. <https://doi.org/10.1016/j.tics.2007.04.005>
- Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology, 24*, 41–47. <https://doi.org/10.1016/j.copsyc.2018.04.009>
- Konrad, K. A., Lohse, T., & Qari, S. (2014). Deception choice and self-selection - The importance of being earnest. *Journal of Economic Behavior and Organization, 107*(PA), 25–39. <https://doi.org/10.1016/j.jebo.2014.07.012>

- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. <https://doi.org/10.1523/JNEUROSCI.3016-12.2012>
- Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. <https://doi.org/10.1177/1529100620946707>
- Kringelbach, M. L., & Radcliffe, J. (2005). *The human orbitofrontal cortex: linking reward to hedonic experience*. *6*(September), 691–702. <https://doi.org/10.1038/nrn1748>
- Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, *72*(5), 341–372. <https://doi.org/10.1016/j.pneurobio.2004.03.006>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.
- Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167. <https://doi.org/10.1038/nn.4064>
- Lee, D., & Seo, H. (2016). Neural Basis of Strategic Decision Making. *Trends in Neurosciences*, *39*(1), 40–48. <https://doi.org/10.1016/j.tins.2015.11.002>
- Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1–9. <https://doi.org/10.1038/s41562-017-0067>
- Leong, Y. C., Hughes, B. L., Wang, Y., & Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. *Nature Human Behaviour*, *3*(9), 962–973. <https://doi.org/10.1038/s41562-019-0637-z>
- Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. <https://doi.org/10.1016/j.conb.2012.06.001>
- Li, Y., Li, Y., Sescousse, G., Sescousse, G., Amiez, C., Dreher, J. C., & Dreher, J. C. (2015). Local Morphology Predicts Functional Organization of Experienced Value Signals in the

- Human Orbitofrontal Cortex. *Journal of Neuroscience*, 35(4), 1648–1658. <https://doi.org/10.1523/JNEUROSCI.3058-14.2015>
- Lindeman, M., Svedholm-Häkkinen, A. M., & Lipsanen, J. (2015). Ontological confusions but not mentalizing abilities predict religious belief, paranormal belief, and belief in supernatural purpose. *Cognition*, 134, 63–76. <https://doi.org/10.1016/j.cognition.2014.09.008>
- Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences*, 44(7), 1585–1595. <https://doi.org/10.1016/j.paid.2008.01.014>
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1), 75–98. <https://doi.org/10.1037/0033-2909.116.1.75>
- Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37, 205–220. <https://doi.org/10.1146/annurev-neuro-071013-014017>
- Maier, S. R. (2005). Accuracy matters: A cross-market assessment of newspaper error and credibility. *Journalism and Mass Communication Quarterly*, 82(3), 533–551. <https://doi.org/10.1177/107769900508200304>
- Mansouri, F. A., Freedman, D. J., & Buckley, M. J. (2020). Emergence of abstract rules in the primate brain. *Nature Reviews Neuroscience*, 21(11), 595–610. <https://doi.org/10.1038/s41583-020-0364-5>
- Marsh, E. J., Cantor, A. D., & M. Brashier, N. (2016). Believing that Humans Swallow Spiders in Their Sleep: False Beliefs as Side Effects of the Processes that Support Accurate Knowledge. *Psychology of Learning and Motivation - Advances in Research and Theory*, 64, 93–132. <https://doi.org/10.1016/bs.plm.2015.09.003>
- Martin-Soelch, C., Missimer, J., Leenders, K. L., & Schultz, W. (2003). Neural activity related to the processing of increasing monetary reward in smokers and nonsmokers. *European Journal of Neuroscience*, 18(3), 680–688. <https://doi.org/10.1046/j.1460-9568.2003.02791.x>
- Mazaika, P. K., Hoefft, F., Glover, G. H., & Reiss, A. L. (2009). Methods and Software for fMRI Analysis of Clinical Subjects. *NeuroImage*, 47, S58. [https://doi.org/10.1016/s1053-8119\(09\)70238-1](https://doi.org/10.1016/s1053-8119(09)70238-1)
- McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural

- systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503–507. <https://doi.org/10.1126/science.1100907>
- McNair, D. M. (1971). Profile of mood states instrument. In *Manual for the profile of mood states*.
- Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667. <https://doi.org/10.1007/s00429-010-0262-0>
- Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*, *21*(2), 103–122. <https://doi.org/10.1037/gpr0000111>
- Metereau, E., & Dreher, J.-C. (2013). Cerebral Correlates of Salient Prediction Error for Different Rewards and Punishments. *Cerebral Cortex*, (February 2013:23), 477–487.
- Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. <https://doi.org/10.1016/j.neuron.2015.09.039>
- Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *65*, 276–291. <https://doi.org/10.1016/j.neubiorev.2016.03.020>
- Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*(2), 265–284. [https://doi.org/10.1016/S0896-6273\(02\)00974-1](https://doi.org/10.1016/S0896-6273(02)00974-1)
- Moore, D. A., & Healy, P. J. (2008). The Trouble With Overconfidence. *Psychological Review*, *115*(2), 502–517. <https://doi.org/10.1037/0033-295X.115.2.502>
- Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, *6*(8), 771–781. <https://doi.org/10.2139/ssrn.1804189>
- Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in Cognitive Sciences*, *11*(11), 489–497. <https://doi.org/10.1016/j.tics.2007.08.013>
- Murray, R. J., Schaer, M., & Debbané, M. (2012). Degrees of separation: A quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation



- between self- and other-reflection. *Neuroscience and Biobehavioral Reviews*, 36(3), 1043–1059. <https://doi.org/10.1016/j.neubiorev.2011.12.013>
- Nagypál, A., & Wood, R. I. (2007). Region-specific mechanisms for testosterone-induced Fos in hamster brain. *Brain Research*, 1141(1), 197–204. <https://doi.org/10.1016/j.brainres.2007.01.022>
- Naqvi, N. H., & Bechara, A. (2009). The hidden island of addiction: the insula. *Trends in Neurosciences*, 32(1), 56–67. <https://doi.org/10.1016/j.tins.2008.09.009>
- Nelson, R. J., & Trainor, B. C. (2007). Neural mechanisms of aggression. *Nature Reviews Neuroscience*, 8(7), 536–546. <https://doi.org/10.1038/nrn2174>
- Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2021). Reuters Institute Digital News Report 2021: 10th Edition. *Reuters Institute for the Study of Journalism, University of Oxford*, 164.
- Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron*, 75(6), 1114–1121. <https://doi.org/10.1016/j.neuron.2012.07.023>
- Nieschlag, E., & Behre, H. M. (n.d.). Pharmacology and clinical uses of testosterone. In *Testosterone*. Springer, Berlin, Heidelberg.
- O’Doherty, J. P., Deichmann, R., Critchley, H. D., & Dolan, R. J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron*, 33(5), 815–826. [https://doi.org/10.1016/S0896-6273\(02\)00603-7](https://doi.org/10.1016/S0896-6273(02)00603-7)
- O’Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., McGlone, F., Kobal, G., ... Ahne, G. (2000). Sensory-specific satiety-related olfactory activation of the human orbitofrontal cortex. *NeuroReport*, 11(4), 893–897.
- Oei, N. Y. L., Veer, I. M., Wolf, O. T., Spinhoven, P., Rombouts, S. A. R. B., & Elzinga, B. M. (2012). Stress shifts brain activation towards ventral “affective” areas during emotional distraction. *Social Cognitive and Affective Neuroscience*, 7(4), 403–412. <https://doi.org/10.1093/scan/nsr024>
- Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yücel, M., & Lorenzetti, V. (2018). The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human Brain Mapping*, 39(8), 3398–3418.



<https://doi.org/10.1002/hbm.24184>

- Öngür, D., & Price, J. L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral Cortex*, *10*(3), 206–219. <https://doi.org/10.1093/cercor/10.3.206>
- Op De MacKs, Z. A., Moor, B. G., Overgaauw, S., Gürolu, B., Dahl, R. E., & Crone, E. A. (2011). Testosterone levels correspond with increased ventral striatum activation in response to monetary rewards in adolescents. *Developmental Cognitive Neuroscience*, *1*(4), 506–516. <https://doi.org/10.1016/j.dcn.2011.06.003>
- Ossewaarde, L., Qin, S., Van Marle, H. J. F., van Wingen, G. A., Fernández, G., & Hermans, E. J. (2011). Stress-induced reduction in reward-related prefrontal cortex function. *NeuroImage*, *55*(1), 345–352. <https://doi.org/10.1016/j.neuroimage.2010.11.068>
- Packard, M. G., Schroeder, J. P., & Alexander, G. M. (1998). Expression of testosterone conditioned place preference is blocked by peripheral or intraaccumbens injection of  $\alpha$ -flupenthixol. *Hormones and Behavior*, *34*(1), 39–47. <https://doi.org/10.1006/hbeh.1998.1461>
- Park, J., Kim, H., Sohn, J. W., Choi, J. R., & Kim, S. P. (2018). EEG beta oscillations in the temporoparietal area related to the accuracy in estimating others' preference. *Frontiers in Human Neuroscience*, *12*(February), 1–11. <https://doi.org/10.3389/fnhum.2018.00043>
- Park, S. A., Goïame, S., O'Connor, D. A., & Dreher, J. C. (2017). Integration of individual and social information for decision-making in groups of different sizes. *PLoS Biology*, *15*(6), 1–28. <https://doi.org/10.1371/journal.pbio.2001958>
- Park, S. A., Sestito, M., Boorman, E. D., & Dreher, J. C. (2019). Neural computations underlying strategic social decision-making in groups. *Nature Communications*, *10*(1), 1–12. <https://doi.org/10.1038/s41467-019-12937-5>
- Patton, J., Stanford, M., & Barratt, E. (1995). Patton et al., (1995) Factor structure of the barratt impulsiveness scale.pdf. *Journal of Clinical Psychology*, Vol. 51, pp. 768–774.
- Pennycook, G., Allan, J., Nathaniel, C., Derek, B., & Fugelsang, K. J. A. (2015). *On the reception and detection of pseudo-profound bullshit Gordon*. *10*(6), 549–563.
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of

- headlines without warnings. *Management Science*, 66(11), 4944–4957. <https://doi.org/10.1287/mnsc.2019.3478>
- Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, 7(1), 1–13. <https://doi.org/10.1525/collabra.25293>
- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making*, 15(4), 476–498.
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590–595. <https://doi.org/10.1038/s41586-021-03344-2>
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- Pennycook, G., & Rand, D. G. (2019a). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188(September 2017), 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188(September 2017), 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2), 185–200. <https://doi.org/10.1111/jopy.12476>
- Pennycook, G., & Rand, D. G. (2021a). Nudging social media sharing towards accuracy. *The Annals of the American Academy of Political and Social Science*, 1–23.
- Pennycook, G., & Rand, D. G. (2021b). The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., ... del Millán, J. R.

- (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8382–8390. <https://doi.org/10.1073/pnas.1918335117>
- Persoskie, A., Ferrer, R. A., & Klein, W. M. P. (2014). Association of cancer worry and perceived risk with doctor avoidance: an analysis of information avoidance in a nationally representative US sample. *Journal of Behavioral Medicine*, *37*(5), 977–987. <https://doi.org/10.1007/s10865-013-9537-2>
- Plassmann, H., O’Doherty, J., Shiv, B., & Rangel, A. (2008). Marketing actions can modulate neural representations of experienced pleasantness. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(3), 1050–1054. <https://doi.org/10.1073/pnas.0706929105>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. <https://doi.org/10.1037/a0019737>
- Plummer, M. (2003). *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling JAGS: Just Another Gibbs Sampler*. (Dsc).
- Poon, J., Niehaus, C. E., Thompson, J. C., & M., C. T. (2019). Adolescents’ pubertal development: Links between testosterone, estradiol, and neural reward processing Jennifer. *Horm Behav.*, *176*(5), 139–148. <https://doi.org/10.1016/j.yhbeh.2019.02.015>.Adolescents
- Popper, K. (1959). *The logic of scientific discovery* (H. & Co, Ed.).
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745–2752. <https://doi.org/10.1523/JNEUROSCI.4286-07.2008>
- Rademacher, L., Krach, S., Kohls, G., Irmak, A., Gründer, G., & Spreckelmeyer, K. N. (2010). Dissociation of neural networks for anticipation and consumption of monetary and social rewards. *NeuroImage*, *49*(4), 3276–3285.

- <https://doi.org/10.1016/j.neuroimage.2009.10.089>
- Radke, S., Volman, I., Mehta, P., Van Son, V., Enter, D., Sanfey, A., ... Roelofs, K. (2015). Testosterone biases the amygdala toward social threat approach. *Science Advances*, 1(5). <https://doi.org/10.1126/sciadv.1400074>
- Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, 5(3), 184–194. <https://doi.org/10.1038/nrn1343>
- Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7), 545–556. <https://doi.org/10.1038/nrn2357>
- Rapp, D. N., & Salovich, N. A. (2018). Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from the Behavioral and Brain Sciences*, 5(2), 232–239. <https://doi.org/10.1177/2372732218785193>
- Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, 9(5), 347–356. <https://doi.org/10.1111/1467-9280.00067>
- Redlawsk, D. P., Civettini, A. J. W., & Emmerson, K. M. (2010). The Affective Tipping Point: Do Motivated Reasoners Ever “Get It”? *Political Psychology*, 31(4), 563–593. <https://doi.org/10.1111/j.1467-9221.2010.00772.x>
- Redouté, J., Stoléru, S., Grégoire, M. C., Costes, N., Cinotti, L., Lavenne, F., ... Pujol, J. F. (2000). Brain processing of visual sexual stimuli in human males. *Human Brain Mapping*, 11(3), 162–177.
- Ren, Z. (Bella), Dimant, E., & Schweitzer, M. E. (2021). Social Motives for Sharing Conspiracy Theories. *SSRN Electronic Journal*, (January 2021). <https://doi.org/10.2139/ssrn.3919364>
- Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, 57, 199–226. <https://doi.org/10.1146/annurev.psych.57.102904.190208>
- Roiser, J. P., Stephan, K. E., den Ouden, H. E. M., Friston, K. J., & Joyce, E. M. (2010). Adaptive and aberrant reward prediction signals in the human brain. *NeuroImage*, 50(2), 657–664. <https://doi.org/10.1016/j.neuroimage.2009.11.075>
- Röttger, Paul; Vedres, B. (2020). *The Information Environment and its Effects on Individuals*

- and Groups: An Interdisciplinary Literature Review*. (April).
- Rouault, M., & Fleming, S. M. (2020). Formation of global self-beliefs in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27268–27276. <https://doi.org/10.1073/pnas.2003094117>
- Rusch, T., Steixner-Kumar, S., Doshi, P., Spezio, M., & Gläscher, J. (2020). Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, *146*, 1–44. <https://doi.org/10.1016/j.neuropsychologia.2020.107488>
- Russell, B. (1912). *The problems of philosophy*. New York: Henry Holt and Company.
- Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *Journal of Neuroscience*, *30*(40), 13525–13536. <https://doi.org/10.1523/JNEUROSCI.1747-10.2010>
- Sapienza, P., Zingales, L., & Maestripieri, D. (2009). Gender differences in financial risk aversion and career choices are affected by testosterone. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(36), 15268–15273. <https://doi.org/10.1073/pnas.0907352106>
- Scholz, C., Baek, E. C., Brook O'Donnell, M., & Falk, E. B. (2020). Decision-making about broad- and narrowcasting: a neuroscientific perspective. *Media Psychology*, *23*(1), 131–155. <https://doi.org/10.1080/15213269.2019.1572522>
- Scholz, C., & Falk, E. B. (2020). The Neuroscience of Information Sharing. *The Oxford Handbook of Networked Communication*, 284–307. <https://doi.org/10.1093/oxfordhb/9780190460518.013.34>
- Scholz, C., Jovanova, M., Baek, E. C., & Falk, E. B. (2020). Media content sharing as a value-based decision. *Current Opinion in Psychology*, *31*, 83–88. <https://doi.org/10.1016/j.copsyc.2019.08.004>
- Schultz, W. (2016). Dopamine reward prediction- error signalling: a two-component response. *Nature Reviews*.
- Schulz, L., Fleming, S. M., & Dayan, P. (2021). *Metacognitive computations for information search: confidence in control*.
- Schürmeyer, T., & Nieschlag, E. (1984). Comparative pharmacokinetics of testosterone

- enanthate and testosterone cyclohexanecarboxylate as assessed by serum and salivary testosterone levels in normal men. *International Journal of Andrology*, 7(3), 181–187. <https://doi.org/10.1111/j.1365-2605.1984.tb00775.x>
- Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., ... Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9), 2349–2356. <https://doi.org/10.1523/JNEUROSCI.5587-06.2007>
- Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., & Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology*, 114(3), 224–241. [https://doi.org/10.1002/1096-8644\(200103\)114:3<224::AID-AJPA1022>3.0.CO;2-I](https://doi.org/10.1002/1096-8644(200103)114:3<224::AID-AJPA1022>3.0.CO;2-I)
- Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, 346(6207), 340–343. <https://doi.org/10.1126/science.1256254>
- Serra-Garcia, M., & Gneezy, U. (2021). Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies. *American Economic Review*, 111(10), 3160–3183. <https://doi.org/10.1257/AER.20191295>
- Sescousse, G., Barbalat, G., Domenech, P., & Dreher, J. C. (2013). Imbalance in the sensitivity to different types of rewards in pathological gambling. *Brain*, 136(8), 2527–2538. <https://doi.org/10.1093/brain/awt126>
- Sescousse, G., Caldú, X., Segura, B., & Dreher, J. (2013). *Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies*.
- Sescousse, G., Li, Y., & Dreher, J. C. (2013). A common currency for the computation of motivational values in the human striatum. *Social Cognitive and Affective Neuroscience*, 10(4), 467–473. <https://doi.org/10.1093/scan/nsu074>
- Sescousse, G., Redouté, J., & Dreher, J. C. (2010). The architecture of reward value coding in the human orbitofrontal cortex. *Journal of Neuroscience*, 30(39), 13095–13104. <https://doi.org/10.1523/JNEUROSCI.3501-10.2010>
- Shalvi, S., Soraperra, I., van der Weele, J. J., & Villeval, M. C. (2019). *Shooting the*



*Messenger ? Supply and Demand in Markets for Willful Ignorance \**.

- Shamay-Tsoory, S. G., Saporta, N., Marton-Alper, I. Z., & Gvirts, H. Z. (2019). Herding Brains: A Core Neural Mechanism for Social Alignment. *Trends in Cognitive Sciences*, 23(3), 174–186. <https://doi.org/10.1016/j.tics.2019.01.002>
- Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, 20(1), 25–33. <https://doi.org/10.1016/j.tics.2015.11.002>
- Sharot, T., & Sunstein, C. R. (2020). How people decide what they want to know. *Nature Human Behaviour*, 4(1), 14–19. <https://doi.org/10.1038/s41562-019-0793-1>
- Shepperd, J. A., Waters, E. A., Weinstein, N. D., & Klein, W. M. P. (2015). A Primer on Unrealistic Optimism. *Current Directions in Psychological Science*, 24(3), 232–237. <https://doi.org/10.1177/0963721414568341>
- Simerly, R. B., Swanson, L. W., Chang, C., & Muramatsu, M. (1990). Distribution of androgen and estrogen receptor mRNA-containing cells in the rat brain: An in situ hybridization study. *Journal of Comparative Neurology*, 294(1), 76–95. <https://doi.org/10.1002/cne.902940107>
- Simos, P. G., Breier, J. I., Fletcher, J. M., Foorman, B. R., Castillo, E. M., & Papanicolaou, A. C. (2002). Brain mechanisms for reading words and pseudowords: An integrated approach. *Cerebral Cortex*, 12(3), 297–305. <https://doi.org/10.1093/cercor/12.3.297>
- Sinclair, D., Purves-Tyson, T. D., Allen, K. M., & Weickert, C. S. (2014). Impacts of stress and sex hormones on dopamine neurotransmission in the adolescent brain. *Psychopharmacology*, 231(8), 1581–1599. <https://doi.org/10.1007/s00213-013-3415-z>
- Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, 13(8), 334–340. <https://doi.org/10.1016/j.tics.2009.05.001>
- Sirlin, N., Epstein, Z., Arechar, A. A., & Rand, D. G. (2021). Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review*, 2(6), 1–13. <https://doi.org/10.37016/mr-2020-83>
- Small, D. M., Zatorre, R. J., Dagher, A., Evans, A. C., & Jones-Gotman, M. (2001). Changes in brain activity related to eating chocolate: From pleasure to aversion. *Brain*, 124(9), 1720–1733. <https://doi.org/10.1093/brain/124.9.1720>



- Smith, A. R., Steinberg, L., & Chein, J. (2014). The role of the anterior insula in adolescent decision making. *Developmental Neuroscience, 36*(3–4), 196–209. <https://doi.org/10.1159/000358918>
- Smith, D. V., Hayden, B. Y., Truong, T. K., Song, A. W., Platt, M. L., & Huettel, S. A. (2010). Distinct value signals in anterior and posterior ventromedial prefrontal cortex. *Journal of Neuroscience, 30*(7), 2490–2495. <https://doi.org/10.1523/JNEUROSCI.3319-09.2010>
- Snyder, P. J., & Lawrence, D. A. (1980). Treatment of male hypogonadism with testosterone enanthate. *Journal of Clinical Endocrinology and Metabolism, 51*(6), 6–1339. <https://doi.org/10.1210/jcem-51-6-1335>
- Spielberg, J. M., Forbes, E. E., Ladouceur, C. D., Worthman, C. M., Olino, T. M., Ryan, N. D., & Dahl, R. E. (2013). Pubertal testosterone influences threat-related amygdala-orbitofrontal cortex coupling. *Social Cognitive and Affective Neuroscience, 10*(3), 408–415. <https://doi.org/10.1093/scan/nsu062>
- Stanton, S. J., Liening, S. H., & Schultheiss, O. C. (2011). Testosterone is positively associated with risk taking in the Iowa Gambling Task. *Hormones and Behavior, 59*(2), 252–256. <https://doi.org/10.1016/j.yhbeh.2010.12.003>
- Stanton, S. J., Mullette-Gillman, O. A., McLaurin, R. E., Kuhn, C. M., LaBar, K. S., Platt, M. L., & Huettel, S. A. (2011). Low- and high-testosterone individuals exhibit decreased aversion to economic risk. *Psychological Science, 22*(4), 447–453. <https://doi.org/10.1177/0956797611401752>
- Steer, R. A., & Beck, A. T. (1997). *Beck Anxiety Inventory*.
- Stenstrom, E., & Saad, G. (2011). Testosterone, Financial Risk-Taking, and Pathological Gambling. *Journal of Neuroscience, Psychology, and Economics, 4*(4), 254–266. <https://doi.org/10.1037/a0025963>
- Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences, 131*(3), 139–148. <https://doi.org/10.1007/s12064-011-0142-z>
- Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated reasoning and public opinion. *Journal of Health Politics, Policy and Law, 36*(6), 935–944. <https://doi.org/10.1215/03616878-1460524>

- Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, 6(5), 363–375. <https://doi.org/10.1038/nrn1666>
- Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How People Update Beliefs about Climate Change: Good News and Bad News Cass. *Cornell L. Rev.*, (102).
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (M. Press, Ed.).
- Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., ... Nakahara, H. (2012). Learning to Simulate Others' Decisions. *Neuron*, 74(6), 1125–1137. <https://doi.org/10.1016/j.neuron.2012.04.030>
- Taber, C. S., & Lodge, M. (2006). of Political Beliefs. *American Journal of Political Science*, 50(3), 755–769.
- Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153. <https://doi.org/10.1080/21670811.2017.1360143>
- Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Bayesian or biased? Analytic thinking and political belief updating. *Cognition*, 204(June), 104375. <https://doi.org/10.1016/j.cognition.2020.104375>
- Tappin, B. M., Pennycook, G., Rand, D. G., & Hanser, P. (2021). Rethinking Sophistication and Motivated Reasoning. *Forthcoming in the Journal of Experimental Psychology: General*, 1–60.
- Tavassoli, A., & Ringach, D. L. (2010). When your eyes see more than you do. *Current Biology*, 20(3), 93–94. <https://doi.org/10.1016/j.cub.2009.11.048>
- Teresa Arnedo, M., Salvador, A., Martínez-Sanchis, S., & Pellicer, O. (2002). Similar rewarding effects of testosterone in mice rated as short and long attack latency individuals. *Addiction Biology*, 7(4), 373–379. <https://doi.org/10.1080/1355621021000005955>
- Thaler, M. (2021). The Supply of Motivated Beliefs. *ArXiv Preprint*. Retrieved from <http://arxiv.org/abs/2111.06062>
- Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5), 1–38.

<https://doi.org/10.18637/jss.v059.i05>

- Toelch, U., & Dolan, R. J. (2015). Informational and Normative Influences in Conformity from a Neurocomputational Perspective. *Trends in Cognitive Sciences*, 19(10), 579–589. <https://doi.org/10.1016/j.tics.2015.07.007>
- Toledano, R., & Pfaus, J. (2006). The sexual arousal and desire inventory (SADI): A multidimensional scale to assess subjective sexual arousal and desire. *Journal of Sexual Medicine*, 3(5), 853–877. <https://doi.org/10.1111/j.1743-6109.2006.00293.x>
- Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *Annals of the American Academy of Political and Social Science*, 700(1), 136–151. <https://doi.org/10.1177/00027162221087936>
- Travison, T. G., Morley, J. E., Araujo, A. B., O'Donnell, A. B., & McKinlay, J. B. (2006). The relationship between libido and testosterone levels in aging men. *Journal of Clinical Endocrinology and Metabolism*, 91(7), 2509–2513. <https://doi.org/10.1210/jc.2005-2508>
- Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Annals of the International Communication Association*, 44(2), 157–173. <https://doi.org/10.1080/23808985.2020.1759443>
- Tucker, J., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., ... Nyhan, B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3144139>
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1), 273–289. <https://doi.org/10.1006/nimg.2001.0978>
- Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. *AAAI Fall Symposium - Technical Report, FS-12-06*, 37–42.
- Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning Memory and*

- Cognition*, 33(1), 219–230. <https://doi.org/10.1037/0278-7393.33.1.219>
- Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11), 913–916. <https://doi.org/10.1016/j.tics.2021.07.013>
- Van Den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S., & Hulshoff Pol, H. E. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Human Brain Mapping*, 30(10), 3127–3141. <https://doi.org/10.1002/hbm.20737>
- Van Honk, J., Schutter, D. J. L. G., Hermans, E. J., Putman, P., Tuiten, A., & Koppeschaar, H. (2004). Testosterone shifts the balance between sensitivity for punishment and reward in healthy young women. *Psychoneuroendocrinology*, 29(7), 937–943. <https://doi.org/10.1016/j.psyneuen.2003.08.007>
- van Marle, H. J. F., Hermans, E. J., Qin, S., & Fernández, G. (2010). Enhanced resting-state connectivity of amygdala in the immediate aftermath of acute psychological stress. *NeuroImage*, 53(1), 348–354. <https://doi.org/10.1016/j.neuroimage.2010.05.070>
- Van Wingen, G. A., Zylicz, S. A., Pieters, S., Mattern, C., Verkes, R. J., Buitelaar, J. K., & Fernández, G. (2009). Testosterone increases amygdala reactivity in middle-aged women to a young adulthood level. *Neuropsychopharmacology*, 34(3), 539–547. <https://doi.org/10.1038/npp.2008.2>
- van Wingen, G., Mattern, C., Verkes, R. J., Buitelaar, J., & Fernández, G. (2010). Testosterone reduces amygdala-orbitofrontal cortex coupling. *Psychoneuroendocrinology*, 35(1), 105–113. <https://doi.org/10.1016/j.psyneuen.2009.09.007>
- Vermeulen, A., Verdonck, L., & Kaufman, J. M. (1999). A critical evaluation of simple methods for the estimation of free testosterone in serum. *Journal of Clinical Endocrinology and Metabolism*, 84(10), 3666–3672. <https://doi.org/10.1210/jcem.84.10.6079>
- Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., & Buckner, R. L. (2008). Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *Journal of Neurophysiology*, 100(6), 3328–3342. <https://doi.org/10.1152/jn.90355.2008>
- Volman, I., Toni, I., Verhagen, L., & Roelofs, K. (2011). *Endogenous Testosterone Modulates*

- Prefrontal--Amygdala Connectivity during Social Emotional Behavior*. *Cerebral Cortex*.
- Washburn, A. N., & Skitka, L. J. (2018). Science Denial Across the Political Divide: Liberals and Conservatives Are Similarly Motivated to Deny Attitude-Inconsistent Science. *Social Psychological and Personality Science*, 9(8), 972–980. <https://doi.org/10.1177/1948550617731500>
- White, J. K., Bromberg-Martin, E. S., Heilbronner, S. R., Zhang, K., Pai, J., Haber, S. N., & Monosov, I. E. (2019). A neural network for information seeking. *Nature Communications*, 10(1), 1–19. <https://doi.org/10.1038/s41467-019-13135-z>
- Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, 2(3), 125–141. <https://doi.org/10.1089/brain.2012.0073>
- Wilson, R. P., Colizzi, M., Bossong, M. G., Allen, P., Kempton, M., Abe, N., ... Bhattacharyya, S. (2018). The Neural Substrate of Reward Anticipation in Health: A Meta-Analysis of fMRI Findings in the Monetary Incentive Delay Task. *Neuropsychology Review*, 28(4), 496–506. <https://doi.org/10.1007/s11065-018-9385-5>
- Wood, R. I. (2008). Anabolic-androgenic steroid dependence? Insights from animals and humans. *Frontiers in Neuroendocrinology*, 29(4), 490–506. <https://doi.org/10.1016/j.yfrne.2007.12.002>
- Wood, R. I., Johnson, L. R., Chu, L., Schad, C., & Self, D. W. (2004). Testosterone reinforcement: Intravenous and intracerebroventricular self-administration in male rats and hamsters. *Psychopharmacology*, 171(3), 298–305. <https://doi.org/10.1007/s00213-003-1587-7>
- Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 71, 101–111. <https://doi.org/10.1016/j.neubiorev.2016.08.038>
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>
- Zajkowski, W., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex

- in directed, but not random, exploration. *Proceedings of ICCM 2017 - 15th International Conference on Cognitive Modeling*, 79–84.
- Zald, D. H., McHugo, M., Ray, K. L., Glahn, D. C., Eickhoff, S. B., & Laird, A. R. (2014). Meta-Analytic Connectivity Modeling Reveals Differential Functional Connectivity of the Medial and Lateral Orbitofrontal Cortex. *Cerebral Cortex*, *24*(1), 232–248. <https://doi.org/10.1093/cercor/bhs308>
- Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1419–1424. <https://doi.org/10.1073/pnas.1116783109>
- Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *NeuroImage*, *34*(3), 1310–1316. <https://doi.org/10.1016/j.neuroimage.2006.08.047>
- Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370. <https://doi.org/10.1111/j.1600-0447.1983.tb09716.x>