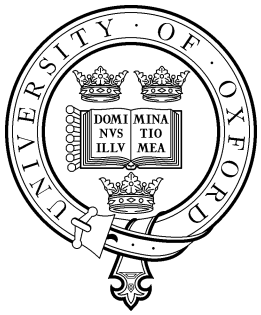
On the Use of Metacognitive Signals to Navigate the Social World



Niccolò Pescetelli
Christ Church College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Experimental Psychology

Trinity Term 2017

Abstract

On the use of Metacognitive Signals to Navigate the Social World

Niccolò Pescetelli

Christ Church College
Thesis submitted for the degree of Doctor of Philosophy in Experimental Psychology
University of Oxford
Trinity Term 2017

Since the early days of psychology, practitioners have recognised that metacognition - or the act of thinking about one's own thinking - is intertwined with our experience of the world. In the last decade, scientists have started to understand metacognitive signals, like judgments of confidence, as precise mathematical constructs. Confidence can be conceived of as an internal estimate of the probability of being correct. As such, confidence influences both advice seeking and advice taking while allowing people to optimally combine their views for joint action and group coordination.

This work begins by exploring the idea that confidence judgments are important for monitoring not only uncertainty associated with one's performance but also, thanks to their positive covariation with accuracy, the reliability of social advisers, particularly when objective criteria are not available. I present data showing that, when adviser and advisee's judgments are independent, people are able to detect subtle variations in advice information, irrespective of feedback presence. I also show that, when such independence is broken, the use of subjective confidence to track others' reliability leads to systematic deviations.

I then proceed to explore the differences existing between static and dynamic social information exchange. Traditionally, social and organisational psychology have investigated one-step unidirectional information systems, but many real-life interactions happen on a continuous time-scale, where social exchanges are recursive and dynamic. I present results indicating that the dynamics of social information exchange (recursive vs. one-step) affect individual opinions over and above the information that is communicated. Overall, my results suggest a bidirectional involvement of confidence in social inference and information exchange, and highlight the limits of the mechanisms underlying it.

Acknowledgements

First and foremost, I would like to express my infinite gratitude to my supervisor Nick Yeung. I cannot think of another person who influenced me more than he did, with his scrupulous methodology, the sharp thinking, patience, humour, and calming attitude, but most of all his warm and welcoming presence. Nick never misses an occasion to make you think, to teach you something new, to bring your best inner scientist out, in a vaguely spooky Socratic way. If Italian was the language of science, I could not come up with a better word to describe him than *maestro*.

Secondly I would like to express my gratitude to my other mentor, Bahador Bahrami. Undeniably, he was my first true supervisor. He was the first to make time for me and my ideas, the first to patiently deal with my little problems, my anxieties, my questions. And of course the first to show me that no matter how wacky your interests are you can do science with them. But also one apology is due: For not finding more time to stay in touch during this DPhil. I have no doubt that great work could have come out from our collaboration, but it's unbelievable how quick four years pass by.

And then to all my collaborators, my fantastic colleagues over the years. To Lucie and Annika, the first metacognition team. To Marike and Raluca, for sticking around. To Chia-Lun, Santiago and Hildward, for being companions in this adventure. To Dan, for showing me that the same question is better answered together. To Michael, Naomi, Joshua, Aaron, Jan and Fabrice, for the interesting conversations, the lunches and the many visits to the pub. To the Department, the Clarendon Fund and Christ

Church College, who all contributed to my research. And to our undergraduate and graduate students who, over the years, have helped us so much.

Finally this work, would have not been possible, without my friends and family. Thank you for the support, the stimulation and for creating the perfect physical and mental environment to nourish and sustain me over the years. I am grateful for having all of you. To my gorgeous sisters, I look forward to seeing how amazing you both become and the great things you'll surely achieve (if my advice does not spoil you). To nonna Adriana, for her incredible generosity and open heart. To my many friends from my earlier university years, now spread around the world. My incredible flatmate Manuel, and our "other" mate Joshua, you both made my time in Oxford truly fun and you are the reason my house felt like a home. And finally my muse Georgina, for being always there even though we were often miles apart, and of course, for proofreading this page.

This thesis is dedicated to my parents, Tiziana and Goffredo.

For supporting me in all my enterprises. Even the most crazy ones.

They are the best parents that a child, affected by curiosity and love for discovery, could wish for.

Can you step back from your own mind

and thus understand all things?

明白四達，能無知乎?

— Tao Te Ching

道德經

Contents

List of Figures

1

Introduction

'Don't you think if I were wrong I'd know it?'

– Sheldon Cooper

Chapter Abstract

The two overarching themes of this thesis are (1) describing the bidirectional relation between metacognition and the social world and (2) understanding social information sharing during decision making. In this Chapter I will provide an overview of the literature on metacognition, underlining the fact that metacognitive phenomena have been described in very distant realms of psychology, and I will present confidence judgments in perceptual decisions as a working case for the present work. I will stress the idea that confidence is better described as a dynamic information process, rather than a static read-out of evidence. I will then discuss the literature on group decision making and show how its modern development puts it in close proximity to research on metacognition. I next introduce the two lines of experiments presented in this thesis, the first one investigating how we learn about the reliability of social information and the second one investigating the differences existing between static and dynamic sharing of information. I will conclude by describing the scope of this work and outlining its development across the chapters.

Confidence: An Example of Metacognition

Preamble. A search on Google of the word "confidence" outputs a comprehensive list of websites that promise to boost your success in life. Pages like "How to Fake It When You're Not Feeling Confident", "Building Self-Confidence - Stress Management Skills from Mind Tools" and "Confidence Coaching to Build Self-Esteem & Self-Belief" inform you that "Confidence gives you the power to conquer the world"[1]. At first glance, this aspect of confidence seems to be detached from anything related to the cognitive sciences, where confidence has a completely different meaning, in terms of inverse uncertainty about internal state variables or subjectively estimated probability of a correct decision. However, such dissociation raises the question in the scholar's mind of why the popular interpretation of confidence has been linked with life success and well-being. The research presented in this thesis hopes to show the reader that confidence is indeed intertwined with the way we learn from our experiences in everyday life and how we act in our social world. And perhaps, the next time, the popular interpretation of confidence will be less surprising.

Work overview. In this thesis, I explore the role of confidence in social information sharing, in which I study how people share and use information from other agents in simple, carefully controlled perceptual decisions, and contrast these behaviours with the predictions of normative probabilistic frameworks of the decision process. In this Introduction, I first discuss empirical and theoretical work that characterises confidence as a subjective estimate of accuracy. I then review studies from organizational psychology looking at advice taking behaviour, which typically used general knowledge tasks. I then discuss studies of group decision making looking at how groups of individuals can operate as powerful information aggregation systems.

[1] quote from: `www.inc.com/peter-economy/5-powerful-ways-to-boost-your-confidence.html`

I then outline the work reported in this thesis, which synthesises these approaches in carefully-controlled perceptual decision making paradigms by: (1) manipulating advice characteristics (including the presence or absence of feedback) and measuring the impact of these on explicit and implicit measures of trust; (2) manipulating the dynamics of information sharing between individuals, thus isolating the effect of communication means from communication content.

Metacognition

When living our daily life, a sense of confidence often accompanies our decisions. When deciding whether to bring an umbrella to work, our decision is informed by how sure or unsure we are about the possibility of rain. This sense of certainty or uncertainty called *confidence* belongs to a class of cognitive phenomena that usually goes under the general term of *metacognition*. The first use of the term is attributed to John Flavell, who within the context of learning defines it as "knowledge and cognition about cognitive phenomena" (Flavell, 1979, p.906). Thus, any act of thinking (conscious or unconscious) that has as its object one's thinking itself can be ascribed to metacognition.

In cognitive science, confidence refers to one of the three main metrics characterising a decision, together with judgment accuracy and reaction times (Peirce & Jastrow, 1884). Although confidence has not traditionally received the same attention given to its two siblings, the past ten years have seen a steady resurgence of interest, and confidence studies have grown in number.

Several different flavours

Metacognitive constructs have been investigated in multiple fields within psychology, each using different methods and often different terminologies. Within psychology of learning, Flavell found that older students could accurately tell when they had learnt a list of names. On the contrary, younger students' perception of their learning was

uncorrelated with their objectively measured performance in a recall task. Several other disciplines have made extensive use of self-assessing judgments that can be interpreted as metacognitive processes. For example, studies on conscious perception have primarily used self reports to measure visual awareness (Persaud, McLeod, & Cowey, 2007; Ramsoy & Overgaard, 2004; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010) and the distinction between perceptual awareness and metacognition is still under debate (Charles, Van Opstal, Marti, & Dehaene, 2013; Dehaene, Charles, King, & Marti, 2014; Rausch & Zehetleitner, 2016; Seth, 2008). In social and organizational psychology, researchers have investigated the role of confidence in leadership and influence within a group. Decision makers were found to be more willing to accept advice from a confident source even if this did not scale with objective accuracy (Price & Stone, 2004; Sniezek & Van Swol, 2001). Zarnoth and Sniezek (1997) found that confidence drove influence and predicted accuracy only on tasks where a correct solution could be objectively demonstrated. Meanwhile, results from forensic psychology showed that confident testimony is usually more influential on juries' verdicts (Penrod & Cutler, 1995). Tenney, MacCoun, Spellman, and Hastie (2007) however showed that confident but uncalibrated advisers quickly lost the trust of the jury once their accuracy was proved to be unrelated to their confidence (see also Tenney, Spellman, & Maccoun, 2008).

In the 1990s, the debate generated from the discovery of cognitive biases touched on confidence judgments as well. Gigerenzer, Hoffrage, and Kleinbölting (1991) showed that judgments of confidence come about from a host of circumstantial information that is often not strictly task-relevant, but emerges from the interaction of proximal cues that are learnt to co-vary with the measure of interest (Gigerenzer, 2008). For example, if participants are asked to compare the population of two cities they tend to use auxiliary information like whether the city has a football team or not, and base their confidence on the perceived or experienced reliability of those

cues (Pescetelli, Rees, & Bahrami, 2016). Although such contextual cues offer a fast-and-frugal useful heuristic, they can also lead to overconfident mistakes.

More recently, research in perceptual decision-making has gained interest in metacognition (Bach & Dolan, 2012). Here, the last decade has produced important contributions for the development of a general theory of metacognition. Precise computational models (Pleskac & Busemeyer, 2010; Vickers, 1979) as well as theories about its neural implementation (Fleming & Dolan, 2012; Fleming, Huijgen, & Dolan, 2012; Kiani & Shadlen, 2009) have resulted in quantitative descriptions of the mechanisms of confidence formation, as described below.

Finally, metacognitive processes have been investigated also in non-human animals where a host of paradigms have been developed to allow the study of metacognitive representations using purely behavioural measures instead of verbal reports (Kepecs & Mainen, 2012; J. D. Smith, Couchman, & Beran, 2012). Common methods include (1) adding of an opt-out option to avoid a punishment or to get a smaller but sure reward (Kepecs & Mainen, 2012; Kiani & Shadlen, 2009); (2) using post-decision wagering techniques (Persaud et al., 2007); and (3) measuring the willingness to wait for a reward instead of initiating a new trial (Kepecs, Uchida, Zariwala, & Mainen, 2008). Evidence exists indicating that several species - including rats, rhesus monkeys and dolphins - might be able to engage in metacognitive processes, although the topic remains a matter of debate (Le Pelley, 2012).

Two main observations emerge from the extreme diversity of fields that have made use of metacognitive judgments. First, the sheer number of disciplines that touch upon this topic suggests that metacognitive processes might be at play in a wide range of tasks and species. Second, this very diversity of disciplines should also warn researchers to be aware of differences when trying to use a common language. Comparing results together is often made more difficult by the use of diverse tasks, measures and experimental conditions. A challenging task is to understand

whether this diversity is generated by similar cognitive functions and whether similar mechanisms can explain it.

Properties of Confidence

Notwithstanding the diversity of psychological research areas that have investigated confidence and metacognitive phenomena, common properties are observed. An important feature of confidence, often reported in several domains, is its co-variation with accuracy within each individual (Roediger III, Wixted, & Desoto, 2012), a relationship called "calibration". If you think about your own decisions, you will notice that you are more likely to be right when you are confident than when unsure. Mathematically, calibration measures how well confidence judgments linearly scale with actual performance (Fleming & Lau, 2014). We will see later how calibration makes confidence important in both learning and social interactions. At the same time, it makes confidence difficult to disentangle from measures of performance itself, because any effect of interest attributed to confidence could also be due to differences in accuracy. In §*How to Measure Confidence* I will describe methods used to deal with this issue.

Although confidence calibration is positive *within* participants, it also varies considerably *across* individuals (Ais, Zylberberg, Barttfeld, & Sigman, 2016; Kruger & Dunning, 1999; Song et al., 2011), such that some people systematically are over- or under-confident with respect to their real underlying accuracy. On top of that, the degree to which people are over- or under-confident depends on task difficulty, a phenomenon termed "the hard-easy effect" (Gigerenzer et al., 1991): people are typically found to be under-confident when the task is easy and over-confident when the task is difficult. Similar effects are common in different research areas (Koriat, 2012a; Pleskac & Busemeyer, 2010) raising the possibility of common mechanisms.

Common principles and mechanisms

A general theory of metacognitive judgments is still missing, but attempts in this direction have been made by several authors. For example, Koriat (2012a) suggested a self-consistency model to explain confidence and confidence calibration results in the general knowledge and memory domains as well as in perceptual comparison tasks. According to this proposal, confidence emerges from sampling different pieces of evidence from memory and evaluating the agreement between different samples. Another influential framework was proposed by Nelson and Narens (1990), who framed different results from metamemory research in terms of the interplay between control and monitoring processes. Applying a similar logic, Yeung and Summerfield (2012) suggested that the error-monitoring literature (Rabbitt, 1966) and decision-making confidence literature (Vickers, 1979) can both be interpreted as a continuous process of uncertainty monitoring acting also after a decision is made. In the confidence judgments literature, an integrative account of confidence judgments in perceptual tasks has been offered by Pleskac and Busemeyer (2010), who showed how appropriately modified accumulation of evidence models (e.g., drift diffusion models reviewed below) can explain different results in the confidence literature, including changes of mind, lower confidence in error than correct, post-decisional processes and reaction times. Finally, Fleming, Dolan, and Frith (2012) recently suggested a descriptive taxonomy of research on metacognition, where individual studies can be classified along three relevant dimensions, in terms of conscious access, type of behaviour (first vs. second order) and type of representation (object- vs. meta-level).

In the rest of the present work, I will refer to confidence as the estimated subjective probability of having selected the correct answer. I will use the term *belief* and *opinion* interchangeably and interpret them as a probability distribution over possible outcomes or responses. In binary choices (e.g. 2-alternatives forced-choice tasks), such distributions can be thought of as a discrete probability distribution over two

alternative outcomes with probability p and $1 - p$. This definition offers an easy and direct way to use subjective confidence reports in formal models of social interaction. I will not make any strong claim about how these estimates are generated internally, but I will provide below a synthesis of current models attempting to do so. My aims in doing so here is to clarify how confidence can be thought of as a mathematically tractable variable that carries information about the state of external variables.

Confidence in Decision Making

Models of Confidence

In this section I will describe the three main models of decision confidence that exist in the literature: Signal Detection theoretic models (SDT), Drift Diffusion models (DDM) and Bayesian models. They provide a useful reference to place the present work in context and interpret its results, through formalising the notion that confidence judgments are a subjective evaluation of the probability that a given choice is correct.

Signal Detection models. Signal Detection theory is a well developed framework to describe the process underlying correct responses and errors in traditional 2AFC tasks (Green & Swets, 1966; Macmillan & Creelman, 2005). It is a special case of Bayesian decision theory (Lau, 2008), where the information a decision is based upon is represented as two partially overlapping distributions over evidence x underpinning a decision (which is usually, but not always, a decision about a perceptual stimulus); the two distributions can represent signal and noise (detection tasks) or two different signals (discrimination tasks). A decision criterion c can be arbitrarily set to separate two portions over x. For example, in a detection task, evidence falling on the left of c would favour decision A "The evidence comes from the noise distribution", otherwise decision B "The evidence comes from the signal distribution". According to Bayesian decision theory (Macmillan & Creelman, 2005), a discriminability measure called d'

can be determined that is proportional to the strength of the signal (μ) but inversely so to its uncertainty (σ):

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}(\sigma_S^2 + \sigma_N^2)}} \tag{1.1}$$

where S and N represent the signal and the noise distributions respectively, in a detection task (Figure 1.1). An intuitive representation of confidence within this framework is the distance from the decision boundary c that the evidence falls on a particular trial. Thus, when the observed evidence falls far below the decision boundary, one can be confident that the signal is absent; correspondingly, when the observed evidence falls well above the boundary, one can be confident that the signal is present. For a binary confidence scale (e.g. Sure *vs.* Unsure), we can add two more arbitrary boundaries $\pm\theta_{\text{confidence}}$ on either side of c so that: if $|x| < |\theta_{confidence}|$, then express low confidence, otherwise express high confidence (Figure 1.1). This framework naturally creates a factorial combination of decision type (present/absent) and confidence (high /low). Conservativeness in reporting high confidence is in this model manipulated by increasing the distance between the decision boundary and the confidence boundary.

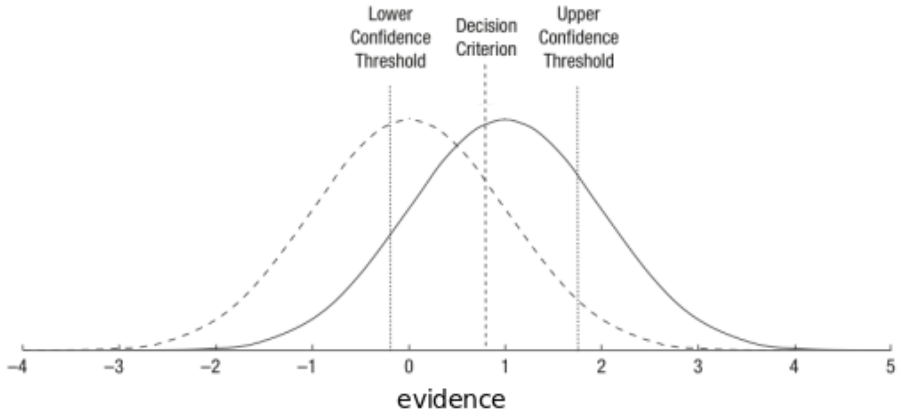


Figure 1.1: Signal detection model of confidence in a binary response task with binary confidence levels.

Multiple confidence levels can similarly be formalised in terms of setting additional confidence thresholds, such that thresholds that are further away from the criterion c

generate higher confidence levels. The model nicely accounts for the fact that errors are characterised by lower confidence levels - because it is less likely for an error to reach the confidence boundary - and that better discrimination generates higher confidence - because the two distributions are now shifted away from one another. Signal detection models of confidence have been used to characterise phenomena like blindsight (Ko & Lau, 2012). Notice that in this instance of the model, confidence is generated from the same signal generating the first order decision (Fleming, Dolan, & Frith, 2012). Galvin, Podd, Drga, and Whitmore (2003) expanded this to include meta-level representations by using distributions derived from first-order decisions instead of the first order signal and noise distributions. Notwithstanding their appeal, SDT models have the disadvantage of characterising the decision process as a static read-out of perceptual (object-level) or metacognitive (meta-level) signals (Yeung & Summerfield, 2012). Thus, dynamic components of a decision - e.g. reaction times, changes of mind, belief updating and uncertainty monitoring - cannot readily be incorporated.

Evidence accumulation models. A different approach is taken by evidence accumulation models. This class of models describe decisions as dynamic processes of evidence accumulation over time (see Pleskac & Busemeyer, 2010, for a detailed review). In a perceptual task, evidence accumulation can be thought of as sequential sampling from a sensory stimulus. For example, in a "balance of evidence" model (Vickers, 1979), evidence is accumulated separately for the two response options (e.g., the two distributions in Figure 1.1), thus effectively being equivalent to a continuous-time version of the SDT model. This accumulation process creates a random walk of two decision variables representing the evidence favouring each response. At every time point, each decision variable is updated with a new sample. The process continues until one of two variables hit the boundary θ, thus triggering the corresponding response (Figure 1.2, panel A). The strength of the evidence is represented by the rate

of accumulation, the drift rate δ, for each accumulator in the race. Greater drift rates generate shorter reaction times as the boundary is reached more rapidly. Confidence can be defined as the difference in the amount of accumulated evidence between the chosen and unchosen options ($\Delta\epsilon$) at the moment that the first decision variable hits the decision boundary.

A related and perhaps more influential model of decision making is the drift diffusion model (DDM), where only one decision variable is accumulated representing the difference between two alternative options (Figure 1.2, panel B). The advantage of this model is that it can be thought of as an implementation of the sequential probability ratio test (SPRT) and alternative models of 2AFC decisions can be reduced to it (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; P. L. Smith & Vickers, 1988). This model correctly accounts for properties of RTs such as distribution skewness and correct responses-errors asymmetry. However, confidence in the drift diffusion model is more difficult to define as the term $\Delta\epsilon$ is missing. Nevertheless, different cues have been proposed that could contribute to generate a confidence signal, including the time to hit boundary θ (or a mixture of evidence and decision time (Kiani, Corthell, & Shadlen, 2014)), the number of vacillations (Audley, 1960), and signal strength and variability (Boldt, de Gardelle, & Yeung, 2017). Evidence from neural recordings in monkeys and rodents (Kepecs et al., 2008; Kiani & Shadlen, 2009) suggested that confidence represents an interaction between drift rate δ and boundary θ, namely weighing the quality of the evidence by its quantity (Yeung & Summerfield, 2012). In these experiments, neurons in lateral intraparietal and orbito-frontal cortices exhibited firing rates consistent with an accumulation to bound process. Furthermore, firing rate was reduced in opt-out trials, supposedly when the animal recognised its own uncertainty.

Notwithstanding the success of these accounts, some authors (Pleskac & Busemeyer, 2010; Yeung & Summerfield, 2012) have argued that any model that describes confidence as generated at the moment of the decision ("decisional locus" models)

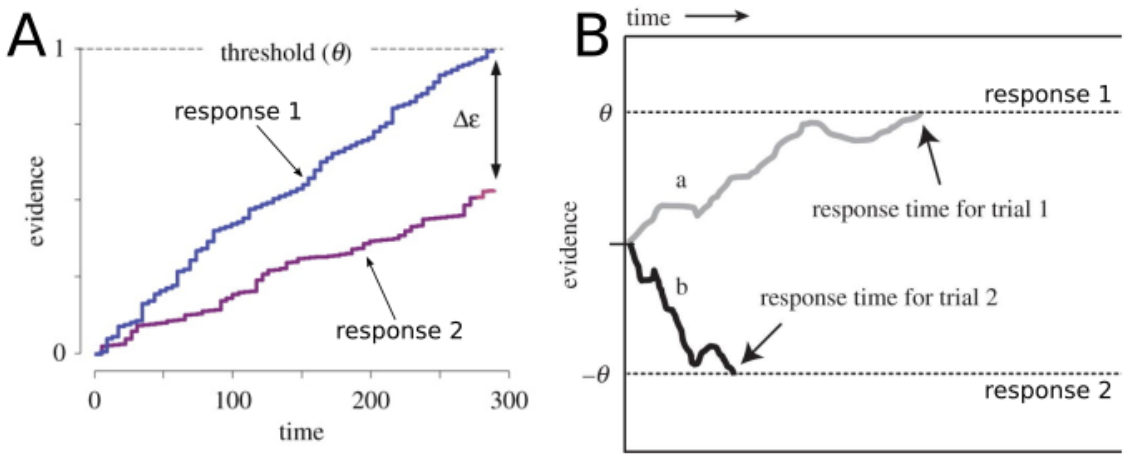


Figure 1.2: Two evidence accumulation models: A) Balance of evidence model. Figure adapted from (Kepecs & Mainen, 2012); B) Drift diffusion model of decision. Figure adapted from Yeung and Summerfield (2012).

cannot explain for the fact that confidence in correct answers is usually higher than for errors and that changes of mind often occur after a decision is initiated (particularly after incorrect decisions). To account for these phenomena, it has been suggested (Pleskac & Busemeyer, 2010) that the accumulation process must continue until the confidence judgment is given ("post-decisional locus" models). Indeed, recent evidence from behavioural and brain imaging studies has provided clear support for post-decisional view of evidence accumulation (Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010; Yu, Pleskac, & Zeigenfuse, 2015, Charles & Yeung in prep.), although the debate is still ongoing about what quantity is being accumulated (Fleming, 2016): Some studies advocate that the same decision variable is accumulated post-decisionally (van den Berg et al., 2016), others suggest that different signals with separate neural sources might be involved (Murphy, Robertson, Harty, & O'Connell, 2015).

Bayesian models. The idea that the brain computes Bayesian statistics to make inferences about the world has been around for some time (Knill & Pouget, 2004), fostered by the finding in multisensory perception and motor learning that humans

integrate different modalities in an optimal manner, namely weighting evidence by uncertainty (Ernst & Banks, 2002; Körding & Wolpert, 2004; Kvam & Pleskac, 2016). Although higher order decisions, like economic decisions, did not appear for a long time to benefit from the same mechanisms (Ariely, 2008; Kahneman & Tversky, 1979) recent work shows that these decisions too may be explained by optimal Bayesian inference (Oaksford & Chater, 1991; Tsetsos et al., 2016). Bayesian decision theory, for which SDT represents a specific instance, describes the rules by which prior beliefs, expressed probabilistically, should be updated as new evidence is collected. Formally:

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{1.2}$$

where θ represents a set of beliefs expressed as probability distributions over states or variables and $\mathbf{D}$ represents incoming new data. The variance of the distributions quantifies the uncertainty over a given state or variable value. Thus, confidence is inherently embedded in this formulation as the inverse variance or "precision" of the distribution (Meyniel, Sigman, & Mainen, 2015; Pouget, Drugowitsch, & Kepecs, 2016; Yeung & Summerfield, 2012). Recent efforts have tried to describe confidence within this framework. Aitchison, Bang, Bahrami, and Latham (2015) showed that when people are asked to rate their confidence on a previously reported binary response, a Bayesian mapping function from sensory data to confidence reports was better able to fit human behaviour compared to a heuristic approach where confidence was simply proportional to sensory evidence. Instead of simply scaling confidence c with sensory evidence $\mathbf{x}$, a Bayesian agent would compute the probability of being correct given the sensory evidence and its decision d as $P(correct|\mathbf{x}, d)$. The bold face represents the fact that the sensory evidence can be multidimensional[2]. When confidence reports and decisions were reported at the same time, rather than confidence judgments provided separately after their binary choice, some participants

[2]In fact, in this study using multidimensional stimuli was the only way to discriminate between the two models.

were better described by a heuristic approach. A similar idea was brought forward by Pouget et al. (2016) who recently suggested a distinction between uncertainty as noise over sensory states (named *certainty*) and uncertainty about the correctness of the decision response (named *confidence*).

A slightly different definition was given by Meyniel et al. (2015) who suggested that confidence reports can be interpreted as a summary statistic (e.g. *arithmetic mean*) of the distribution over sensory states $\mathbf{x}$ or, alternatively, as the precision (i.e. inverse variance) of the same distribution (cf. Yeung & Summerfield, 2012). The first suggestion is similar to SDT approaches while the second is inspired by biologically plausible implementations of Bayesian inference (Ma, Beck, Latham, & Pouget, 2006). In either case, confidence is a scalar readout of distribution $\mathbf{x}$ and is not necessarily dependent on decision d.

An important challenge is that internal representations (e.g. distribution, certainty, etc.) and external reports (e.g. readouts and confidence reports) are difficult to disentangle and are rarely studied within the same species. Moreover, although extremely promising, Bayesian approaches lack the temporal dimension that DDMs focus on (although see Kvam and Pleskac (2016) for an exception). Future work could usefully address how Bayesian models of confidence can represent the evolution of decision variables as a function of continuous changes in sensory evidence.

How to Measure Confidence

As described in previous paragraphs, one of the most common findings in confidence literature is its covariation with accuracy (i.e. confidence calibration). Any attempt to quantify the accuracy of confidence reports thus must take into account the fact that participants who perform well in the first-order task (i.e. classifying the sensory stimulus) will automatically be better at the second-order task (i.e. saying whether they are performing accurately) (Kruger & Dunning, 1999). A first attempt to measure confidence calibration used a variation of the SDT method of determining the

discriminability d' between two distributions (N vs. S or S_1 vs. S_2). Instead of using first-order hits and false alarm rates, type II d' uses second-order hits and false alarms: a hit is the classification of a correct trial by means of a high confidence judgment and a false alarm is reporting a high confidence in an incorrect trial (Galvin et al., 2003). Similarly, type-II Area Under the Receiver Operating Characteristic (A_{ROC}) curve can be calculated to quantify (second-order) performance in ways that are independent of setting specific criteria (Macmillan & Creelman, 2005). This method however makes first and second order decisions difficult to compare. Maniscalco and Lau (2012) introduced the measure of *meta-d'*, that addresses the issue by expressing metacognitive sensitivity (i.e. calibration) in relation to first order available information. Thus a meta-d' of one would mean that no information was lost between first-order and second-order decisions - that is, the calibration of confidence is as good as can be expected given the accuracy of the decision task (see Fleming and Lau (2014) for a review of the topic). Different suggestions have been proposed to understand the neural mechanisms underlying inter-individual differences in calibration (Fleming, Huijgen, & Dolan, 2012; Song et al., 2011), but for the scope of the present work it is sensible to assume that they stem from internal noise in the representation of sensory information.

Confidence in Social Information Sharing

The role of confidence in our lives has been extensively investigated in social, organizational and forensic psychology. Below is an overview of some key findings from these literatures. Collectively, these findings show that, to the extent that confidence is calibrated, it provides useful information about the value and reliability of shared information.

Judge-Adviser Systems.

Research in organizational psychology has systematically explored the conditions under which people ask for and make use of advice (Brehmer & Hagafors, 1986). The

typical paradigm used in this literature is the judge-adviser systems (JAS from here onwards) (see Bonaccio & Dalal, 2006, for an exhaustive review). Several versions exist but all share similar components. In its most general version, a participant ("judge") is asked to make an initial estimate about a given question, then receives advice from one or more real or virtual sources ("advisers"), and then is finally given the option of revising their initial judgment. Typical questions for the judge fall into two categories: judgment tasks and estimation tasks. The former are characterised by questions with categorical responses, such as "What is France's most populated city?". The latter are characterised by continuous (or at least ordinal) response variables, such as "What is the population of France's most populated city?". Estimation tasks can provide a more sensitive measure of advice-taking because subtle changes of mind can be registered that fall in between the judge's initial answer and the adviser's recommendation. However, one common issue faced by both types of task - and that the present work has addressed - is the difficulty in comparing together different questions (for example because some might be more difficult than others) and different judges (for example because of people's differing expertise in the topic). Thus advice-taking measures, like (Harvey & Fischer, 1997, 's), are often computed for each question separately and analyses are done across participants instead of within each judge. Bonaccio and Dalal (2006) notice that measures of advice taking used in many studies conflate changes of mind with changes of confidence (Bonaccio & Dalal, 2006, p. 141). For the present work, it is important to note that even if the original response is maintained, advice could impact the confidence in that response.

Egocentric Bias.

One of the most replicated results using JAS tasks is the presence of a robust advice-discounting bias, whereby judges weigh their own opinion more than their advisers', typically updating their initial estimate between 20% and 30% of the difference between their and their adviser's opinion. The bias is less evident if the judge is not

able to form an initial judgment prior receiving advice (Harvey & Harries, 2004). Some results suggest that the ego-centric bias might be mediated by confidence. For instance, Yaniv (2004b) showed that more knowledgeable individuals discounted advice more in a general knowledge task, and that the weight of advice decreased as a function of the distance from the judge's initial opinion. Similarly, power manipulations between judge and advisers, as well as observational studies, have shown that power is negatively associated with advice taking, such that more powerful individuals discount advice more. The effect is mediated by a higher confidence in the judge's own judgment, which has in turn a negative impact on the judge's final accuracy (See, Morrison, Rothman, & Soll, 2011; Tost, Gino, & Larrick, 2012). These results are interesting if we consider the fact that using advice is typically beneficial for the judge in JAS paradigms: it significantly improves accuracy (Yaniv, 2004a), it allows for sharing accountability (Harvey & Fischer, 1997), it makes estimates closer to normative predictions, and it forces judges to consider different problem strategies (Schotter, 2003).

Several explanations of the egocentric bias exist. A simple explanation is that judge's initial estimate (if available) provides a reference to which further information is anchored to (Tversky & Kahneman, 1974). Alternatively, Yaniv and Kleinberger (2000) have argued that judges have privileged access to their internal reasons supporting their own opinion, but not to their advisers', and so give more weight to the former. However, later results confirmed the existence of a genuine preference for the self: participants egocentrically up weigh opinions that have been (correctly or incorrectly) labelled as their own, but average opinions when the self is not involved (Harvey & Harries, 2004; Soll & Mannes, 2011). Two untested explanations would make the egocentric bias less "irrational": (1) asymmetric regret might exist between errors following self-initiated changes of mind and errors following advice-dependent changes of mind (Coricelli et al., 2005); (2) learning might be reduced after advice-dependent errors due to a reduced sense of agency (Khalighinejad & Haggard, 2016;

Metcalfe & Greene, 2007; Moore, Dickinson, & Fletcher, 2011; Murty, DuBrow, & Davachi, 2015).

Influence of Advice.

The results just described suggest that confidence is inversely proportional to advice taking. The other side of the same coin is that advice confidence is directly proportional to its influence. This phenomenon has been particularly investigated within the legal domain where eyewitness testimony is often the primary method to identify potential suspects. Confident witnesses are trusted more (Penrod & Cutler, 1995) and confident testimony is more influential on juries, while identifications made with low confidence often never make it to the courtroom (Roediger III et al., 2012). Due to the negative consequences of a poor correlation between confidence and accuracy, research in this field has thus put great emphasis on confidence calibration. Similar results have been found using a JAS paradigm in the knowledge domain, where Sniezek and Van Swol (2001) showed that confident advice had larger impact on judges' opinions than uncertain advice. People have been repeatedly shown to adopt a confidence heuristic, defined as a preference to trust and rely more on confident opinions, which are in turn perceived as more reliable (Price & Stone, 2004; Yates, Price, Lee, & Ramirez, 1996). Notice how a Bayesian framework can naturally account for these results as it prescribes that opinions should be aggregated by weighting individual judgments by associated inverse uncertainty or confidence (Bahrami et al., 2010), which naturally leads to less advice influence when the advisee is confident and more advice influence when the adviser is more confident.

Tenney et al. (2007) confirmed that confident testimony is more influential but expanded existing results by showing that trust in the reliability of the source is quickly lost if a confident witness is proved wrong - (see also Yaniv & Kleinberger, 2000, for asymmetric reputation revision) - or in other words if advice is uncalibrated (Fleming & Lau, 2014). However, evidence suggests that people value calibration only

if feedback is readily available. In a series of experiments, Sah, Moore, and Maccoun (2013) manipulated the availability of feedback in an advice-taking task and showed that, when feedback was removed or costly to acquire, people reverted to a confidence heuristic and disregarded advice calibration. Interestingly, the use of calibration to form estimates of advisers' reliability seems to develop later in life compared to more explicit cues of reliability such as confidence and accuracy (Tenney, Small, Kondrad, Jaswal, & Spellman, 2011).

Notably, the original results from Yates et al. (1996) suggest that over- or under-confidence does not hurt the reputation of the adviser as long as confidence judgments have a good *resolution* (Fleming & Lau, 2014), where resolution is defined as the discriminability of correct and error trials on the basis of confidence. For example, an adviser who is 90% correct and who expresses a confidence of 60% whenever wrong and 70% whenever correct, shows good resolution but poor calibration. The results of Yates et al. (1996) seem thus to suggest that some form of normalisation of advice received is performed by the advisee and that people show sensitivity to the informativeness of the advice rather than calibration *per se*.

Interaction.

Advice taking seems to be also affected by how judges and advisers share information. JAS studies have been performed using very diverse procedures, some involving unstructured face-to-face interaction, some involving semi-structured interaction (e.g., through writing or computer interface), others involving no interaction at all (Bonaccio & Dalal, 2006; Sniezek & Van Swol, 2001; Swol & Sniezek, 2005). Systematic manipulations of the interactive medium, however, have rarely been carried out. One study, Van Swol (2011) (Experiment 1), compared situations where judge and adviser communicated face-to-face versus in writing, but found no effect of communication medium, both in terms of trust and advice influence.

Although JAS tasks have consistently shown the presence of an ego-centric bias, research on group decision making and small groups has shown that interaction among peers often leads to polarisation and opinion drifts - also known as "risky shift" (Kerr & Tindale, 2004; Moscovici & Zavalloni, 1969; Myers & Lamm, 1976). Using a sports prediction task, Heath and Gonzalez (1995) showed that interaction did not affect accuracy in task performance or the accuracy of metacognitive judgments, but simply made decision makers overall more confident. The authors suggested that this effect was due to the fact that when interacting, people implicitly create rationales for their opinions thus making them more confident (the "rationale construction" hypothesis). The result is in agreement with an argumentative view of interaction suggesting that better arguments are elicited when people have the possibility to discuss (Mercier & Sperber, 2011; Trouche, Sander, & Mercier, 2014). The argumentative framework, however, cannot be reconciled with perceptual judgment tasks, for which arguments are impossible to define.

What is Confidence *for*?

The studies reviewed above clearly show the importance of confidence (real or reported) on advice taking and opinion formation in social contexts. Nonetheless, they remain agnostic about why its role is so important. One recent hypothesis, developed within the cognitive sciences, suggests an intriguing view on the topic. Shea et al. (2014) discuss a symmetry existing between theory of mind and metacognition: While metacognition concerns representations an agent has about its own mental content, theory of mind concerns representations about other agents' mental content. The authors highlight that automatic metacognitive representations (Bach & Dolan, 2012; Pouget et al., 2016) are necessary for a number of tasks, including cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001), uncertainty monitoring (Yeung & Summerfield, 2012), optimal integration of different sensory modalities (Morgan, DeAngelis, & Angelaki, 2008) and optimal action selection (Körding &

Wolpert, 2004). At the individual level, optimal action policies can quickly be learnt as an agent interacts with its environment by minimising prediction errors (Sutton & Barto, 1998). But humans have been exposed to great evolutionary pressure to live in societies (Dunbar, 2003; Dunbar & Shultz, 2007) where coordinated action is often more important than individual action. Here, prediction errors take the form of a discrepancy between sensory data and expectations on others' actions and can be minimised by inferring the *intentions* of social others (Friston & Frith, 2015). A huge literature body on theory of mind has described the mechanisms (and failures) of inferring intentions from other's actions (C. D. Frith, 1999, 2007; U. Frith & Frith, 2003; Kilner, Friston, & Frith, 2007). Shea et al. (2014) suggest that coordinated actions among individuals can occur by sharing explicit metacognitive representations (for example, in the form of language) across members of a group. Thus, optimal action selection and cognitive control at the group level are made possible by creating a shared metacognitive space, that grounds within the cognitive sciences earlier work in social psychology on shared task representations and mental models (Kerr & Tindale, 2004).

The decision making literature and the advice taking literature described so far are characterised by a focus on individual cognitive processes. Information is perceived by a receiver and a response is produced. But how do people make decisions *together*? In the next section I will be describing the existing literature on group decision making. I will present the main findings, the traditional interpretation of those and then more recent developments.

Group decisions

Crowds: the Traditional View(s)

A philosophical tradition that goes back to Plato's Republic, passing through Locke, Hobbes and Rousseau, has tried to characterise the optimal way to regulate a group

of people, being this a city or a country. The first empirical investigation of the topic was done by the Marquis de Condorcet (1785), who in his *Essai* describes a probabilistic analysis of integrating a "plurality of views" from members of a jury. He shows that, if we assume that people's views have a positive correlation with the truth, then convergence to that truth is expected in the limit of increasing group size. The first empirical proof of the concept was given by Francis Galton (1907) who (reluctantly) demonstrated that the median value of a distribution of estimates coming from independent naïve observers was remarkably close to the true value. However, many collective failures produced by crowds were also well known at that time (Mackay, 1841). The fact that group decisions can turn disastrously wrong is sadly well-known. Economic bubbles, the Challenger's launch and the Bay of Pigs invasion are commonly reported as examples of group choices gone wrong. Modern social psychology has tried to define the conditions leading to crowds' "madness" or "wisdom".

Madness

Postwar Years. Continuing a tradition started with Mackay (1841) and Le Bon (1895), and fostered by the events of the first half of the 20th century, many social psychologists showed how social groups impair the rational abilities and empathy of the individual, leading to conformism and groupthink (Asch, 1956; Haney, Banks, & Zimbardo, 1973; Latane, Williams, & Harkins, 1979). Janis (1972) hypothesised that under some conditions - like strong leadership, cohesion and encapsulation from external information or experts' opinion - groups can show extreme consensus-seeking leading to extreme mistakes. Although the model has been extremely influential (Turner & Pratkanis, 1998), researchers have shown that results are more ambiguous than originally thought and similar conditions can produce also collective gains (Kerr & Tindale, 2004).

The Information Shift. In the 1990s, groups research saw a shift towards how information is distributed and shared among individuals, thus beginning to conceive groups as information processing units (Hinsz, Tindale, & Vollrath, 1997). A paradigmatic study is represented by Stasser and Titus (2003)'s hidden profile experiment, in which individual members of a group are given different pieces of information that, if shared, allow the group to come to the correct decision. Importantly, some pieces of information are given to several members while others are only given to no more than one individual. A robust effect is that group members are more willing to discuss shared information than unshared, even when doing so leads to errors. A model of the effect suggests that groups sample information from group members so that shared evidence is more likely to be discussed early on and to be perceived as reliable because multiple (supposedly independent) sources support it (Kerr & Tindale, 2004).

Wisdom

Stationary Aggregation Methods. Research on groups has also highlighted that groups can sometimes outperform individuals, an idea recently popularised with the term *Wisdom of Crowds* (Surowiecki, 2004). As in Galton (1907)'s original experiment, groups of individuals have been repeatedly shown to be extremely good at estimation tasks and forecasting. Several methods developed to harvest this power rely upon taking the (absolute or weighted) mean or the median of individual estimates and using other simple combination algorithms (Mannes, Soll, & Larrick, 2014; Soll & Larrick, 2009). One of the most common aggregation strategies is the majority rule (i.e., aggregation of opinions by counting individual votes), which has been shown to be robust compared to other integration strategies across a variety of situations (Hastie & Kameda, 2005).

Sorkin, Hays, and West (2001) applied a signal detection analysis to provide a normative description of information integration among different individuals performing a difficult visual detection task. The model defines the weights that should be given to

each opinion given individual sensitivity *d'* so as to maximise group-*d'*. The authors show that a majority rule outperforms other aggregation rules like supermajority, in which an option is taken if more than half individuals composing the group support it and a default option is taken otherwise, and a unanimity rule, that an option is taken if all members of the group support it and a default option taken otherwise. Empirical groups were also tested and free communication among members was allowed. The authors found that group-*d'* was greater than individual-*d'*, although lower than the optimal benchmark.

Interestingly, however, Lorenz, Rauhut, Schweitzer, and Helbing (2011) showed evidence that the traditional wisdom of crowds effect is progressively reduced as a function of repeated exposure to other group members' judgments, suggesting that breaking of independence between individual estimates was damaging. Notably, and importantly for the present work, others' opinions were presented without any face-to-face interaction nor they were accompanied by confidence judgments, raising the question of whether collective losses could have been mitigated by these factors (Gürçay, Mellers, & Baron, 2015).

Dynamic Aggregation Methods. A common view held in economics considers prices as aggregate measures that dynamically reflect the distributed information about goods and economic players that is currently available (Hayek, 1945). Developing this idea further, people have started to explicitly trade information about future events using prediction markets. Prediction markets are markets where people can buy "stocks" of future events according to what they think is the most likely outcome. Interestingly, stock prices - which are determined by the popularity of the stock as in traditional markets - have been shown to reflect and track over time the probability associated with possible outcomes (Dreber et al., 2015; Wolfers & Zitzewitz, 2004), with accuracy as high as experts' forecasts and other forecasting methods (Graefe & Armstrong, 2011).

Contrary to static aggregation methods, prediction markets show that information can also be adaptively represented by group dynamics and members' interaction. Dynamic information aggregation strategies have been investigated also in non-human animals, most commonly in social insects, birds and fish (Bonabeau, Dorigo, & Theraulaz, 1999; Couzin, 2009). This line of research has shown that many emergent properties of these swarming systems can be explained by the aggregate of simple individual rules that govern interaction at a local scale (Couzin, 2009). Recent web implementations of swarm dynamics have been applied also to humans to facilitate forecasting (Rosenberg, 2015).

Two common explanations. Groups seem to be somehow extremely efficient in spreading information (and misinformation) across a network of agents (Mason & Watts, 2012; Moussaïd, Brighton, & Gaissmaier, 2015). The efficiency of information aggregation systems - like prediction markets, averaging and majority rule - has traditionally been explained using two common mechanisms. The noise cancelling hypothesis (NCH) assumes that different individuals' beliefs can be described by a shared signal component and unshared uncorrelated noise. Increased aggregate accuracy is observed because uncorrelated noise across individuals cancels out (Surowiecki, 2004). The bracketing hypothesis (BH) suggests that when people interact with social partners, their final estimate improves as long as their initial estimates "bracket" the true value on a continuous variable scale (Larrick & Soll, 2006). These proposed mechanisms nicely explain why - for example - the diversity of the group (Hong & Page, 2004) or the diversity of the models used (N. Silver, 2012), and the independence of group members (Lorenz et al., 2011), predict group performance. But a different story, fostered by metacognitive research in the cognitive sciences, has recently emerged.

Crowds: An Alternative Explanation

Results by Sorkin et al. (2001), already described, showed that most groups did not achieve the nominal collective accuracy that was predicted by their model. The authors attribute this collective loss to a social loafing effect where individuals exert less effort if in a group than alone (Latane et al., 1979). However, Bahrami et al. (2010) showed that the effect could be explained by a confidence sharing model where a two-member group ("dyad") is characterised as a Bayesian unit that aims to integrate information optimally across the two members, i.e. by scaling evidence (choice) by uncertainty (inverse confidence). The confidence sharing model correctly predicts greater information loss and thus smaller collective benefit as the difference in sensitivity *d'* between the two members becomes progressively larger. Further language analyses performed on speech recorded during the task confirmed that participants tended to discuss the uncertainty associated with their subjective experiences (Fusaroli et al., 2012). Face-to-face interaction seems to be the best way participants are able to optimally share their personal information. Sharing choices only, i.e. strengths but not uncertainties, is not enough to attain collective benefit. Sharing confidence levels, even if only through a computer indicator, is enough for the dyad to surpass its best member (Bahrami et al., 2012b; Migdal, Raczaszek-Leonardi, Denkiewicz, & Plewczynski, 2012); a Maximum Confidence Slating (MCL) aggregation strategy, namely selecting the most confident opinion available, is often sufficient to reach greater-than-individual-accuracy (Bang et al., 2014; Koriat, 2012b).

Thus, in simple binary perceptual tasks such as the one used in the present work's experiments, confidence sharing seems to be a necessary and sufficient condition to attain collective benefit. Similar results were replicated in higher order cognitive tasks such as enumeration (Bahrami, Didino, Frith, Butterworth, & Rees, 2013), object-recognition (Brennan & Enns, 2015) and knowledge-based questions (Koriat, 2012b).

Mechanisms of Social Interaction Under Confidence Sharing

As we saw in the previous section on confidence, covariation between confidence and accuracy within each individual is a very robust finding. Thanks to this property, individuals are able to determine on a given trial who among them is more likely to be correct by simply comparing their confidence levels (Pescetelli et al., 2016). This strategy seems to be an efficient heuristic to adopt in many situations (Koriat, 2012b). However, characteristics of the confidence distribution - e.g. the average confidence - are highly variable across individuals (Ais et al., 2016), and dependent on contingent factors such as gender (Barber & Odean, 2001), profession (Broihanne, Merli, & Roger, 2014), mental-health (Huq, Garety, & Hemsley, 1988), personality (Campbell, Goodie, & Foster, 2004) and culture (Mann, 1998). Thus, for the confidence heuristic - i.e., using confidence as an indicator of reliability (Price & Stone, 2004) - to be a successful communication strategy, people need to reduce confidence differences arising from contingent factors while maintaining variance due to task-relevant external signal (Bang, Aitchison, et al., 2017).

Another and related potential issue arising from the adoption of a confidence heuristic is that confidence calibration also greatly varies across individuals (Song et al., 2011). We should then expect that whenever people with poor confidence calibration interact they should incur in a collective loss. This is exactly what was recently found (Pescetelli et al., 2016). Although the rational strategy for the dyad in these situations is to adopt the judgment of the most accurate participant irrespective of confidence, empirical dyads seem to adopt an equality bias that over-weights the less accurate member and under-weights the most accurate one even when feedback is given on every trial (Mahmoodi et al., 2015). Face-to-face real-time social interaction seems to reduce the negative consequences of the equality bias if members are not allowed to form a prior opinion (Hertz, Romand-Monnier, Kyriakopoulou, & Bahrami, 2016), although the mechanisms underlying the effect remain unclear.

Open Questions

The literature reviewed so far shows that different fields of investigation have highlighted different aspects of confidence and confidence sharing. First, studies from meta-memory and decision making have shown that confidence can be conceived as an internal estimate of probability correct. This literature has also defined formal models that mathematically describe how judgments of confidence are formed. Second, the organisational psychology literature investigating advice taking behaviour has shown that confidence predicts lower influence if advice is received (i.e., confident people seek for and listen to advice less than uncertain ones) and greater influence if advice is given (i.e., confident people are more influential and perceived as more reliable). Third, the group decision making literature has shown that sharing confidence can lead to optimal decisions in social contexts, by weighting judgment estimates by their associated uncertainty. Effective use of social information thus depends on knowing the relative reliability of different opinions. The literature reviewed so far prompted me to investigate the two questions that motivate the present work: (a) How do we learn about the reliability of social information, particularly when objective feedback is absent; (b) How does the opportunity for dynamic interaction affect the reliability of social information sharing. These two lines of investigation are described below in two separate sections.

First line: Learning about others in the absence of feedback

Confidence in Learning

Reinforcement learning is a well-developed class of models characterising decision-making (Sutton & Barto, 1998). In a classical paradigm, subjects make serial decisions while experiencing a sequence of rewards or punishments. Learning occurs by matching internal expectations with external reward or punishment contingencies.

The difference between the two generates a prediction error signal that proportionally guides the update of expectations. Importantly, this framework provides a way to characterise both Pavlovian associative learning and instrumental conditioning by defining complementary roles within the ventral (Schultz, Apicella, Scarnati, & Ljungberg, 1992) and the dorsal striatum (O'Doherty et al., 2004). According to the standard interpretation, the former system (the "critic") predicts future outcomes in relation to states of the environment, whereas the latter (the "actor") learns action policies for particular states of the environment based on experienced outcomes (O'Doherty et al., 2004), with arbitration between the two defined by their associated uncertainty (Daw, Niv, & Dayan, 2005). Importantly for this work, it is common to discriminate between uncertainty arising from noise in the outcomes and noise in the rule (Bach & Dolan, 2012). In other words, a discrepancy between predicted and actual outcome could be observed even when the subject has learnt the rules in its environment and has responded accordingly. For example, this is the case with probabilistic learning. Imagine that by trial t you have learnt that an urn gives you a reward on 80% of trials. On trial $t + 1$ you will still be marginally surprised to see either outcome. This residual uncertainty has been defined "expected uncertainty" and it does not produce a change in the behavioural policy. If, however, the same urn starts giving you punishments one after the other, you will accumulate enough prediction errors that might tell you that the urn content has changed. This has been named "unexpected uncertainty" and it guides learning the rules governing what action policy to adopt (Bach & Dolan, 2012). Thus, it is important to bear in mind that several levels of uncertainty governing learning and behaviour exist and can be represented as a distribution probability over states (e.g. outcomes or rules) instead of as scalar values. For instance, if a reward always follows a correct response and punishment always follows an incorrect response, then feedback (i.e. outcome) is totally informative on the state of the world and can be represented as a scalar. For example, getting an electric shock from a loose wire always informs you that the

wire was electrified. Probabilistic feedback on the contrary is characterised by less informativeness, as the exact state of the world cannot be inferred from one single trial. Think for example of getting stuck in traffic on a new road to work. This could either mean that the new road suffers more traffic jams so it should not be taken any more, or alternatively could mean that the road is actually a short cut that only today happens to be overcrowded. The uncertainty about possible states of the world characterising probabilistic feedback makes it impossible for the subject to learn which state they are in from a single observation. Instead, learning must occur by accumulating outcomes over time (Schiffer, Siletti, Waszak, & Yeung, 2016).

Confidence as Internal Probabilistic Feedback

Nevertheless, learning seems to happen even when no external feedback of *any* sort is available. Indeed, most daily decisions that people face are made in situations where no objective reward or punishment is available or, if it is, its outcome is too delayed in time to be effectively used in a classical RL fashion. Take as an example a junior doctor making a medical diagnosis. The doctor will know about whether the diagnosis was correct (i.e. the outcome for the patient) months or years after it was made. In the meantime however, the same doctor is continuing diagnosing patients and improving her diagnostic skills. In these contexts, decision makers ought to rely on other types of cues, which however must correlate with accuracy for learning to be successful.

In the context of social information sharing, this issue can be reformulated into the problem of how we learn about the reliability of social information sources, like advisers or other group members, when feedback about our and their actions cannot be used to form an objective estimate of reliability. Weiss and Shanteau (2003) suggested that when no feedback is available, expertise should be defined on the basis of the ability of a judge to discriminate stimuli belonging to different categories and the ability of being consistent in treating stimuli belonging to the same class. The

authors' work correctly characterises expertise among different agents, in categorisation tasks when feedback is absent. However, it does not tell us about how agents in similar feedback-free scenarios might *acquire* such expertise. In other words, the approach taken by Weiss and Shanteau (2003) does not offer alternative feedback-like signals that can be reconciled with traditional reinforcement learning literature.

In an attempt to overcome these limitations, the central hypothesis of the first line of investigation in the present work suggests that variance in internally generated confidence signals carries useful probabilistic information about the state of the world (e.g., a correct or incorrect response), which can be exploited to guide learning through similar mechanisms as the ones described in the literature for external probabilistic feedback (Sutton & Barto, 1998). In a Bayesian framework, confidence represents the subjectively estimated probability of a given outcome (e.g., a correct response or a given reward). Once estimated, this quantity could in turn be used to infer the probability that the advice received is correct or wrong, thus allowing the accumulation of a probabilistic learning signal about adviser reliability over time. For instance, imagine our junior doctor thinks, after observing certain symptoms, that the probability of a patient having Alice in Wonderland Syndrome (Todd, 1955) is 98%. Imagine now that, in the absence of any other objective reference, she observes that another junior doctor also thinks that the patient has AiWS. Our junior doctor might confidently arrive to the conclusion that her peer's judgment is likely correct and indeed evidence of the peer's good diagnostic skills.

The use of such agreement-in-confidence heuristic - not to be counfounded with a confidence heuristic whereby the *adviser*'s confidence is used as an indirect signal for their accuracy (Price & Stone, 2004) - can be thought of as a cognitive short-cut to the otherwise intractable problem of learning without feedback. Work on heuristics and biases (Tversky & Kahneman, 1974) has shown that, when facing computationally intractable problems, the brain often uses fast-and-frugal heuristics that simplify the problem space, often reaching near-optimal results (Gigerenzer & Brighton, 2009;

Oaksford & Chater, 1991). These cognitive short-cuts work most of the time but lead, under specific circumstances, to characteristic deviations from rationality. Thus, a central prediction is that using an agreement-in-confidence heuristic will enable participants to learn the reliability of the advice received but, under special experimental manipulations, systematic biases will be observed.

This idea was recently supported by brain-imaging data showing that mesolimbic brain areas - typically associated with reward prediction error signals (Schultz et al., 1992) - encode both anticipation and prediction error of confidence in the absence of an objective external feedback, such that actions associated with higher confidence produced event-related BOLD time courses with a positive deflection from baseline compared to trials characterised by low confidence (Guggenmos, Wilbertz, Hebart, & Sterzer, 2016). The authors showed that a modified reinforcement learning algorithm using confidence instead of actual rewards can fit the behavioural and brain imaging data. These results suggest that learning may occur even in the absence of external feedback, simply by building an association between external states (e.g. stimuli) and internal states (e.g. metacognitive judgments).

The first line of experiments will be described in Chapters 2-4. Chapter 2 introduces the paradigm, which is a modified JAS task in which participants perform sequential perceptual decisions and receive advice from independent advisers over repeated interactions. Trust and advice influence are inferred for advisers with different objective reliability and accuracy, which show that participants are able to discriminate subtle advice cues even in the absence of feedback. Chapter 3 introduces a simple formal model of how learning in the absence of feedback can take place, and shows that the model reproduces qualitative patterns of the behavioural data. Chapter 4 presents two experiments where the independence of participants' judgments from the advice received is manipulated, and shows that both model and participants' pattern of trust in the advisers are consistent with an agreement-in-confidence strategy to estimate advice reliability. An adapted reinforcement learning model is also described

in Appendix A to show that our conclusions hold also when applying a similar logic to qualitatively different models.

Second line: Effects of dynamic interaction on social information sharing

An Experimental Dilemma: the Control-Validity Trade-off

One of the most difficult and consequential decisions that the scientist interested in social phenomena faces when starting a new study is what degree of control to exert on the social system under investigation. The dilemma lies in the fact that although the scientist's instinct would be to try to attain perfect control over all the variables of interest, the social system under study becomes less and less realistic the more control is exerted. High degrees of control over social variables often means creating computerised social characters, deceiving participants into believing that a real person is participating in an adjacent room, limiting the interaction to sharing tokenised pieces of information and/or controlling the direction of the information flow (Bonaccio & Dalal, 2006; Edelson, Sharot, Dolan, & Dudai, 2011; Soll & Mannes, 2011). On the other hand preserving ecological validity often involves allowing free unstructured interaction that in the best scenario gets videotaped and independently analysed, but with the associated burdens and challenges of quantification (Bahrami et al., 2010; Fusaroli et al., 2012; Haney et al., 1973).

Studies in the field have fallen on either side of the divide, but criticisms can be made of both. On one hand, social interaction does not happen in a vacuum and very rarely a distinction can be drawn between forming an opinion, collecting information from others, and integrating them with one's own to form an updated view. Interaction is essentially a dynamic system composed of two people who act and react on each others' signals, so studying it using staged paradigms and/or one-participant tasks risks an experimental setting that is sterile at best and inaccurate at

worst. On the other hand, allowing interaction to happen freely without any control limits the scope for careful manipulations and causal inference, and creates a host of potential confounds that cannot easily be accounted for.

The Need for a Middle Ground

Authors have recently started raising this concerns and asking how we can study social phenomena without being limited by our methodology. Schilbach et al. (2013) suggested the introduction of a second-person neuroscience, as opposed to first-person neuroscience, characterised by imagining one self in the other person's shoes (Rizzolatti & Sinigaglia, 2010), and third-person (Theory-of-Mind-based) neuroscience, characterised by inferring intentions from *observing* others (U. Frith & Frith, 2003; Gallagher & Frith, 2003). Both previous approaches limited social investigation to the use of non-interactive stimuli. Second person approaches emphasise the active engaged role of a social agent in contrast to a passive spectator. Studies on joint action have tried to answer similar questions for some time. Analysis of the dependencies in behaviour and neural correlates within and between musicians playing together suggested these settings to be a good environment for studying real-time social interactions (D'Ausilio, Novembre, Fadiga, & Keller, 2015). Use of multi-brain imaging techniques or hyperscanning has also been a useful resource for the investigation of the neural correlates of joint actions (Dumas, Nadel, Soussignan, Martinerie, & Garnero, 2010). The development of more ecologically valid paradigms does not necessarily need to go astray from high degrees of control. For instance a recent fully computerised paradigm (Auvray, Lenay, & Stewart, 2009; Froese, Iizuka, & Ikegami, 2014) successfully showed that real-time recursive interaction using haptic avatars was necessary for detecting the presence of others.

A recent interesting theoretical advancement has been to recreate similar interactive scenarios by pairing a human participant with a dynamic model of human

behaviour. Dumas, de Guzman, Tognoli, and Kelso (2014) suggested a Human Dynamic Clamp (HDC) paradigm inspired by the dynamic clamp method used in biology. Here, a human participant interacts with a model of human behaviour in a real-time closed loop system. The authors showed that oscillating dynamics of synchronous and asynchronous hand movements emerged when asking humans to imitate the model and the model to do the opposite. The HDC paradigm promises to be a powerful tool for model validation of real-time interactions.

Finally, in psychiatry the use of virtual reality environment has also been suggested to allow ecological validity while at the same time exerting high degrees of control (Mattout, 2012; Schilbach et al., 2006). This leaves us with an unsolved question: what does really make a social stimulus *social*?

What is ultimately "social"?

An intuitive idea of what a "social stimulus" is is possessed by everyone. But to proceed forward in the investigation of social constructs it is important to make a formal definition explicit. By looking at existing social experiments we can start to distil the features needed for a phenomenon to be social. Most studies in cognitive social neuroscience have investigated social cognition using human faces as stimuli to prompt social constructs (Edelson et al., 2011). There seems to be some agreement on the fact that images of conspecifics and faces in particular are processed differently in the brain than other stimuli (Farah, Wilson, Drain, & Tanaka, 1998; Simion, Macchi-Cassia, Turati, & Valenza, 2001). Whether presenting a face is a necessary requirement for a stimulus to be defined "social" remains however unclear. Indeed simply telling participants they are interacting with social advisers but not showing any face activates adjacent but separate brain areas compared to informationally identical non-social stimuli (Behrens, Hunt, Woolrich, & Rushworth, 2008). One of the most famous experiments in social psychology demonstrated that people can draw a great deal of information, infer intentions and create entire narratives by

simply looking at animated cartoons of geometric shapes (Heider & Simmel, 1944). It thus seems that stimuli do not need necessarily have to possess human body features to be perceived as social.

One could argue that the distinctive feature of a social stimulus is the knowledge that we are interacting with social others (Behrens et al., 2008). However, also this interpretation is problematic. More than half century ago Alan Turing (1950) suggested that the definition of agency and consciousness in social others should entirely be based on one criterion: behaviour. Any agent, either human or machine, should be regarded as conscious if it acts like one. Thus, defining an agent "social" only based on the participant's belief of who he or she is interacting with and disregarding its behaviour might not well be a viable alternative. On the other hand, defining it purely based on behaviour is also inappropriate. Studies have shown that people treat very differently humans and machines even if the observed behaviour is identical, usually in the direction of a greater punishment for machine errors (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Dietvorst, Simmons, & Massey, 2015).

The recent developments in second-person neuroscientific approaches and virtual reality experiments (Auvray et al., 2009; D'Ausilio et al., 2015; Mattout, 2012; Schilbach et al., 2013) seem to suggest that the nature of sociality is the interactive dynamic process occurring between agents. If we are going to accept this approach then social phenomena should be considered as complex systems that evolve over time where agents are engaged actors instead of simple spectators of social others (Schilbach, 2014). Agents act and react to each other without any clear definition of who is cause and who is effect. Due to their dynamic nature social interactions are expected to lead to non-linearities and super-additive effects. For example social interaction has been shown to lead to phenomena of escalation where agents reinforce each other's behaviour in a recursive positive feedback loop De Martino, O'Doherty, Ray, Bossaerts, and Camerer (2013); Mahmoodi, Bang, Ahmadabadi, and Bahrami (2013).

Comparing Interactive and Non-Interactive Social Exchange

To understand whether this last definition of social stimuli is an appropriate one, a direct comparison between interactive and non-interactive social exchange is needed. The second line of investigation in the present work aims to compare identical social situations where the information that is shared between the two agents (for example a judge and an adviser) is kept constant but the possibility for interaction is manipulated. The comparison can potentially shed light on the differences between classic JAS paradigms - typically characterised by information exchange that is one-directional, static and structured in stages - and more ecologically valid social situations - characterised by real-time recursive dynamics. This second research line in the present work is exploratory in nature, but existing evidence suggests that dynamic interaction should have effects that differ from more static forms of information sharing. Economic bubbles are a long-known example of situations where recursive interaction among social interacting agents lead to self-reinforcing effects and error magnification (De Martino et al., 2013; Mackay, 1841). Phenomena of confidence escalation are known also in the joint decision-making literature (Mahmoodi et al., 2013) and provide modern developments to previous findings in social psychology and risky-shift literature suggesting the presence of non-linear opinion influence dynamics (Heath & Gonzalez, 1995; Moscovici & Zavalloni, 1969; Myers & Lamm, 1976).

The second set of studies reported in the current work (Chapters 5-7) will describe a novel paradigm that aims to disentangle the effects of the communicative means (interaction vs. non-recursive exchange) from the content of communication (i.e., the information shared). Chapter 5 introduces the paradigm, in which pairs of participants perform series of perceptual discriminations in parallel, expressing each time their confidence in their judgment. On each trial, after they have entered their responses, they are allowed to share their views and are incentivised to continuously monitor their changes in confidence and update their responses accordingly. Importantly, social information is shared either statically, showing only the partner's initial

response, or dynamically by showing the partner's initial response and confidence change. Interaction does matter, in the sense that even if conditions were matched in terms of task-relevant information, dynamically interacting with others produces non-linear patterns like confidence escalation. Chapter 6 and 7 both replicate the results of Chapter 5 and add two control conditions to test alternative more parsimonious hypotheses that key interaction effects are due to limitations in the working memory capacity of dyad members. Results confirmed that the interaction interpretation accounts best for the data.

2

A Judge-Adviser Paradigm

Chapter Abstract

This Chapter describes the two main questions that motivate my first line of work and the paradigm used to adjudicate between alternative answers. The first question is: How does the information provided by social partners impact people's independent beliefs? The second question is: Are internal metacognitive signals useful to inform our judgment about the reliability of social others? I will compare situations where participants have access on a trial-by-trial basis to objective feedback on their partner's and their own performance with situations where such feedback is not available. A first behavioural experiment is presented to provide a proof of concept that people are indeed sensitive to subtle informational differences among their advisers even in situations when objective feedback is not available. In the task participants are asked to provide a confidence judgment on a binary forced-choice perceptual decision. After confirming their answer, they are presented with the judgment and associated confidence of one out of four virtual advisers, whose accuracy and confidence calibration are factorially manipulated. The participant is then allowed to adjust their initial decision and confidence judgment. Feedback is manipulated between-participants: the Feedback group, but not the No-Feedback group, received trial-to-trial feedback that gave information about the accuracy of their own judgment and that of the adviser. Participants' learning about the different advisers was assessed in two distinct

ways. First, explicit subjective trust reports were recorded every other block. Second, implicit advisers' influence on subjective choices and confidence was measured. Results show that participants trust more, and are more influenced by, accurate compared to inaccurate advisers, and by calibrated compared to uncalibrated advisers, and that these differences are observed with and without trial-level feedback. The results demonstrate an ability to discriminate between good and bad advice even in the absence of objective feedback, and raise the question of the strategies that people use to compensate for the lack of feedback when this is not immediately available from the environment. Two alternative hypotheses are suggested that could account for the results.

Problem definition

Situations in which feedback on others' performance is not immediately available are abundant in our life. In many contexts, including in education and health, we rely on advice but may not have immediate feedback or other objective standard with which to judge the reliability of that advice. Yet in these contexts we must learn to distinguish good from bad advice. How we do this, and how reliably we do so, are open questions. One hypothesis is that what defines a reliable source from an unreliable one is the ratio between classification variability among items belonging to different categories and classification variability among items belonging to the same category (Weiss & Shanteau, 2003). One of the limitations of this approach is that it is not based on a learning signal that can be easily reconciled with traditional reinforcement learning models (Sutton & Barto, 1998). The question still remains relatively unexplored.

In many perceptual tasks used in group decision and advice-taking studies, accuracy correlates with confidence - that is, more confidently expressed judgments are, on average, more likely to be correct (Henmon, 1911; Koriat, 2012b). This feature makes confidence judgments valuable when objective feedback is missing, because confidence can serve as a proxy for feedback. The first line of studies in this thesis assesses whether people can detect subtle informational differences among their social partners in the absence of objective feedback and, if so, whether internal metacognitive signals play a role in trust formation and advice reliability estimates. Evidence seems to suggest that this might well be the case. Bahrami et al. (2012a) demonstrated that in the absence of trial-by-trial feedback, participants were able to accrue collective benefit - defined as the difference between group performance and most accurate member's performance - if verbal interaction between participants was allowed. On the contrary providing objective feedback on members' trial accuracy but not allowing for verbal communication was not enough to accrue a collective benefit. This

is an interesting finding considering that what should ultimately drive trust in a partner and the decision to follow their judgment is their accuracy rate. Previous studies had demonstrated that conveying information about confidence between participants is crucial for group performance (Bahrami et al., 2010; Fusaroli et al., 2012; Koriat, 2012b). This raises three main questions: (1) What information is communicated during verbal communication that is not communicated with objective feedback? (2) What cues are available to people to infer the reliability of a social source (e.g., a partner's advice) when feedback is removed? (3) What are the mechanisms underlying opinion change in these contexts?

The hypothesis tested here is that people are able to exploit the trial-by-trial covariation between their own internal sense of confidence and the actual state of the environment to overcome the absence of objective feedback. Confidence has already been shown to be an important factor that allows cognitive control (Botvinick et al., 2001; Fleming & Dolan, 2012), meta-learning (Flavell, 1979) and social coordination (Bahrami et al., 2010; Shea et al., 2014). According to a recent proposal (Bahrami et al., 2010; C. D. Frith, 2012) the main reason we have explicit representations of confidence is so that we can inform others of the degree of belief we hold in our internal decisions. We here explore the idea that confidence can also be a signal that can be fed back to assign value to external sources of information when objective feedback is not readily available. In this respect, confidence might act as an internal probabilistic feedback akin to probabilistic external feedback commonly used in classical reinforcement learning paradigms (Sutton & Barto, 1998).

According to recent developments in confidence studies (Aitchison et al., 2015; Meyniel et al., 2015; Pouget et al., 2016), confidence should be interpreted in a probabilistic framework as the subjectively-estimated likelihood of having made the correct decision, given the internal sensory states and the decision made. In a social scenario and given its probabilistic relation with accuracy subjective confidence could in principle be used to infer the likelihood that a social adviser's belief is correct. External

feedback gives you objective information about whether a piece of advice is correct or incorrect. However confidence too can convey a similar, subjective and probabilistic estimate of accuracy. For example, suppose you are certain that your choice is correct, then you might treat advice that disagrees with your decision as certainly wrong and down-weight your trust in the source of that advice accordingly. If you make a choice with less confidence, you should down-weight disagreeing advice but to a lesser extent.

The next section will present the first experiment of this series. Four virtual advisers with differing reliability were designed. The choice of having virtual advisers was to allow precise control over their advice. They differed in accuracy (proportion correct) and calibration (how expressed confidence scales with probability correct). Participants repeatedly experienced each adviser so to give them the opportunity to learn about adviser individual reliability. Learning was examined in terms of both explicit rating of trust in each adviser, and implicit expression of advice influence. This was done to test possible dissociations between explicit and implicit trust. A key question is whether participants are sensitive to differences in reliability even when not provided with feedback about their performance (No-Feedback group). A subsidiary question is how this sensitivity might differ relative to the case with explicit feedback (Feedback group).

Experiment 1

Methods

Participants

Volunteers (N = 46, females = 26, age = 23 $\pm$ 0.45) were recruited in exchange of monetary compensation or university credits. Half of the participants were assigned to the Feedback condition and the other half to the No-Feedback condition. The study was approved by local ethical committee. All participants gave informed consent prior to participation.

Paradigm

To investigate these questions in a quantitative manner, we implemented a computerised version of the classic Judge-Adviser System paradigm (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). In this task a judge (typically the participant) gives their response to a set of general knowledge questions or estimation tasks. The judge is then given the response of another individual (the adviser) and is asked to revise the original judgment. The size of the opinion update quantifies the influence of the advice.

In our version of the task, participants perform a series of difficult perceptual discriminations, after each one expressing their confidence in their decision. They are then told the opinion of a virtual adviser and are then given the opportunity to update their original decision and confidence rating.

The perceptual task used is a dot-count comparison task already described in Boldt and Yeung (2015). The task consists of rapid visual presentation of two boxes, presented for 160 ms on the left and on the right of a fixation cross, containing dots arranged in random order. Participants are asked to determine which box contains more dots (Figure 2.1). Each box is a grid of 20x20 positions that can each contain a white dot or an empty space. One box is chosen to contain more dots, specifically $ndots = 200 + d$. The other box has $ndots = 200 - d$. By manipulating the d parameter we can control the difficulty of the task. For example if $d = 1$ then the two boxes are equal but for two dots. Difficulty was titrated to each individual sensitivity by applying a 2-down-1-up staircase procedure (Treutwein, 1995) that ensured that all participants experienced on average an equal number of correct responses and errors (nominal accuracy rate = 70.7%). The location of the box with more dots was predetermined in advance by a pseudo-randomisation that ensured the number of left and right correct answers was balanced across the experiment.

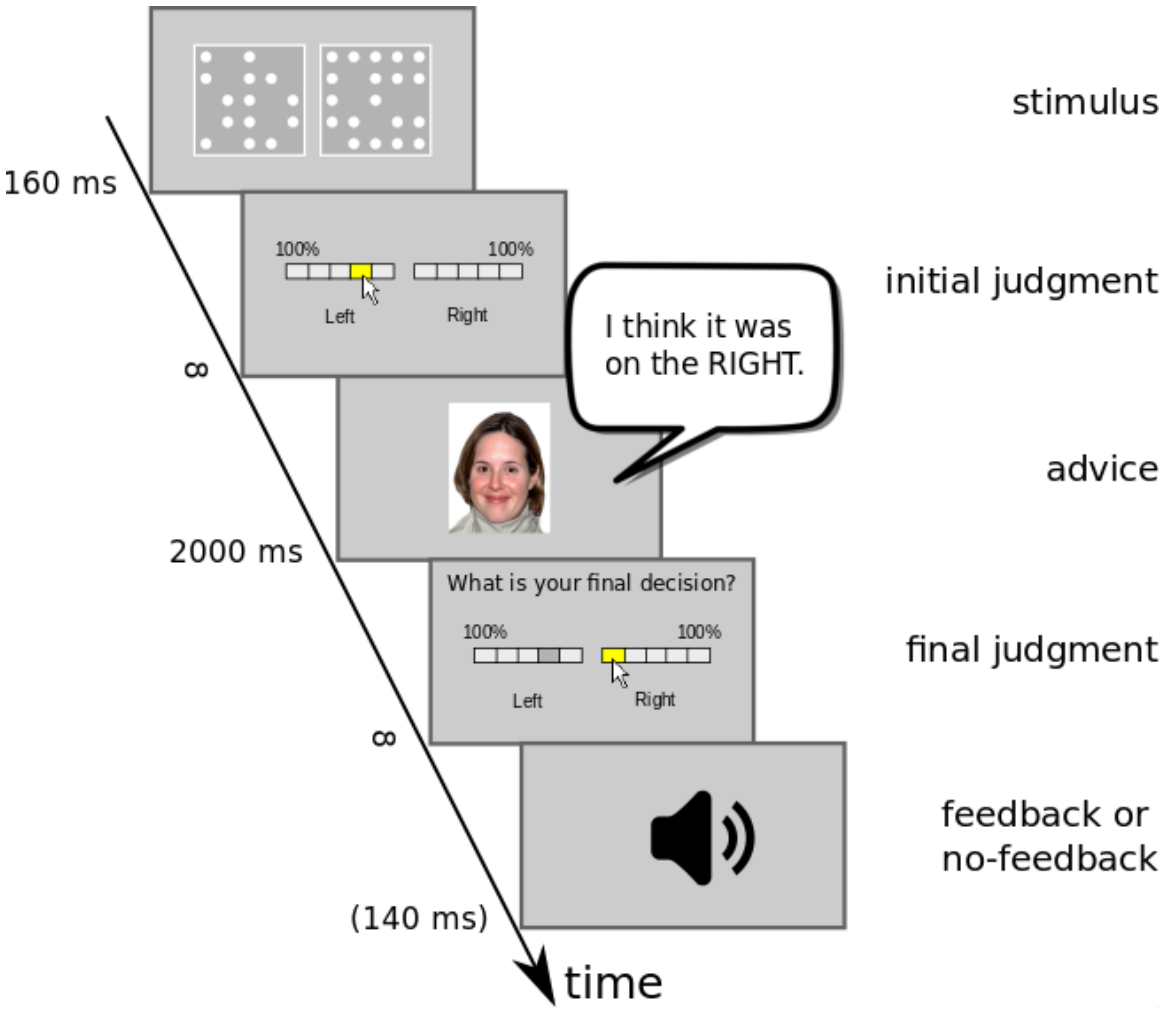


Figure 2.1: The task represents a computerised version of the classical Judge-Adviser System pioneered in organisational psychology (Sniezek & Buckley, 1989). In our version, a participant makes a decision about which of two briefly presented boxes containing dots arranged in random order, contains most dots. The participant expresses his/her opinion on a semi-continuous confidence scale ranging from "100% sure left" to "100% sure right"

After the brief visual presentation of the stimuli (160 ms) participants had unlimited time to enter their response and confidence judgment. They did so in a one-step decision by clicking with the mouse along a semi-continuous scale in 10 steps - ranging from "100% sure left" to "100% sure right" - and confirming their response pressing space bar. Text landmarks signalling 10% increases aided the interpretation of the scale. The middle point of the scale (50% or total uncertainty) was removed and a gap appeared instead, meaning that participants had to commit to one interval

(2-alternatives forced-choice).

After confirming their response, one out of four different advisers appeared at the center of the screen as a head-shot picture. Pictures were selected from the NimStim database (Tottenham et al., 2009) and depicted four Caucasian, smiling female characters that were randomly assigned to accuracy/calibration conditions described below for each participant. Advice was provided as a pre-recorded female voice through active noise-cancelling headphones. Audio tracks were recorded in from native English speakers and their duration was altered with Audacity® so that all lasted for exactly two seconds. The association between adviser voice, face and advice profile was randomised across participants to avoid confounds due to appearance or voice characteristics. Advisers could express a binary level of confidence (low vs. high) and either agree or disagree with the participant's judgment. Low confidence was expressed by the sentences "I think it was on the [LEFT/RIGHT]" and "It was on the [LEFT/RIGHT], I think", with one of the two randomly assigned on every trial to avoid over-repetition of a single sentence. Similarly, high confidence was expressed by the sentences "I'm sure it was on the [LEFT/RIGHT]!" and "It was on the [LEFT/RIGHT], I'm sure!". The selection of LEFT or RIGHT depended on the adviser's choice and accuracy as described below.

After the advice was given, participants had unlimited time to update their decision and confidence level using the same interface and input method as used in the pre-advice period. The question "What is your final decision?" appeared to prompt the update. The pre-advice confidence level remained on the scale as a shaded marker to remind participants of their initial opinion. During the post-advice part participants were allowed to stay in the same position along the scale, increase and decrease their confidence or even change their minds (i.e. changing interval along the scale). Again they confirmed their final decision by pressing space bar. In the Feedback group only, after the final decision was confirmed, a high frequency error tone communicated whether the participant's final decision was incorrect. In the No-Feedback

group, a new trial started immediately after participants had confirmed their final answer.

At the end of each block, a summary on the post-advice percentage accuracy of the participant (but not of the adviser) was provided to both groups. Notice that advisers presentation was balanced within blocks so the feedback participants received at the end of each block could not favour one adviser over the others. Participants performed 500 trials divided in 10 blocks. Two extra blocks with a fifth adviser served as practice and were removed from all the analyses. On each experimental block, each adviser appeared ten times. Ten randomly selected trials within each block were presented with a black silent silhouette and a post-advice decision was not required (null trials). This was done to motivate participants to provide meaningful answers in their pre-advice answers on each trial and avoid pre-advice random guessing.

Manipulation

We orthogonally manipulated the average accuracy of the four advisers and their confidence-to-accuracy calibration (Table 2.1), keeping the raw numbers of confident and unconfident advice equal across advisers (50% high confidence rate). We thus defined two accurate advisers by setting their accuracy to 80% and two inaccurate advisers by setting their accuracy to 60%. Then we set the calibrated advisers to be always accurate whenever confident, and the uncalibrated advisers' confidence to be entirely unpredictive of accuracy. This led to the profiles shown in Table 2.1.

As described in Chapter 1, calibration is a term used in metacognitive studies (Fleming & Lau, 2014) indicating the strength of co-variation between confidence judgments and accuracy. Confidently given advice is more likely to be correct when it comes from a calibrated adviser than an uncalibrated adviser. Several measures exist to quantify calibration (Fleming & Lau, 2014): some (like meta-d') assume a generative model underlying confidence judgments generation, some others (like type 2 A_{ROC}) do not. In this paradigm, uncalibrated advisers all had a type 2 A_{ROC} of 0.5,

	Advisers			
Events count	Accurate Calibrated	Accurate Uncalibrated	Inaccurate Calibrated	Inaccurate Uncalibrated
Incorrect Confident	0	1	0	2
Incorrect Unconfident	2	1	4	2
Correct Unconfident	3	4	1	3
Correct Confident	5	4	5	3
$A_{ROC}II$	.72	.5	0.84	.5
IG	0.29	0.26	0.38	0.08
IG_e	0.063	0.063	0.084	0.021

Table 2.1: Values in the central section represent the number of times each event occurred over the course of a 50-trial block (10 null trials not shown).

meaning that confidence was totally uninformative in predicting the adviser's trial-level accuracy. Due to the experimental design - where overall accuracy of advisers was fixed at 60% or 80%, and where calibrated advisers were always correct when high in confidence - calibrated advisers differed in their metacognitive sensitivity according to this metric (Accurate Calibrated adviser type 2 $A_{ROC} = 0.72$; Inaccurate Calibrated adviser type 2 $A_{ROC} = 0.84$).

We formalised an adviser's informational value as the mean absolute information gained after each possible social event with that adviser, with social events being "Confident disagreement", "Unconfident disagreement", "Unconfident agreement", "Confident agreement" (where "Confident" and "Unconfident" refer to the level of confidence expressed by the adviser on that trial). Information gain is the difference between the posterior and prior probability of participant's correct response:

$$IG = p(d = w|e) - p(d = w) \tag{2.1}$$

where posterior probability correct $p(d = w|e)$ represents the probability that the

decision d is equal to the correct decision w, conditional on the advice received. e represents one of the four possible social events: the adviser (1) confidently disagrees, (2) unconfidently disagrees, (3) unconfidently agrees, (4) confidently agrees. Posterior probability $p(d = w|e)$ was computed with Bayes' theorem and was proportional to participant's prior probability correct $p(d = w)$ and the likelihood of the social event given participant's accuracy $p(e|d = w)$. Given the staircase procedure, we used 70% as prior $p(d = w)$. The probability of agreement (or disagreement) conditional on correct response and the overall probability of agreement (or disagreement) were known by design. The mean absolute information gain so computed (Table 2.1) was lowest for the Inaccurate Uncalibrated adviser (0.08), intermediate for the Accurate Calibrated and Accurate Uncalibrated advisers (0.29 and 0.26 respectively) and the highest for the Calibrated but Inaccurate adviser (0.38). This can be intuitively understood by looking at Table 2.1. Although the Inaccurate Calibrated adviser's accuracy rate is lower than the Accurate Calibrated adviser, outcomes can be better predicted by its judgment. In particular, its judgments correlate strongly positively when sure and strongly negatively when unsure with the correct answer. On the contrary when the Accurate Calibrated adviser is unsure there is a much higher uncertainty about the final outcome. We also computed an expected information gain IG_e for each adviser (Table 2.1) by scaling IG by the overall probability of each event:

$$IG_e = IG * p(e) \tag{2.2}$$

where $p(e)$ is the overall probability of each social event (i.e. Confident disagreement, Unconfident disagreement, Unconfident agreement, Confident agreement). The expected information gain captures the idea that extremely informative but very unlikely events are not very valuable. IG_e values for each adviser were: Accurate Calibrated = .063, Accurate Uncalibrated = .063, Inaccurate Calibrated = .084, Inaccu-

rate Uncalibrated = .021; again suggesting that the Inaccurate Calibrated adviser's advice was the most informative.

Measures of interest

Two measures of interest were assessed to understand how different advice profiles affected their perceived reliability. The first was the explicit trust that participants expressed in the advisers. Every two blocks participants answered a brief questionnaire about their explicit opinions about the four advisers. Four questions asked participants to directly rate on a scale from 1 ("Not at all") to 50 ("Extremely") how much they thought each adviser was accurate (Q1), confident (Q2), trustworthy (Q3) and influential on their own choices (Q4). The first questionnaire was presented immediately after the practice blocks but before any interaction with the advisers took place to provide a baseline measure. Baseline ratings were removed from following ratings to account for confounding factors related to advisers' appearance and inter-individual differences in the use of the scale. A principal component analysis (PCA) was performed for dimensionality reduction on normalised difference scores and the first component was taken as a unitary measure of expressed trust. Question loadings for the Feedback group were: Q1=.52, Q2=.44, Q3=.50, Q4=.52. Loadings for the No-Feedback group were: Q1=.51, Q2=.39, Q3=.54, Q4=.52.

The second measure of interest was an implicit index of adviser's influence on participant's opinions, quantifying participant's confidence changes from pre- to post-advice:

$$\delta_C = C_{post} - C_{pre}. \tag{2.3}$$

where C_{pre} is an integer value between +1 and +5 and C_{post} is an integer value between -5 and +5 (negative C_{post} representing changes of mind). Positive δ_C values mean increases in confidence from pre- to post-advice and negative values represent

decreases in confidence. Notice that δ_C values have a negative skew, ranging from -10 (moving from highest confidence in one judgment to highest confidence in the opposite judgment) to +4 (moving from lowest to highest confidence rating for a single judgment). Agreement and disagreement trials have typically opposite effects on confidence change: agreement usually leads to increases in confidence while disagreement to confidence decreases. The absolute magnitude of confidence shifts in both agreement and disagreement trials can be expected to grow larger as the participant makes more use of the advice received. Thus a unitary measure of influence was obtained by subtracting average δ_C in disagreement from average δ_C in agreement:

$$I = \bar{\delta}_C^a - \bar{\delta}_C^d \tag{2.4}$$

where I assumes greater values as participant's confidence increases in agreement and confidence decreases in disagreement become larger. All the following analyses were performed in MATLAB (The MathWorks Inc., 2016) and R (R Core Team, 2014).

Exclusion criteria

An exclusion criterion was *a priori* set for staircase convergence. Participants who showed progressively increasing thresholds (i.e., increasing d values across the experiment) were eliminated as this indicated that they were randomly guessing. None of the participants had to be removed when this criterion was applied to our sample. At the end of the experiment the average difficulty parameter d across participants (pooled data) was 9.6 ± 2.81.

Results

The following analyses were performed to show that participants were sensitive to all advice dimensions that were manipulated within participants, namely accuracy rate, calibration and advice confidence. Of particular interest is to know whether these

main effects vary as a function of the feedback presence (i.e., the between-participants manipulation). If the pattern of trust and influence changes as a function of feedback presence vs. absence, this would indicate that different advice dimensions are made more or less salient by the presence of feedback. An interaction with feedback would then indicate that the trial-level feedback is important to appropriately evaluate one's advisers. These questions were investigated through the analysis of both explicit trust ratings and implicit influence I measure. Given that in all analyses time was not a significant factor (e.g., when data were pooled across 10 successive blocks to give a factor of time-on-task with 5 levels), the data from all blocks were collapsed together.

Trust ratings

Between-participants analyses. A first analysis focused on participants' explicit ratings of adviser reliability that participants provided at the end of every second experimental block. Ratings were converted into a unitary measure of trust, after preprocessing steps involving normalisation, baseline correction and PCA as described above. A mixed-design ANOVA was run to test whether feedback (between-participants), adviser accuracy (within-participants) and adviser calibration (within-participants) affected explicit trust ratings. Results showed significant main effects for adviser's accuracy ($F(1, 44) = 9.68, p = .003, \eta^2_G = 0.079$) and adviser's calibration ($F(1, 44) = 12.32, p = .001, \eta^2_G = 0.076$), but not for feedback ($F < 1$). Participants trusted accurate over inaccurate advice and calibrated over uncalibrated advice. No interaction term reached significance ($F(1, 44) < 1.9, p > .16$). Importantly, neither of the main within-participants effects interacted with feedback, suggesting that participants were sensitive to the reliability of the advice and that this sensitivity did not depend on the presence or absence of feedback (Figure 2.2).

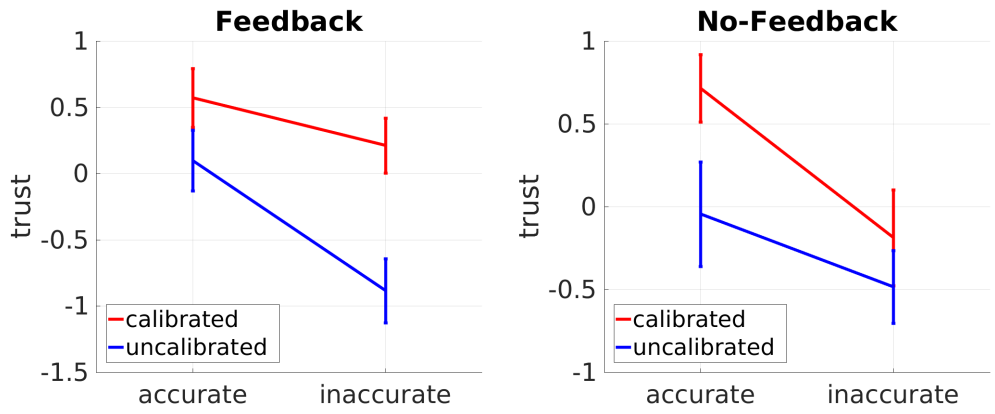


Figure 2.2: Average trust measure in the two feedback groups as a function of adviser accuracy and calibration. Error bars represent s.e.m.

Feedback Condition. We next ran separate analysis for each feedback group individually to identify specific patterns within each group. The trust measure obtained from participants receiving trial-by-trial feedback was subjected to a two-way ANOVA with factors adviser accuracy (low vs. high) and adviser calibration (low vs. high). The analysis revealed reliable main effects of both adviser accuracy ($F(1, 22) = 8.26, p = .008, \eta_G^2 = .09$) and adviser calibration ($F(1, 22) = 8.71, p = .007, \eta_G^2 = 0.12$), but no reliable interaction ($F(1, 22) = 1.24, p = .27, \eta_G^2 = .02$). As shown in Figure 2.2 left panel, when objective feedback is provided on a trial-by-trial basis, trust is determined by both the average accuracy of the advisers but also the confidence-to-accuracy calibration.

No-Feedback Condition. I then looked at the data from the No-Feedback group, where participants could not form an opinion about their partners based on objectively reliable feedback. A 2x2 ANOVA on adviser accuracy (low vs. high) and calibration (low vs. high) showed that both main effects of accuracy ($F(1, 22) = 3.42, p = .07, \eta_G^2 = 0.06$) and calibration ($F(1, 22) = 4.03, p = .05, \eta_G^2 = 0.04$) were marginally significant. The interaction term was not significant ($F < 1$). Importantly however the results do not show a radically different pattern from the Feedback group, sug-

gesting that participants were able to distinguish advisers based on other non-explicit cues.

Influence

Between-participants analyses. Figure 2.3 plots averaged influence (I), the implicit measure of learning about adviser reliability. These results were analysed using a mixed-design ANOVA with the same factors of feedback presence (between-participants), adviser accuracy rate and calibration (both within-participants). For the influence measure an extra within-participants factor was represented by the confidence expressed by the adviser. While both adviser's accuracy and calibration are "trait" variables (with values low vs. high), advice confidence is a trial-by-trial "state" variable (also with values low vs. high). Results showed significant main effects of adviser's accuracy ($F(1, 44) = 14.80, p < .001, \eta^2_G = 0.02$) and calibration ($F(1, 44) = 15.84, p < .001, \eta^2_G = 0.01$), mirroring the observed effects of adviser reliability seen for explicitly expressed trust. A reliable main effect of adviser's confidence ($F(1, 44) = 55.82, p < .001, \eta^2_G = .12$) indicated that more confidently expressed advice had greater influence. There was a significant interaction between calibration and adviser confidence ($F(1, 44) = 9.62, p = .003, \eta^2_G = 0.004$) indicating that the greater influence of calibrated advice over uncalibrated advice was larger when advice was expressed confidently rather than with uncertainty. The results showed also a significant 3-way interaction between adviser's accuracy, calibration and confidence ($F(1, 44) = 4.75, p = .03, \eta^2_G = 8.5e - 04$), indicating that the two-way interaction between calibration and confidence was larger for inaccurate advisers than accurate ones. Importantly, there was no reliable main effect of feedback ($F < 1$), nor any reliable interaction between feedback and any other main effects (all $F(1, 44)s < 1.1$) suggesting again that participants' sensitivity to adviser reliability, here expressed in terms of the influence of their advice, did not depend significantly on the provision of trial-by-trial feedback.

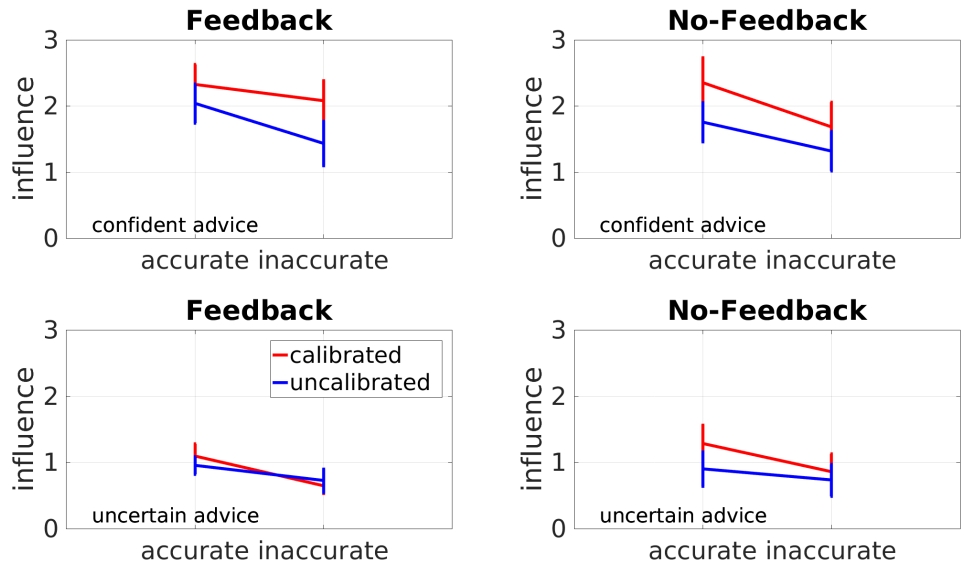


Figure 2.3: The figure shows the effect of adviser accuracy, advice confidence and calibration, and feedback group on influence measure I, which reflects the difference in confidence on trials with agreement vs. disagreement with each adviser. Error bars represent s.e.m.

Feedback Condition. We then looked at each feedback group individually to identify specific patterns within each group. A three-way repeated measures ANOVA on the influence measure I for participants in the Feedback condition with factors adviser's accuracy, adviser's calibration and advice confidence showed significant effects for accuracy ($F(1, 22) = 8.23, p = .008, \eta_G^2 = 0.02$), calibration ($F(1, 22) = 6.35, p = .01, \eta_G^2 = .03$) and advice confidence ($F(1, 22) = 32.92, p < .001, \eta_G^2 = 0.18$). A two-way interaction was found between adviser's calibration and confidence ($F(1, 22) = 7.37, p = .01, \eta_G^2 = .008$) but not between accuracy and calibration nor accuracy and advice confidence (both $F < 1$). As shown in Figure 2.3, the interaction term represents the fact that the average influence difference between calibrated and uncalibrated advisers was mainly shown for highly confident advice compared to uncertain advice. Finally a reliable three-way interaction between accuracy, calibration and adviser's confidence ($F(1, 22) = 8.32, p = .008, \eta_G^2 = .003$) suggests that such interaction was stronger for inaccurate advisers than accurate ones (Figure 2.3).

No-Feedback Condition. I then turned to analyse influence differences in the No-Feedback condition. A 2x2x2 ANOVA showed main effects for advisers' accuracy rate ($F(1, 22) = 6.87, p = .01, \eta^2_G = .02$), calibration ($F(1, 22) = 9.48, p = .005, \eta^2_G = .01$) and advice confidence ($F(1, 22) = 22.95, p < .001, \eta^2_G = .07$). The results show that participants were more influenced by accurate and calibrated advisers, and by confident advice compared to unsure advice, paralleling the effects observed when feedback was provided. None of the two-way interaction terms reached significance (Accuracy rate x calibration: $F(1, 22) = 1.35, p > .25, \eta^2_G = .001$; accuracy rate x advice confidence: $F(1, 22) = 2.90, p = .1, \eta^2_G = .002$; calibration x advice confidence: $F(1, 22) = 2.59, p = .12, \eta^2_G = .001$), and nor was the three-way interaction ($F < 1$). Similarly to the Feedback case, albeit not significant, greater effects of confidence were found for calibrated advisers.

Discussion

The current experiment was an implementation of the classical Judge-Adviser system paradigm (Sniezek & Buckley, 1989; Yaniv & Kleinberger, 2000), in which a participant is asked to form a judgment, then is presented with the opinion of a social partner and is finally asked to provide a final answer. It differs from previous studies in the literature in that it uses highly controlled stimuli for the decision to be based upon, to provide precise control over the amount of information presented to the participants. Furthermore the use of computerised advisers similarly allows a full control over variables of interest, including accuracy rate, calibration and expressed confidence. The results provide evidence that the presence or absence of objective feedback had little consistent effect on participants' estimates of advisers' reliability. This was the case both when looking at explicit reports of trust, as prompted by our questionnaires, as well as when looking at more subtle measures of implicit advice influence on opinion change.

The high degree of control over variables of interest allowed us to have participants with different sensitivities in the primary perceptual task converging to the same accuracy rate. This ensured that advice was equally useful for all participants. It also allowed us to know in advance the expected accuracy rate of each participant and thus to formally estimate the information that the participant could have gained by listening to the advice. The fact that advisers did not differ in terms of the number of times they expressed confident vs. uncertain advice excludes the possibility that participants relied on adviser's over- or under-confidence to form a stable opinion about adviser's reliability (Price & Stone, 2004).

Participants seemed to trust more, and be more influenced by, advice characterised by high confidence, high accuracy and high calibration. Although the first two dimensions are not a surprise (Sniezek & Van Swol, 2001), the fact that participants also value advice calibration is not as well documented. Tenney et al. (2007) showed that in mock jury decisions confident but inaccurate witnesses are discredited more and more quickly than equally inaccurate but less confident ones. Sah et al. (2013) further confirmed that when feedback is readily available confident but incorrect advice backfires on the adviser's reputation. However when feedback was not available the same authors found evidence for a confidence heuristic (Price & Stone, 2004) whereby confident advisers were more influential regardless of accuracy. The current study extends this literature in two respects. First, it replicates in a highly controlled perceptual task setting the finding that judges value adviser's calibration. By allowing repeated interactions with the same advisers and by fixing advisers' average expressed confidence, the current paradigm was able to quantify advice calibration and show that people value it over and above confidence. Secondly, it shows that people trust more, and are more influenced by, calibrated advisers even when feedback is removed but advisers could not be distinguished based on a confidence heuristic (Price & Stone, 2004; Sah et al., 2013).

Participants were sensitive to the same dimensions of advice both with and without feedback, although effects of adviser accuracy and calibration on explicit trust ratings were only marginally significant in the No-Feedback condition. This raises the question of what cues participants must have used to estimate advice reliability in the absence of objective feedback. Two alternative explanations are here offered. One possibility is that people simply use the agreement rates of their partners; i.e., how often the adviser agrees with one's own judgment. This follows the intuition that agreement tends to correlate with accuracy assuming independent observations. As an example, consider two perfectly accurate partners performing the task and sharing their views. They will agree on 100% of the trials due to their perfect detection of the correct stimulus. On the other side two partners who are at chance have an expected agreement rate of 50%. This hints to the fact that people could in principle use agreement rate as a heuristic cue for reliability (assuming their performance is above chance[1]) and thus overcome the absence of objective feedback.

An alternative explanation is that people solve the task using a more nuanced strategy. Internal sense of confidence is also another variable that is known to co-vary with objective accuracy (Henmon, 1911; Koriat, 2012b; Roediger III et al., 2012). In many perceptual tasks, people are known to be more likely to give a correct answer when their confidence is high compared to when they are uncertain. Thus participants in the current task might have been able to associate the trial-by-trial variability in their internal confidence with the advice received and so form an estimate of each partner's reliability. Thus a more nuanced model of trust formation would not only take into account raw agreement rates, but also its covariation with metacognitive signals. The following Chapter will describe how to disentangle these two alternative explanations using two simple models of trust formation in judge-adviser systems.

[1]Indeed, any two below-chance participants will tend to agree on the wrong answer.

Conclusions

The present paradigm was implemented to test whether people are able to detect subtle advice differences such as its calibration in a classical Judge-Adviser System paradigm. Specifically, it tested whether the same factors affecting trust formation and opinion change were responsible for the behaviour observed in situations when trial-to-trial feedback was readily available and situations when it was absent. The results show that feedback availability had little effect on final trust formation and opinion change, raising the question of what alternative cues may hint to advisers' underlying reliability. Two alternatives were suggested that will be more thoroughly explored in the next chapters. The first explanation is that people keep track of the number of times each adviser agreed with their own judgments and use this as an implicit measure of trust. The second explanation suggests that internal metacognitive signals, like confidence, play a role in the formation of trust by weighting agreements and disagreements by trial-level uncertainty.

3

A simple model of reliability estimation in feedback-free scenarios

Chapter Abstract

To understand how metacognitive signals might be used to build representations of advisers' reliabilities, I compare different variants of a parameter-free model that accumulates evidence over learning based on different cues associated with each adviser. When feedback is available on every trial, an *Accuracy* model can use objective feedback to simply count the proportion of correct answers each adviser gives. This base variant provides a baseline against which to compare other special cases. When no feedback is available on the contrary two variants of the baseline model, the *Consensus* model and the *Confidence* model, use non-metacognitive and metacognitive proxies, respectively, to estimate the reliability of each adviser. In this Chapter the variants are presented and compared to the behavioural data of Experiment 1. Their different behaviour will however become apparent only in the experiments presented in the next Chapter.

Model Description

Experiment 1 showed that people are sensitive to similar dimensions in the advice they receive both when objective feedback is available and when it was not provided. But so far it is unclear what cues people are following to estimate their partners' reliability when feedback is taken away. Two simple explanations can be offered. The first one is that different advisers agreed differently often with participants and this in turn was taken as an indicator of good performance. In a binary choice task, if we assume that two people's judgments are independent, agreement rate between the two will scale linearly with the accuracy of each individual as long as performance is above chance. Given that participants knew their performance was greater than 50% because they received a summary feedback at the end of each block, they might have used raw agreement with their partner as an indicator of the partner's good performance. Thus, accumulating the number of agreement events over time for each individual adviser separately allows a participant to form a stable opinion about the other person's underlying accuracy.

An alternative strategy participants could have used is to accumulate over time the estimated probability of the advice being correct on a given trial. This quantity could be generated based on the internal metacognitive signals of confidence, which is a noisy representation of the uncertainty associated with a given perceptual judgment. In other words, given that confidence in a decision is a probabilistic representation of the correctness of one's own decision, it can also be used to estimate the likelihood that the advice received is correct or incorrect. Accumulating such evidence over time can help a decision maker to estimate the reliability underlying advice whenever more secure signals are not available.

To formalise such hypotheses we implemented a simple model that uses different pieces of information depending on different experimental conditions to estimate advisers' reliability. This simple model can then be compared with human observers to provide insight into the strategies they are using to evaluate advice reliability.

Three different model variants are described below that account for the Feedback condition, the No-Feedback condition without metacognitive insight and the No-Feedback condition with metacognitive insight respectively. These models were applied to the data from Experiment 1 to generate estimated reliability of the four advisers (Accurate Calibrated; Accurate Uncalibrated; Inaccurate Calibrated; Inaccurate Uncalibrated) according to these three strategies, based on the actual sequence of events experienced by the participants, in terms of their initial judgment and confidence, whether the adviser agreed or disagreed with this judgment, and whether objective feedback was present and, if so, whether it indicated that the participant and/or adviser had made the correct choice.

Accuracy Model

When objective feedback is given to participants by the experimenter, the model can use it to infer the accuracy rate of its advisers. The accuracy of the adviser ($Acc = \{0, 1\}$) is the same as the accuracy of the participant in agreement trials, while is opposite in disagreement. By counting correct and error rates for each adviser separately, the model obtains a trial-by-trial estimation of the adviser's accuracy rate, θ, as the ratio between the number of adviser's correct trials and the total encounters with that adviser:

$$\theta^i = \frac{\alpha^i}{\alpha^i + \beta^i} \tag{3.1}$$

where α^i and β^i are the correct and error counts respectively, during the past trials with adviser i:

$$\alpha^i = \sum_{t=1}^{n} Acc_t \tag{3.2}$$

$$\beta^i = \sum_{t=1}^{n} 1 - Acc_t \tag{3.3}$$

$t = 1$ here represents the first encounter with adviser i while $t = n$ represents the last one. A slight complication in Experiment 1, however, is that advisers also provided a binary confidence judgment associated with the advice. A simple way for the model to make use of adviser's confidence is by treating it as a linear scaler of the advice received. We applied a set of arbitrary weights to the four possible advice scenarios, namely the adviser is (1) correct and confident, (2) correct but unsure, (3) incorrect and unsure and (4) incorrect but confident (Table 3.1). Although arbitrary, any set of weights that preserves the order of such events would result in similar final adviser preferences.

	Event Observed			
	Inaccurate Confident	Inaccurate Unsure	Accurate Unsure	Accurate Confident
Feedback	-1	-0.5	+0.5	+1
	Disgree Confident	Disagree Unsure	Agree Unsure	Agree Confident
No-Feedback	-1	-0.5	+0.5	+1

Table 3.1: Model weights (w) applied to different advice events observed in the Feedback and No-Feedback scenario.

Thus instead of simple accuracies α and β in equations 3.2 and 3.3 can now be reformulated as:

$$\alpha^i = \sum_{t=1}^{n} .5 + .5 * w_t \tag{3.4}$$

$$\beta^i = \sum_{t=1}^{n} .5 - .5 * w_t \tag{3.5}$$

This set of equations results in values of 1, 0.75, 0.25 and 0 for the four events listed above respectively. Although these values could be simply summed to obtain α and β values, the unusual formulation of the equations 3.4 and 3.5 was preferred to be coherent with the equations describing the following models. They show how a

simple model can take into account feedback, advice received and advisers' expressed confidence to track over time the objective reliability of its advisers.

Consensus Model

When feedback is removed from the participants, as in the No-Feedback condition of Experiment 1, the model does not have access to the advisers' objective accuracy. It must then rely on different proxies for objective accuracy and integrate those instead over time. The first cue to underlying accuracy rate we considered is agreement rate. When two independent agents express judgments on a binary task, the agreement rate between the two linearly scales with the accuracy of each whenever the accuracy rate is higher than chance: $Agr = Acc_1 * Acc_2 + (1 - Acc_1) * (1 - Acc_2)$. We thus adapted the equations of the *Accuracy* model above to exploit this covariation. Instead of tracking the accuracy rates of its advisers, the *Consensus* model tracks their agreement rates with subjective judgments. Thus equations 3.4 and 3.5 can be used to estimate a θ value by now using as w_t the scaled agreement observed on encounter t as described in Table 3.1. To take into account the fact that in Experiment 1 advisers expressed a binary confidence judgment themselves associated with the advice, we used the same linear weights applied to the *Accuracy* model also to scale agreement (Table 3.1). In other words this model perfectly conflates accuracy with agreement, assuming that whenever an adviser agrees with the subjective original judgment, the adviser must be correct. Although this clearly is a simplifying assumption, the model offers a useful proof of concept to understand what inferences an agent lacking metacognitive insight can make simply by using heuristics. It can thus provide a benchmark to quantify the information that is present in the advice received.

Confidence Model

A more nuanced strategy that could be employed to estimate advisers' reliability when feedback is not directly available is through use of internal metacognitive signals. Trial-level variability in subjective confidence is known to covary with objective

accuracy in a perceptual task (Henmon, 1911) and it theoretically represents the estimated likelihood of having made a correct judgment and/or selected the correct response (Pouget et al., 2016). Thus, instead of simply using agreement rates as a cue for accuracy rate, a model endowed with metacognitive insight could accumulate over time the subjective probability that an adviser expressed a correct judgment. A *Confidence* model was created under the assumption that the trial-by-trial subjective reports of confidence are directly related to the true underlying estimated probabilities of having chosen the correct answer. On agreement trials the model estimates the probability of the advice being correct as the subjective probability of a correct answer. Conversely on disagreement trials the model estimates the probability of the advice being correct as the probability of having itself made an error. In other words trial-level agreement ($Agr = \{0, 1\}$) is scaled by trial-confidence expressed as a probability over outcomes (correct vs. incorrect response). Thus equations 3.4 and 3.5 above become according to this model:

$$\alpha^i = \sum_{t=1}^{n} .5 + (p_t(corr) - .5) * w_t \tag{3.6}$$

$$\beta^i = \sum_{t=1}^{n} .5 - (p_t(corr) - .5) * w_t \tag{3.7}$$

where w_t represents the scaled trial-level agreement as described in Table 3.1 and $p(corr)$ represents pre-advice confidence. As described below, rather than taking participants' confidence as a pure index of subjective $p(corr)$, we transformed the value to (1) reduce inter-participants variability and (2) increase scale sensitivity. Regardless, the crucial point is that this model capitalises on the fact that being in agreement or disagreement with an adviser is more informative when the model is itself confident that it gave a correct answer than when it is more likely to have made a mistake.

Bayesian update

All model variants can use the current estimated adviser's reliability θ to appropriately update the pre-advice probability of having selected the correct answer $p(corr)$ into a normative posterior, based on the binary advice received A (agree=1 vs. disagree=0):

$$p(corr|A^i) = \frac{p(corr)p(A^i|corr)}{p(corr)p(A^i|corr) + p(err)p(A^i|err)} \tag{3.8}$$

where $p(err)$ is the subjective probability of making a mistake on the current trial and $p(A^i|corr)$ is the probability that adviser i agrees ($A = 1$) or disagrees ($A = 0$) given that the participant's choice is correct. Prior probability $p(corr)$ is estimated from a simple linear transformation of the pre-advice trial-level confidence data obtained from the participants after appropriate pre-processing. Pre-processing consisted of a parameter-free transformation that (a) brings all subjective confidence distributions on to a similar scale thus reducing the inter-participant variability and (b) expands the centre of subjective distributions so to increase the informativeness of the average trial. This operation was inspired by recent models of adaptive information gain control (Cheadle et al., 2014). According to these proposals, the brain adapts the gain of neuronal firing to the range of information available over different time scales and cognitive domains (Carandini & Heeger, 2011; Cheadle et al., 2014). Here it serves the purpose of increasing the discriminability or information gain of different trials so that trials that are close together on confidence scale gets pulled apart on to a probability scale. The transformation uses parameters obtained from the data:

$$\hat{C}_{pre} = N * normcdf(C_{pre}) \tag{3.9}$$

where $normcdf(C)$ is the normal cumulative density function of the pre-advice confidence C ratings distribution, and N is the number of confidence ratings available on each interval of the scale (in Experiments 1,2: N = 5; in Experiment 3: N = 50).

This simple transformation has the property of translating a normal distribution into a uniform distribution in the range $[0, N]$. Notice that this transformation does not affect the ranking of confidence judgments but only their spacing along a probability scale. After pre-processing, confidence ratings were translated into a probability scale with the linear transformation:

$$p(corr) = 0.5 + (0.1 - \epsilon) * \hat{C}_{pre} \tag{3.10}$$

where ϵ is a small jitter ($\epsilon = .002$) introduced to avoid maximum confidence ratings being turned into probability of one and zero, which would in turn cause inconsistencies within the Bayesian formula (e.g., no confidence change regardless of advice reliability). Thus $p(corr)$ represents trial-level confidence on a probability scale, which can be interpreted as the probability that the participant assigns to having given a correct answer on a given trial. From $p(corr)$ we can also derive the subjective probability that a given trial will end up in an error: $p(err) = 1 - p(corr)$.

To estimate the likelihood term $p(A^i|corr)$ in equation 3.8 we applied a simple heuristic that uses the reliability θ of a given adviser:

$$p(A^i|corr) = \theta^A * (1 - \theta)^{1-A} \tag{3.11}$$

The equation above simply states that the probability of observing adviser i's agreement ($A^i = 1$) when the participant is correct is equal to the accuracy rate of the adviser itself. This follows from the fact that the adviser's and the participant's judgments are independent of each other. Instead the probability of observing disagreement ($A^i = 0$) on the same trials is the adviser's error rate. In other words, the probability of agreement in trials when the participant is correct is the probability that the adviser too is correct. Similarly the probability of disagreement in trials when the participant is correct is equal to the probability that the adviser is wrong.

Results

The three variants of the Bayes update model were trained on the data collected for Experiment 1. Analysis looks at how the different variants of the model fared in evaluating different advisers' profiles. The final theta values for each adviser were plotted to compare different model variants' against each other. The aim of this analysis was to verify how the model's estimates of the reliability of its advisers differed when different pieces of information were used to compute θ. For this analysis data from the Feedback and No-Feedback groups were pooled together as the presence of feedback did not reliably affect the variables that model variants were based on, namely advisers' accuracy rates, agreement rates and participant's pre-advice confidence ratings respectively. The weights for all model variants were set according to Table 3.1.

Confidence changes predicted by the model could have been used to plot the model's implicit trust in the advisers in a manner similar to what already shown with the behavioural data. However, contrary to θ values, confidence changes are an implicit measure of trust and they are affected not only by the value assigned to an adviser but also by the pre-advice confidence of that trial. Thus, although the two measures are related, θ represents a direct measure of the model's belief about advisers' reliability (a trial-by-trial measure that is unfortunately not available in the human data).

Accuracy Model

The *Accuracy* variant of the model was first tested on the data collected for Experiment 1. This variant uses each adviser's counts of correct and error trials to estimate adviser reliability. Results are shown in Figure 3.1 below. The *Accuracy* model final θ values, i.e., the model's reliability estimates on the last trial of the experiment, were passed through a 2x2 repeated measures ANOVA with factors of adviser's accuracy (high vs low) and adviser's calibration (high vs. low). Results show a significant

effect of adviser's accuracy ($F(1, 45) = 1.21e + 16, p < .001, \eta^2_G = 1$) and calibration ($F(1, 45) = 6.55e + 15, p < .001, \eta^2_G = 1$) and a significant interaction between the two ($F(1, 45) = 7.28e + 14, p < .001, \eta^2_G = 1$). The fact that for each participant the number of social events were determined *a priori* (Table 2.1) and that the model used only pre-determined variables that were equal for all participants led to a lack of variability across participants in the dependent variable θ. However the results show that when the model was provided with information about the objective performance of the participant (and thus of the advisers), it was able to distinguish advisers both in terms of their accuracy rate and their confidence calibration.

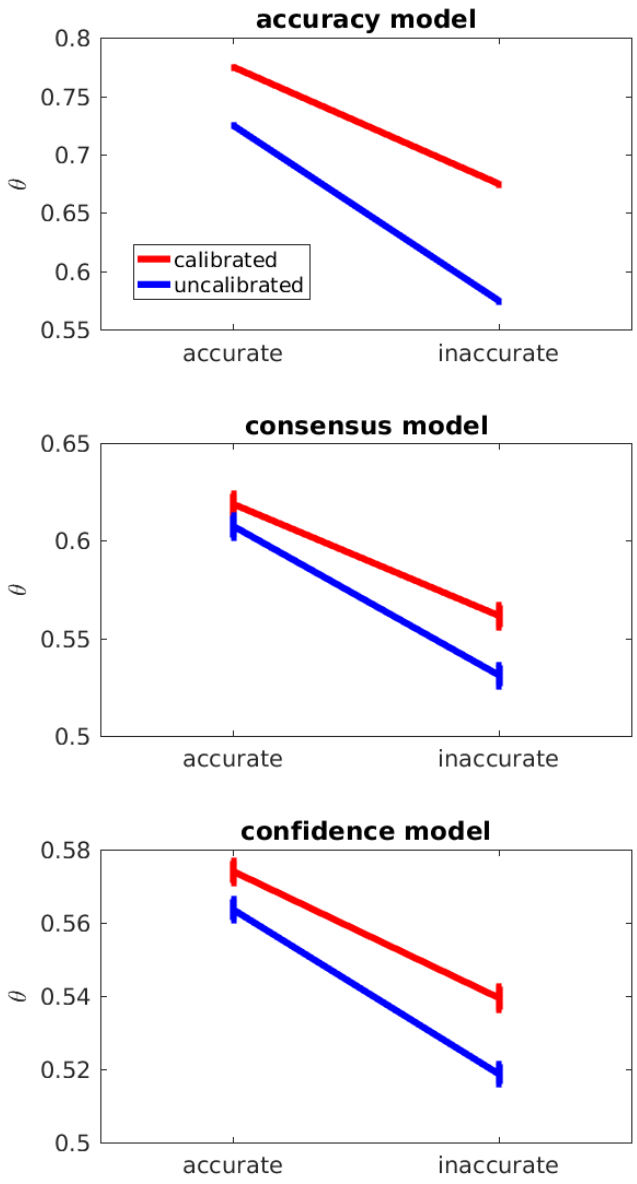


Figure 3.1: Average estimated adviser reliability (θ) according to different variants of the Bayesian model computed from the data of Experiment 1. Notice the different scale on the y-axis. Error bars represent s.e.m.

Consensus Model

Next we tested on the same data the *Consensus* model's reliability values assigned to the advisers. This variant of the model differs from the previous one in the sense that it does not have access to the objective accuracy of the participant, nor consequently of the adviser. It estimates partners' reliabilities by assuming that if an adviser agrees with the participant, the adviser must be correct. Because this model variant conflates agreement with accuracy, different ranges of θ values are output compared to the *Accuracy* variant (and to the *Confidence* one), as shown by the different scales of the y-axes in Figure 3.1. This is due to the fact that agreement rates tend to be lower than accuracy rates[1]. Scaling factors were applied to agreement to take into account the fact that advice was associated a confidence judgment. The resulting θ values from the end of the experiment were passed through a 2x2 repeated measures ANOVA with factors adviser's accuracy (high vs low) and calibration (high vs low). Both a significant effect of accuracy ($F(1, 45) = 165.84, p < .001, \eta^2_G = .44$) and calibration ($F(1, 45) = 14.78, p < .001, \eta^2_G = .07$) were found but no significant interaction between the two ($F(1, 45) = 2.49, p = .12, \eta^2_G = .01$). The results suggest that in the absence of externally given objective feedback, a model could detect small variations in advisers' confidence calibration as well as accuracy rates by simply using weighted agreement rates. Notice that advisers were not constrained to agree with the participant a pre-determined amount of times, thus explaining the variability observed in the dependent variable.

Confidence Model

Finally we tested how the *Confidence* variant of the Bayesian model would have performed in assigning value to each partner's advice. Like the *Consensus* variant, the *Confidence* variant of the model does not have access to adviser's objective accuracy

[1]For example, two agents with an accuracy of 70% will tend to agree only $0.3^2 + 0.7^{=}58\%$ of the times, assuming their judgments are independent.

and must use agreement rates instead. Contrary to the *Consensus* variant, however, it scales trial-level agreement by the subjective probability given by the participant to the adviser decision. For example, instead of simply assigning a value of 0 to a disagreement trial, this model variant assigns a value of 20% if the subjective confidence is 80%, a value of 30% if the subjective confidence is 70% and so on. As with the other models, final θ values were passed through a 2x2 repeated measures ANOVA with factors adviser's accuracy (low vs. high) and calibration (low vs. high). Results show a significant effect for both accuracy ($F(1, 45) = 209.71, p < .001, \eta^2_G = .50$) and calibration ($F(1, 45) = 32.70, p < .001, \eta^2_G = .13$) but no significant interaction between the two ($F(1, 45) = 2.65, p = .11, \eta^2_G = .01$). Results show that the *Confidence* model puts more trust in accurate and calibrated advisers compared to inaccurate and uncalibrated ones. Thus also a model endowed with access to trial-level metacognitive insight is able to perform well at this task and effectively replicate the results both observed in the human data and in the *Accuracy* model.

It thus seems that, at least for Experiment 1, the three models do not make contradictory predictions on which advisers should be trusted. In particular the two No-Feedback models cannot be disentangled using the data collected from this experiment, with both showing sensitivity to both accuracy and calibration of an adviser. This is due to the fact that in the present experiment accuracy, agreement rates and confidence are positively correlated. In the next Chapters however, it will become clearer how - when decorrelating these three variables - the different model variants make different predictions.

Discussion

In this chapter I have introduced a model of the Judge-Adviser task studied in Experiment 1. The model uses available information, depending on condition and metacognitive access, to build over time an estimate of the reliability of each adviser θ. The

results above show that the three variants of the model all trusted accurate advisers more than inaccurate ones, and trusted calibrated advisers over uncalibrated ones, just as did the participants in Experiment 1. In the case of Experiment 1, the three variants do not make different predictions on which advisers should be trusted more. This is particularly important for the two No-Feedback variants of the model, as this makes impossible to disambiguate between the two alternative explanations outlined to explain the results of Experiment 1.

Different models could be compared according to standard AIC/BIC criteria to formally check which model was more likely given the pattern of trust ratings and advice influence shown by human participants. This approach was not however preferred given that different model variants produced different scales in θ value range. The fact that different models computed θ values as accuracy rates, agreement rates and weighted agreement rates respectively, made the variants predict very different posterior confidences, thus making model comparison difficult.

Conclusion

Experiment 1 showed that human observers can overcome the absence of external objective feedback and still form an accurate opinion about their partner profiles, but did not distinguish between alternative explanations of this finding. The modelling here demonstrates that calculating agreement rates and weighting agreement by subjective confidence seem to be similarly effective strategies that can produce estimates of adviser reliability that are similar to those generated when objective feedback is provided. The reason for this is that both agreement and subjective confidence covary with accuracy. Thus, both are useful clues, and the two are difficult to disentangle. To disambiguate between the two hypotheses, one needs to orthogonally manipulate (1) advisers' agreement and accuracy rates and (2) advisers' agreement rates and subjective confidence. Under these conditions, the different model variants

are expected to produce different results. A measure of participants' (and models') opinions about the different advisers so created would be then more informative and could in principle tease these two explanations apart. The next Chapter describes two experiments that were conducted with this purpose in mind.

4

Disentangling two alternative hypotheses

'Whenever people agree with me I always feel I must be wrong.'

– Oscar Wilde

Chapter Abstract

People's judgments are rarely independent. We watch the same news, browse the same websites and are affected by the same cognitive biases. This is likely to influence the extent to which opinions agree or disagree with each other and the information carried by other people's consensus. Two experiments are described in which we manipulated virtual advisers' agreement and accuracy rates (Experiment 2) and different agreement patterns conditional on participants' judgments (Experiment 3). The results of Experiment 2 show that when feedback is available on every trial, both agreement and accuracy are valued in social partners, but that agreement trumps accuracy when feedback is removed, although both seem to be relevant advice dimensions for trust formation. Experiment 3 was designed to isolate the effect of one's own confidence on advice perception from the stronger effect of agreement. In this experiment, three advisers were equal in terms of agreement and overall accuracy, but their agreement probability was conditional on the participant's initial choice and confidence, as if they differed in the extent to which they relied on the same information or biases as

the participant in their decisions. Results show that simple agreement rate *per se* cannot explain the pattern of explicit trust judgments and confidence change that are observed in the data. The model presented in Chapter 3 is here adapted to the new data. A *Confidence* model that weighs agreement by internal confidence provides a better fit to human behaviour than a simple *Consensus* model.

Introduction

The Wisdom of Crowds (WoC) effect (Galton, 1907; Surowiecki, 2004), whereby aggregation of individual opinions results in better-than-expert performance, is often cited as the archetype of collective intelligence. The classic explanation of the effect is that individual judgments are the sum of a signal component and random noise. Averaging individual judgments has the effect of enhancing signal by averaging out noise, in a similar way as repeated measures in statistics (Armstrong, 2001). But individuals' opinions are rarely independent and/or distorted only by random noise (Koriat, 2012b; Krause, Ruxton, & Krause, 2010). People often rely on similar knowledge sources and are affected by the same cognitive biases (Tversky & Kahneman, 1974). Crowds are known to be susceptible to large error cascades (Le Bon, 1895; Mackay, 1841), economic bubbles (De Martino et al., 2013), polarisation (Myers & Lamm, 1976) and *groupthink* (Janis, 1972; Turner & Pratkanis, 1998). A hypothesis is that these examples of error magnification are due to judgment independence breaks.

My investigation started with the puzzle of how, and how reliably, people infer the usefulness of advice in the absence of objective feedback. Experiment 1 showed that people can make subtle distinctions reliably, such that there were no statistically reliable differences in trust and influence with and without feedback. Two model variants capture this effect. One estimates advice reliability in terms of sensitivity to consensus, while the other uses confidence as an internal feedback signal. Crucially, however, both models rely on the association between accuracy and agreement. When judgments are independent, greater agreement rates are observed among accurate observers rather than inaccurate ones. Experiments 2 and 3 aim to extend this approach with two related goals. First, we want to explore the inherent limitations in consensus and confidence-based estimates of reliability. Second, we want to distinguish between

different predictions of the *Consensus* and *Confidence* models and compare these to the observed human behaviour. Key to both aims is breaking the independence of initial participants' judgment and advice. In Experiment 1, these were independent and participants judged evidence reliability regardless of feedback. Here we directly manipulate dependencies between participant's judgment and advice received and two main predictions can be made. First, this manipulation will lead to predictable distortions in trust and influence, that differ from objective reliability of feedback. Second, the patterns of trust and influence are best captured by a *Confidence* model. Understanding of these phenomena promises to shed light onto real-life situations were people are affected by the same errors (Koriat, 2012b; Krause et al., 2010).

Experiment 2

This experiment investigates the question: Do we trust accurate people or people who tend to agree with us? Experiment 1 showed that people are able to detect subtle variations in the reliability of the advice they receive even when objective feedback is not available on a trial-level basis. This finding raises the question of what cues people are sensitive to in these contexts. As described in the previous chapters, the frequency of agreement with the advice might be used as a cue to the reliability of the advice. It is thus interesting to ask whether participants are only sensitive to agreement rate when feedback is unavailable or instead can use subtler cues. We can look at this question by fixing the agreement rate of the advice with participant's judgments and varying its accuracy.

In Experiment 2 four adviser profiles are created from the intersection of two orthogonal dimensions: advice accuracy and advice agreement. In many daily life situations accuracy and agreement among social agents covary whenever the information sources are independent in their judgments. Two highly accurate individuals will tend to agree (on the correct judgment) more often than two individuals who

are less accurate. However, if the two judgments are not independent, the coupling between agreement and accuracy can be broken. It is commonly assumed that judgments are a mixture of signal and random errors, which can be improved by averaging (Armstrong, 2001). But errors can also come from similar sources across observers, thus creating dependencies between judgments. One source of dependence between opinions can arise from the use of similar cognitive heuristics (Tversky & Kahneman, 1974). Other sources can arise from using similar strategies to sample information from the environment (Vandormael, Herce Castañón, Balaguer, Li, & Summerfield, 2017) (e.g., in the current task, if two observers sample from the same portion of visual field), being exposed to similar information (Kao & Couzin, 2014; Kao, Miller, Torney, Hartnett, & Couzin, 2014) or belonging to the same network clique (Jamieson & Cappella, 2008; Jasny, Waggle, & Fisher, 2015; Sunstein, 2001). Common error sources may lead to greater consensus among observers but to a decorrelation from actual accuracy (e.g., by agreeing on errors) (Koriat, 2012b).

In the present experiment, adviser's accuracy is manipulated so that two advisers are on average highly accurate (around 80% accuracy) and two advisers are on average relatively inaccurate (around 60% accuracy). Orthogonally, adviser agreement rate with the participant's judgment is manipulated so as to create two advisers who agree with the participant frequently (around 80% of trials) and two advisers who tend to have a lower agreement rate with the participant (around 60%). This creates an adviser who is highly accurate and tends to agree with the participant, an adviser who is highly accurate but tends to disagree with the participant, an adviser who is not very accurate but still tends to agree with the participant and one adviser who is neither accurate or agreeing. The two intermediate profiles are an important test case where the two dimensions of accuracy and agreement work against each other.

As highlighted above, a situation where accuracy and agreement are disentangled is not simply an experimental artifice, but instead it has several implications for

daily life scenarios. Indeed, it may be rare for two agents to be entirely independent in their judgments outside of the lab. Equally accurate observers who share a common bias will tend to agree more often than two individuals with differing biases but same underlying average accuracy. Knowing what advice characteristics are made salient by the presence of feedback can shed light on the mechanisms of trust formation and advice influence.

In scenarios where agreement and accuracy are disentangled but feedback is available we expect participants to be sensitive to the reliability (i.e., accuracy) of the advice. In scenarios where feedback is absent, on the contrary, both the *Consensus* and the *Confidence* variants of the model predict that participants will be sensitive to the agreement rate of the advice, but only the *Confidence* variant predicts that participants will also be sensitive to its accuracy.

Methods

Participants

The experiment included 46 participants equally divided between the two feedback groups (37 females in total, 18 of whom were in the Feedback group, mean age: 21.63± 3.02). Participants were recruited through Oxford University's online recruitment platform and local news advertisement. Participants were compensated either with money (10£/hour) or course credits. Prior to the beginning of the experiment all participants signed an informed consent form. All procedures were approved by the local ethical committee.

Paradigm

Each trial began with the dot-count task with presentation parameters as described in Experiment 1. Participants performed ten blocks of 44 trials each. Stimulus presentation, response modality, advice presentation and confidence update took place

with identical modalities as described in Experiment 1. Advice was presented in the form of a binary judgment. Advisers expressed their judgments using the sentences "I think it was on the [LEFT/RIGHT]" and "It was on the [LEFT/RIGHT], I think", depending on the adviser's opinion and a random selection of one of the two sentences across trials. For participants in the Feedback group only, after the final choice was confirmed, a high frequency beep sound (duration 140 ms) was played whenever the participant's final choice was incorrect.

At the end of every second block, a computer-based questionnaire was presented to participants prompting them to evaluate the four advisers on a 50-point scale (1="Not at all", 50="Extremely"). Four questions investigated participants' beliefs about the accuracy (Q1), likeability (Q2), trustworthiness (Q3) and influence on own judgments (Q4) of each adviser. The questions were the same as those used in Experiment 1 apart from question two that was replaced to take into account the fact that advisers did not express a confidence judgment in the present experiment. A baseline questionnaire was presented before the beginning of the experiment to collect participants' evaluations of the advisers before any interaction took place.

In each block, the presentation of the four advisers was randomly shuffled across trials, with each adviser appearing exactly ten times. In four randomly selected trials no advice was given and the trial ended immediately after the participant confirmed the pre-advice judgment (null trials), to encourage participants to register meaningful initial confidence judgments as well as final (post-advice) judgments. The difficulty of the first-order task was titrated using a 2-down 1-up procedure so that all participants converged to 70.7% level of accuracy, independent of their sensitivity in the perceptual task. Participants completed two practice blocks, with ten practice trials each, with a fifth adviser. Practice trials were removed from all analyses.

Manipulation

To disentangle the agreement rate from the accuracy of the advisers, the probability of agreement conditional on the participant's choice accuracy was manipulated.

Through the staircase procedure it was expected that all participants would converge to an accuracy level of about 70%. This enabled the experimenter to manipulate the adviser's accuracy and agreement rate by pre-determining the probability of agreement after a participant's error or a participant's correct response. Both accuracy and agreement were manipulated to have two levels (high=80% and low=60%). This gave rise to the four adviser profiles defined in Table 4.1. Probabilities are expressed as a fraction over the number of participants' expected correct (7) and incorrect (3) judgments, during the number of encounters with one adviser in each block (10).

	Advisers			
	High Accuracy High Agreement	High Accur. Low Agreem.	Low Accur. High Agreem.	Low Accur. Low Agreem.
$p(Agree\|Cor_{subj})$	6.5/7	5.5/7	5.5/7	4.5/7
$p(Agree\|Incor_{subj})$	1.5/3	0.5/3	2.5/3	1.5/3
Expected Accuracy rate	80%	80%	60%	60%
Expected Agreement rate	80%	60%	80%	60%
IG	0.28	0.27	0.03	0.06
IG_e	0.09	0.13	0.01	0.03

Table 4.1: Experiment 2 advisers' profiles. Accuracy rate and agreement rates of different advisers are disentangled by manipulating the probability of the advice agreeing with the participant, conditional on the participant's accuracy. In the table, probabilities are expressed as a fraction of the number of participants' expected correct (7) and incorrect (3) judgments, during the number of encounters with one adviser (10). IG and IG_e indicate average informational value of the advice, computed as information gain and expected information gain respectively.

The dissociation between accuracy and agreement was possible by making the disagreeing but accurate adviser disagree more often on trials when the participant had made an incorrect initial decision (meaning a correct response for the adviser). Similarly, an inaccurate but agreeing adviser was created by making the adviser more likely to agree when the participant's initial judgment was incorrect. For the two

advisers for whom accuracy and agreement work in the same direction - i.e., accurate and agreeing vs. inaccurate and disagreeing - clear predictions can be made about the expected trust pattern that the participant will show. However when the accuracy and the agreement pattern work in opposite directions - as it is the case for the agreeing inaccurate and the disagreeing accurate advisers - it can be expected that the availability of external feedback will have different impact on different advice dimensions and the two feedback groups are expected to diverge.

Similar to Experiment 1, we used (as listed in Table 4.1) conditional probabilities and the participants' expected accuracy to compute the informational value of each adviser. Advisers' mean absolute information gain IG and expected information gain IG_e were computed as in the previous experiment. Contrary to Experiment 1, however, advisers did not express different levels of confidence. This created only two possible situations on each trial (instead of four as in the previous experiment), namely either agreement or disagreement. As can be seen in Table 4.1, information gain was highest for the accurate advisers and the lowest for the agreeing but inaccurate adviser. This adviser is in fact counter-intuitively anti-predictive of the correct answer: in contrast with the other advisers, this adviser's probability of agreement is higher when the participant's initial judgment is incorrect than when it is correct. Thus, an optimal observer would decrease their confidence when in agreement and increase it when in disagreement with this adviser. For all other advisers on the contrary, agreement should increase confidence and disagreement should decrease it.

Measures of interest

The same two measures of interest as in Experiment 1 were defined. The first one measures participants' explicit ratings of reliability. A trust index was created as described in Experiment 1. Questionnaire ratings from the four questions about adviser accuracy, likeability, trustworthiness and influence on own judgments were combined into one measure by applying PCA to normalised baseline-corrected scores, resulting

in loadings of 0.53 (Q1), 0.39 (Q2), 0.53 (Q3), 0.52 (Q4) for the Feedback condition and 0.51 (Q1), 0.41 (Q2), 0.53 (Q3), 0.51 (Q4) for the No-Feedback condition. Time was not a significant factor in the analyses, so scores from different blocks were collapsed together in calculating condition averages. The second measure quantifies the impact that advice has on participants' judgments. Given that agreement tends to make participants increase their initial confidence and disagreement tends to reduce it or even lead to changes of mind, an influence measure was created by subtracting the average confidence decrease in disagreement from the average confidence increase in agreement as described by equation 2.4.

For pairwise comparisons, both p-values and Bayes factors are reported. Bayes factors are reported using the notation suggested in Dienes and Mclatchie (2017). The priors on the effect sizes to compute Bayes factors are informed by results of Experiment 1.

Exclusion criteria

The first two experimental blocks were removed from the analysis to allow the staircase procedure to fully adapt to each individual's threshold. This was necessary given that our manipulation was heavily dependent on the expected accuracy rate of the participants. A further exclusion criterion was set to exclude all participants whose threshold never converged, which suggests a random response strategy. None of the participants had to be removed on the basis of this criterion. Average difficulty parameter d was 9.93 ± 2.96 (pooled data).

Results

A first set of analyses explored the effect of our manipulation on the trust and influence measures as a function of whether objective trial-level feedback was provided. In particular we are interested in whether any of the manipulated within-participants factors (advice agreement rate and accuracy) interacted with feedback presence. An

interaction term between a within-participants advice dimension with feedback would indicate that the specific advice dimension's effect on reliability estimates is modulated by the presence of trial-level feedback.

Trust ratings

Between-participants analyses. A mixed-design ANOVA was run on the trust index with feedback group as a between-participants factor and adviser's accuracy (low vs. high) and adviser's agreement rate (low vs. high) as within-participants factors. This analysis revealed significant main effects for both accuracy ($F(1, 44) = 8.36, p = .005, \eta^2_G = .06$) and agreement ($F(1, 44) = 22.52, p < .001, \eta^2_G = .1$) rates but not for feedback ($F < 1$). A significant interaction was found between feedback and accuracy ($F(1, 44) = 8.41, p = .005, \eta^2_G = .06$) but not between feedback and agreement rate ($F(1, 44) = 1.88, p = .17, \eta^2_G = .01$) or between agreement and accuracy ($F < 1$), nor a significant three-way interaction ($F < 1$). The results suggest that in their explicit ratings of trust, participants reported on average to prefer advice coming from accurate sources and sources that tended to agree with their own judgments. Importantly, however, the effect of accuracy was modulated by the presence of feedback, suggesting that the effect of accuracy on trust was larger in the Feedback condition than in the No-Feedback one (Figure 4.1).

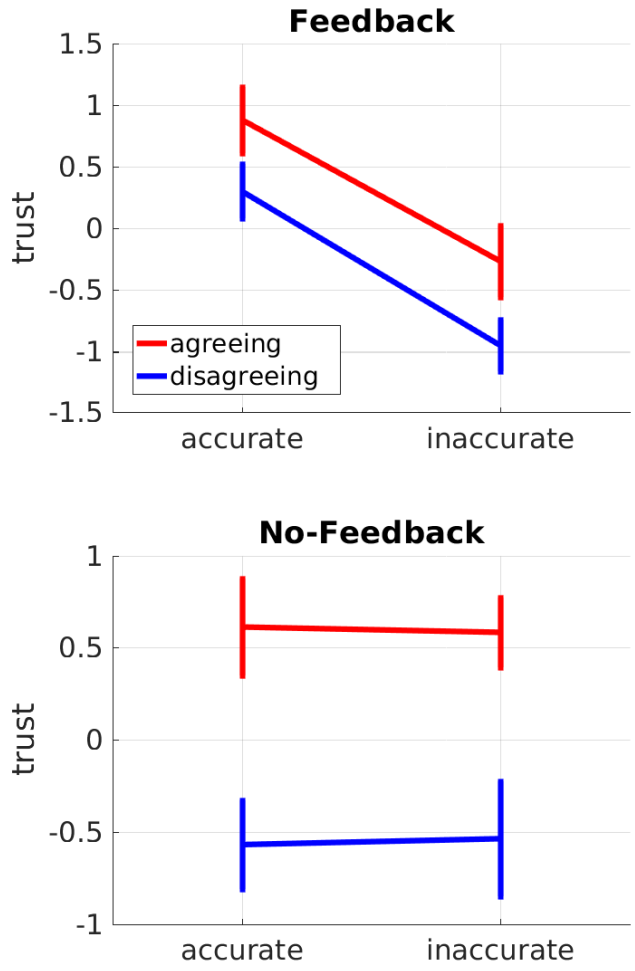


Figure 4.1: Average trust ratings in the two feedback groups, divided by adviser accuracy rate and agreement rate. Error bars represent s.e.m.

Feedback Study. A within-participants ANOVA on trust ratings was run on data from the Feedback group only to test for within condition patterns. Results show a significant effect of adviser's accuracy ($F(1, 22) = 21.36, p < .001, \eta^2_G = .20$) and agreement rate ($F(1, 22) = 5.62, p = .02, \eta^2_G = .06$), but no significant interaction ($F < 1$). These findings indicate that participants showed to explicitly trust more accurate over inaccurate advisers, and agreeing over disagreeing advisers. The effect of accuracy was however much stronger than the one observed for agreement as indicated by the generalised eta squared values (Bakeman, 2005).

No-Feedback Study. A within-participants 2x2 ANOVA was then run on the data from the No-Feedback group to test how the trust measure was affected by the experimental manipulation. Results show that agreement ($F(1, 22) = 18.91, p < .001, \eta^2_G = .18$) but not accuracy ($F < 1$) affected participants' explicit reliability judgments. No significant interaction was found between the two factors ($F < 1$). These findings suggest that when feedback was not directly available to estimate partners' reliability, participants expressed to trust more agreeing advisers over disagreeing ones, but not accurate advisers over inaccurate ones. This pattern contrasted with expressed trust reports from the Feedback group who reported to prefer advice characterised by both dimensions.

Influence

Between-participants analyses. A mixed-design ANOVA on the influence measure was run with feedback as a between-participants factor and adviser's accuracy (low vs. high) and agreement rate (low vs. high) as within-participants factors. This analysis revealed a significant main effects for adviser's accuracy ($F(1, 44) = 14.79, p < .001, \eta^2_G = .04$) and agreement rate ($F(1, 44) = 13.91, p < .001, \eta^2_G = .03$), and marginally so for feedback ($F(1, 44) = 3.19, p = .08, \eta^2_G = .04$). The marginal effect of feedback indicates that participants tended to be more influenced when feedback was not available (Figure 4.2). Importantly no significant interactions were found between accuracy and feedback ($F < 1$), between agreement and feedback ($F(1, 44) = 2.01, p = .16, \eta^2_G = .005$) and between accuracy and agreement rate ($F < 1$). No significant three-way interaction was found either ($F < 1$). The results show that participants were more influenced by accurate advisers and by advisers characterised by high agreement rates with their own judgments. Importantly, however, the effects did not interact reliably with feedback, suggesting that similar effects were found in the two condition groups independently of feedback availability. We then looked at individual pattern of results within each experimental group.

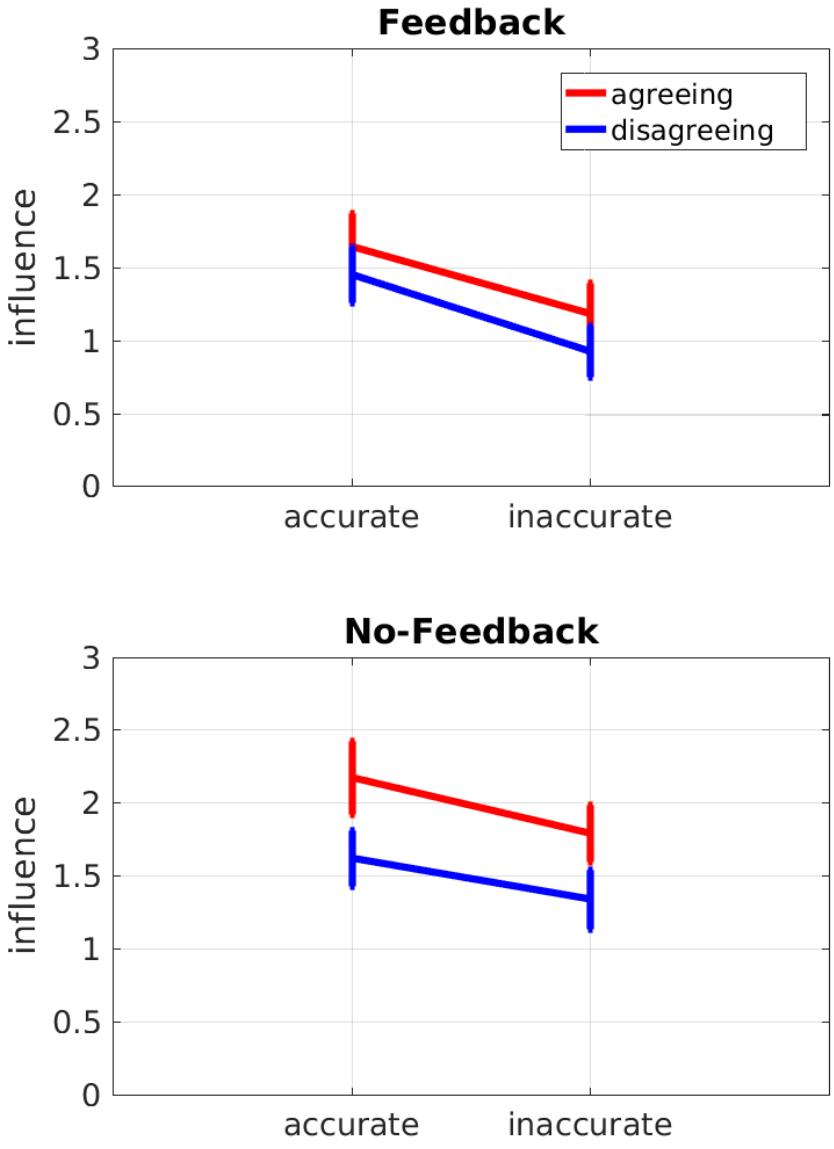


Figure 4.2: The figure above shows the influence that advice had on participants' opinions in the two feedback groups and divided by adviser accuracy rate and agreement rate. Error bars represent s.e.m.

Feedback Study. We then analysed the within-participants influence data from the Feedback group. Results showed significant effects for both adviser's accuracy ($F(1, 22) = 12.71, p = .001, \eta_G^2 = .06$) and agreement rate ($F(1, 22) = 5.81, p = .02, \eta_G^2 = .01$). No significant interaction was observed between the two factors ($F <$

1). As expected, when looking at opinion change, participants' opinions were more affected by accurate advisers over inaccurate ones. Interestingly, participants were also affected by advisers who tended to agree with their own judgment, although feedback indicated equivalent accuracy rates for pairs of advisers characterised by different agreement rates.

No-Feedback Study. When looking at the No-Feedback condition, a similar pattern emerged. A 2x2 repeated measures ANOVA on influence showed significant main effects for both adviser's agreement rates ($F(1, 22) = 8.60, p = .007, \eta^2_G = .06$) and accuracy ($F(1, 22) = 4.09, p = .05, \eta^2_G = .02$), although the effect size of agreement was the greater of the two. No significant interaction was observed ($F < 1$). Thus, in the No-Feedback group, the results suggest a dissociation between what people reported in the questionnaires, with higher trust reported for agreeing advisers regardless of accuracy, and people's performance in the task, where influence showed effects of both accuracy and agreement rate.

Model

The Bayesian model described in Chapter 3 was adapted to Experiment 2, to understand whether the *Consensus* and *Confidence* model variants behaved differently in scenarios where advice accuracy and advice agreement rate are dissociated. Each model variant was run on each participant's expressed confidence, experienced advice and, in the case of the *Accuracy* model, advice objective accuracy. The variants capture what those participants should represent about advisers if they were only sensitive to objective accuracy (*Accuracy* model), simple agreement (*Consensus* model) or agreement graded by confidence (*Confidence* model). In this experiment, advisers did not express a confidence judgment about their opinions. Thus, all model variants could be simplified by not taking into account advice confidence. The *Accuracy* model could be simplified using equations 3.2 and 3.3 to compute α and β for each adviser

instead of equations 3.4 and 3.5. Similarly, the *Consensus* model now computes α and β values for each adviser i separately as:

$$\alpha_i = \sum_{t=1}^{n} .5 + .5 * Agr_t \tag{4.1}$$

$$\beta_i = \sum_{t=1}^{n} .5 - .5 * Agr_t \tag{4.2}$$

where Agr_t is the partner's consensus ($Agr = \{-1, 1\}$) observed on encounter t. Finally, the simplified *Confidence* model computes α and β values as:

$$\alpha = \sum_{t=1}^{n} .5 + (p(corr) - .5) * Agr_t \tag{4.3}$$

$$\beta = \sum_{t=1}^{n} .5 - (p(corr) - .5) * Agr_t \tag{4.4}$$

where $p(corr)$ is the pre-advice confidence expressed in probability scale as described in equation 3.10.

Accuracy Model

Data from the two feedback groups were pooled together to test the model's predictions. Trial-by-trial pre-advice confidence and advice were input to each of the three model variants and resulting θ-values for each adviser were then compared (Figure 4.3). A 2x2 repeated measures ANOVA on *Accuracy* model's θ values as recorded on the last trial of the experiment was run with factors adviser's accuracy and agreement rate. Results show a strongly significant effect of accuracy rate ($F(1, 45) = 618.96, p < .001, \eta_G^2 = .78$) but no effect of agreement rate and no interaction between the two ($F < 1$). Not surprisingly, the results suggest that when provided with objective feedback on trial-by-trial performance, a simplified model of reliability estimation was able to dissociate advisers based on their accuracy but not based on their agreement profile.

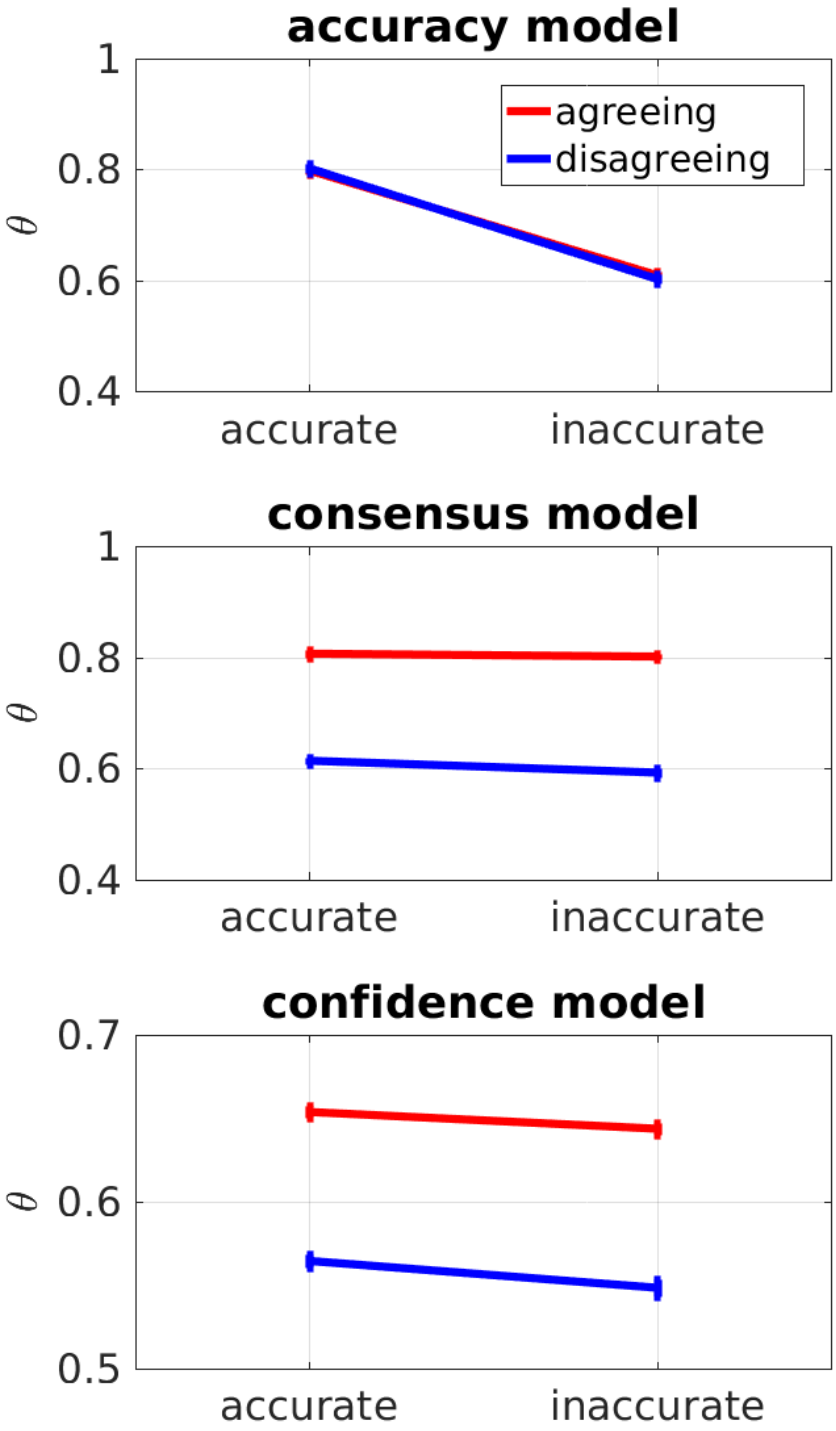


Figure 4.3: Final θ-values for each model variant as a function of adviser accuracy and agreement rate. Notice the different θ ranges on y-axis. Error bars represent s.e.m.

Consensus Model

A similar 2x2 repeated measures ANOVA was run on the *Consensus* model's θ-values with factors adviser's accuracy and agreement profile. Results showed a significant main effect on θ-value for adviser's agreement rate ($F(1, 45) = 847.41, p < .001, \eta^2_G = .82$) and only a marginally significant effect for accuracy rate ($F(1, 45) = 3.22, p = .07, eta^2_G = .02$), but no reliable interaction between the two ($F(1, 45) = 2.04, p = .15, \eta^2_G = .007$). The *Consensus* model trusted more agreeing advisers over accurate ones.

Confidence Model

Finally a *Confidence* model was run on the pooled data and the θ-values recorded on the last trial were passed through a 2x2 repeated measures ANOVA with factors adviser's accuracy and agreement profiles. Results show a significant effect of accuracy ($F(1, 45) = 8.85, p = .004, \eta^2_G = .05$) and agreement rate ($F(1, 45) = 434.76, p < .001, \eta^2_G = .72$) but no significant interaction between the two ($F < 1$). Results suggest that the *Confidence* model was sensitive to both advice dimensions although agreement rate still remained the strongest effect as indicated by the effect sizes.

Experiment discussion

To summarise, the results of Experiment 2, we orthogonally manipulated the agreement rate and accuracy of advice presented to participants. Behavioural results show that when participants had access to trial-by-trial feedback, both their explicit trust reports and their observed confidence changes indicated greater reliability estimates for accurate advisers and advisers who tended to agree with their own judgments. On the contrary, when trial-level feedback was removed, a dissociation seemed to emerge between explicit trust reports and observed confidence change. In particular, when directly asked, participants reported more trust in advisers characterised by high agreement rates, and showed little trust for advisers characterised by high

accuracy over those with lower accuracy. When looking at participant's task performance, however, the influence of an adviser depended on their accuracy as well as their agreement rate.

A Bayesian model which tracks advice reliability showed different patterns of adviser preference depending on what pieces of information were available to form a trust judgment. An *Accuracy* model showed a strong preference for accurate advisers while ignoring differences based on agreement rates. A *Consensus* model on the contrary showed a strong preference for advisers who tended to agree with participant's choices compared to advisers characterised by lower agreement rates. This model only marginally distinguished accurate from inaccurate partners. Finally, a *Confidence* model distinguished both dimensions of advice, namely agreement and accuracy, by scaling agreement with internal metacognitive signals.

Unsurprisingly, adviser accuracy affected reported trust and influence measure in the Feedback group (Behrens et al., 2008). The result that in this condition participants showed also an agreement effect is, however, more surprising. The finding seems to suggest that agreement is perceived as a strong indicator of the reliability of advice, independent of the actual underlying accuracy. A possible explanation for this effect is that participants show a confirmation bias (Nickerson, 1998) by which confirming evidence (in this case an agreement trial) is weighted more than disproving evidence (a disagreement trial). The *Accuracy* model created to describe trust formation in cases when feedback is provided by the experimenter is thus not enough to describe participants' behaviour as it fails to show a mediating effect of agreement. A combination of accuracy and internal confidence could perhaps explain the actual mechanism giving rise to participants' actual behaviour.

When feedback was unavailable, participants' observed behaviour (i.e., influence measure) and expressed trust in their partners seemed to dissociate. Although

both measures were sensitive to agreement rate, with more frequently agreeing advisers rated as more reliable and their advice having greater influence, only confidence changes were sensitive to differences in advice accuracy. One possibility is that the preprocessing performed on original questionnaire ratings through PCA to aggregate different questions together caused a loss of the information originally communicated by participants. Looking at the accuracy question only would verify whether participants were indeed sensitive to advice accuracy when directly prompted. A two-way ANOVA on average accuracy ratings showed a significant effect of agreement ($F(1, 22) = 20.84, p < .001, \eta^2_G = .21$), but no effect of accuracy (($F < 1$) and no interaction ($F < 1$), rejecting the loss of information hypothesis.

An alternative explanation is that explicit beliefs about partners' reliability were indeed dissociated from actual behaviour. This might be due to the fact that the two advice dimensions were differently weighted: when looking at the effect sizes of advice agreement and accuracy it is clear that agreement represents the strongest effect. Although this might not have had an effect on trial-by-trial confidence change behaviour, it might have resulted in a halo-dumping effect (Clark & Lawless, 1994) whereby, when prompted to discriminate among advisers, participants used only the most accessible dimension.

The fact that participants' confidence changes were sensitive to the advice accuracy even in the absence of objective feedback suggests that some proximal cues to accuracy were available to them. It is unlikely that this cue was the agreement of the advice, as this dimension was orthogonal to accuracy by design. A more plausible hypothesis is that participants accumulated over time some internal uncertainty quantity (e.g. confidence) for each adviser separately. As shown by the *Confidence* model, this strategy makes it possible for an agent to discriminate between equally agreeing partners who have different underlying accuracy. Table 4.1 helps to understand this finding. Consider for example the two agreeing advisers. The table shows

that the accurate adviser tends to agree more often than the inaccurate adviser when the participant is objectively correct (6.5 times out of 7 against 5.5 times out of 7) and less often when the participant is objectively wrong (1.5 thirds against 2.5 thirds). Trials when participants' initial judgment is correct are expected to be associated with greater confidence ratings, due to the fact that accuracy and confidence typically correlate (Fleming & Lau, 2014; Henmon, 1911; Koriat, 2012b). A strategy of reliability estimation relying on confidence would thus exploit this covariation to detect differences in accuracy, notwithstanding equal agreement rates.

Findings of Experiment 2 demonstrated that agreement is a crucial cue to reliability, such that its effects are seen even when objective feedback is provided. Advice accuracy, independent of agreement, also affected trust and influence when feedback was provided (Behrens et al., 2008). When feedback was unavailable, on the contrary, participants reported to trust more agreeing advisers. The influence that advice had on their confidence updates was however affected by both advice accuracy as well as agreement rate.

The fact that agreement was such a strong predictor of advice impact on participants' opinions made it difficult to isolate the effect of internal confidence on advice evaluation. Experiment 3 was designed to prevent this issue and to match all advisers in terms of agreement and accuracy rates while at the same time varying their pattern of bias with the participant.

Experiment 3

Experiment 3 aimed to provide stronger evidence that subjective confidence is implicated in the formation of judgments about advice reliability. In this experiment, instead of manipulating the probability of agreement conditional on participant's accuracy (as it was the case in the previous experiment), I manipulate the probability of agreement conditional on the participant's initial confidence. This creates

the possibility of advisers who differ in terms of agreement patterns relatively to the participant's opinion but who are otherwise equal in terms of average accuracy and agreement rate. Three advisers were created by making advisers' judgments more or less biased towards the participant's confidence judgments in trials when the participant's initial judgment is correct: (1) an unbiased adviser who tends to agree with the participant's choice about 70% of the time, independent of participant's confidence; (2) a bias-sharing adviser who tends to agree with the participant's initial choice particularly when the participant is confident; (3) an anti-bias adviser who tends to agree with the participant particularly when the participant is unsure. The manipulation is a "special case" of the general idea of bias sharing among observers that was described above. It aims to study real-life situations where observers' errors are correlated due to the presence of shared biases (as opposed to errors due to random noise, as in traditional Wisdom-of-Crowds effects (Armstrong, 2001; Rauhut & Lorenz, 2011)). One added feature, not present in Experiment 2, is that advisers cannot be discriminated based on agreement rate or accuracy.

Methods

Participants

50 participants were tested and divided in the two experimental groups. Due to unforeseen circumstances 24 participants ended in the No-Feedback group and 26 in the Feedback group. Participants were recruited through local advertisement and the Oxford University online recruitment platform. Prior the start of the experiment all participants signed a consent form approved by the local ethical committee. Volunteers were compensated for their time via monetary compensation (10 £/hour) or university credits.

Paradigm

The experiment consisted of 12 experimental blocks of 35 trials each (5 null trials). The perceptual task was the same as for Experiments 1 and 2. Display parame-

ters, input modality and stages within a trial remained unaltered. The experimental manipulation required the experimenter to have access to the exact confidence distribution of the participant. Thus the confidence scale was changed to a semi-continuous scale including 50 points per interval. Three advisers were designed for Experiment 3. All participants' stimuli were titrated to converge to a 70.7% accuracy level using a 2-down 1-up procedure. Two blocks of 25 trials served as the practice blocks and used a fourth "practice" adviser.

Manipulation

The advice profiles of the three advisers were manipulated so that all advisers converged to the same level of accuracy and agreement rate with the participant's judgments. The pattern of agreement was manipulated, however, such that the three different advisers agreed with the participant's initial judgments on trials characterised by different initial confidences. To this end, the probability of agreement conditional on the participant's pre-advice accuracy and confidence was manipulated across advisers as described in Table 4.2 and illustrated in Figure 4.4. The distribution of the participant's pre-advice confidence judgments was recorded from the first two practice blocks and used as a reference distribution. The reference distribution was divided into three confidence bins: the low, middle and high confidence bin including 30%, 40% and 30% of the trials respectively. On trials where the participant's pre-advice judgment was incorrect all advisers had a probability of agreement of 30% independent of the participant's confidence level. On trials where the participant's judgment was correct on the contrary the three advisers had different agreement patterns. A unbiased adviser had a probability of agreement of 70% independent of the participant's confidence. A bias-sharing adviser was defined as agreeing around 80% of the time when the participant was highly confident and 60% of the time when s/he was uncertain. An anti-bias adviser was designed so to agree 60% of the time when the participant was highly confident and 80% of the time when s/he was uncertain.

Importantly, however all advisers had an equal chance of agreement when the participant's pre-advice confidence fell in the middle bin (70%). This ensured that all advisers were matched in terms of average agreement rate ($0.7*0.7+0.3*0.3 \sim 0.58$) and accuracy rate ($0.7 * 0.7 + 0.3 * 0.7 \sim 0.7$) across all trials. By limiting analyses to trials within the intermediate confidence bin, we could compare advisers on trials where they were all virtually identical. This allowed us to avoid confounds arising from comparing different trials together. For example, asymmetric confidence changes in agreement and disagreement and larger advice influence in uncertain trials might contribute to adviser-specific distortions in the influence measure, if all trials were pooled together. The confidence reference distribution was updated on every block to take into account possible shifts of confidence during the course of the experiment, to reflect the distribution of confidence judgments provided during the previous two blocks.

	Advisers		
	Bias-sharing	Unbiased	Anti-bias
$p(Agree\|Correct^s, Confidence^s_{low})$	60%	70%	80%
$p(Agree\|Correct^s, Confidence^s_{mid})$	70%	70%	70%
$p(Agree\|Correct^s, Confidence^s_{high})$	80%	70%	60%
$p(Agree\|Incorrect^s)$	30%	30%	30%
$AUC(IG_E)$	14.63	14.74	15.78

Table 4.2: Agreement probability of different advisers is manipulated conditional on the participant's pre-advice confidence and accuracy. This manipulation allowed to create three different advisers who were matched in terms of agreement rate and accuracy, but who differed in terms of information as quantified by the area under the expected information gain curve.

Although the average accuracy and agreement rate of the advisers were the same, the information gained by participants receiving advice was not. In the present experiment, computing adviser's information gain from prior knowledge of their conditional probability was not possible as this was dependent on the subjective confidence distribution. However, according to a Bayesian normative model, greater changes along

the confidence scale should take place when subjective initial confidence is low, and lower shifts should be observed when initial confidence is high. The anti-bias adviser is in this respect in the good position of giving supporting evidence (agreement) when it is needed the most (low subjective initial confidence). The bias-sharing adviser on the contrary tends to agree when their judgments are less effective (high subjective initial confidence). This intuition was confirmed using a numerical simulation based on an ideal Bayesian observer performing the task, with a Gaussian distribution of confidence centred on 25 and with a standard deviation of 10. For each initial confidence judgment, the information gained from observing agreement or disagreement was computed for each adviser as the difference between posterior confidence and prior confidence. An expected information gain was computed by multiplying the information gain so obtained by the normalisation term in the Bayes formula. This produced a curve of expected information gains after agreeing or disagreeing with each adviser (Figure 4.4, lower panels). The average area under the expected information gain curve was 14.74 for the unbiased adviser, 14.63 for the bias-sharing adviser and 15.78 for the anti-bias adviser. This procedure thus quantified and confirmed the intuition that the anti-bias adviser provided the most informative advice.

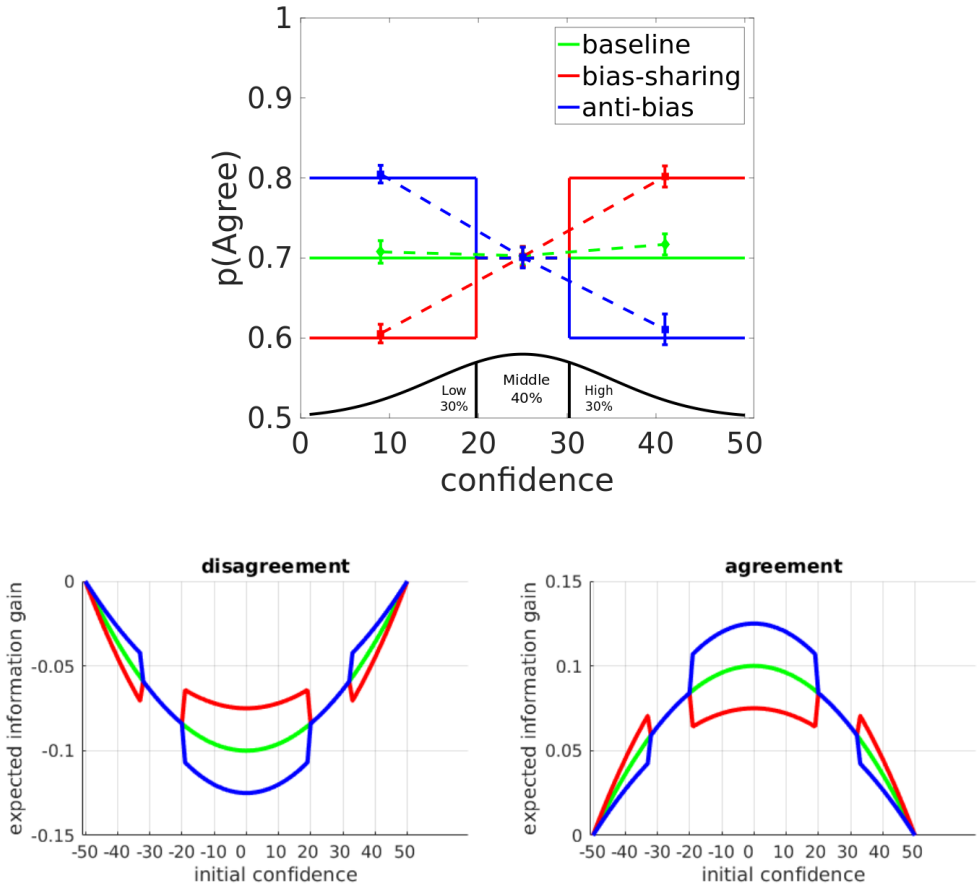


Figure 4.4: Experimental manipulation of Experiment 3. The probability of adviser's agreement conditional on participant's accuracy and confidence was manipulated so that the three advisers differed in their pattern of agreement despite being equal on average agreement rate and accuracy rate. Top panel: continuous lines represent expected agreement rates, dashed lines represent empirical data pooled across the two feedback groups. Lower panels: expected information gain in agreement and disagreement depending on the initial level of confidence. Expected information gain is defined as the difference between posterior and prior confidence scaled by probability of the observed event (agree vs disagree). The area under the curves quantifies the average information of the advice.

Following Experiment 2, we expect different patterns of results to emerge from feedback and feedback-free conditions. Specifically, based on the *Confidence* model, we predict that when feedback is not available trust reports and influence measure will favour the bias-sharing adviser over the other advisers. This follows from the fact that both participants and the *Confidence* model will experience more high-confidence agreements with the bias-sharing adviser compared to the non-bias shar-

ing one. This will happen even if advisers are matched in terms of accuracy and agreement, because participants and the *Confidence* model will learn the association between internal metacognitive states (like confidence) with external contingencies (like agreements/disagreements). Neither the *Consensus* or the *Accuracy* models are expected to distinguish among advisers.

When objective feedback is provided, we might expect that participants will not distinguish between advisers given that they are all equally reliable ($\sim 70\%$ accuracy).

Measures of interest

The same measures of interest adopted in the first two experiment were adopted here to compare how different advisers' profiles were perceived by the participants. First, an explicit trust measure was created by preprocessing questionnaire responses through normalised baseline correction and PCA. The first PCA component was used as a unitary measure of explicit trust. The same four questions asked in Experiment 2 were used here. Questionnaires were presented at the beginning of the experiment (baseline questionnaire) and every other experimental block. PCA loadings for the Feedback group were: 0.53 (Q1), 0.42 (Q2), 0.53 (Q3), 0.50 (Q4); and for the No-Feedback group: 0.48 (Q1), 0.47 (Q2), 0.53 (Q3), 0.50 (Q4).

Second, an implicit trust measure was obtained that quantified the impact of each adviser's opinion on the participant's judgments. As in the previous experiments, this influence measure was obtained by the difference between average confidence change in agreement and average confidence change in disagreement when comparing participants' pre- vs. post-advice judgments on each trial.

Exclusion criteria

An exclusion criterion based on staircase convergence was set so to exclude all participants who appeared to have a random guessing strategy. Application of this criterion

resulted in the exclusion of one participant from the Feedback group and one participant from the No-Feedback group, leaving a total of 23 and 25 participants in these groups, respectively. Average difficulty parameter d was 9.98 ± 2.82 (pooled data).

Results

The following analyses were performed to show whether an effect of adviser type on the measures of explicit trust and implicit opinion change could be detected. Furthermore we were interested in knowing whether this effect was modulated by the presence of feedback. Feedback acted as a between-participants factor while adviser type acted as a within-participants factor in the following mixed-design ANOVAs. For within-participants variables, degrees of freedom were corrected for violations of sphericity according to the Greenhouse-Geisser procedure, with epsilon values reported as appropriate.

Trust ratings

Between-participants analyses. A mixed-design ANOVA on trust measures showed no significant effect of adviser type ($F(2, 92) = 1.66, p = .19, \eta^2_G = .03, \epsilon = 0.99$) or feedback ($F < 1$), but a significant interaction between adviser and feedback ($F(1, 92) = 6.64, p = .002, \eta^2_G = .12, \epsilon = 0.99$). The results suggest that although different advisers were similarly trusted on average, the effect was modulated by the presence of feedback. Figure 4.5 shows how the effect of adviser was influenced by the presence of feedback.

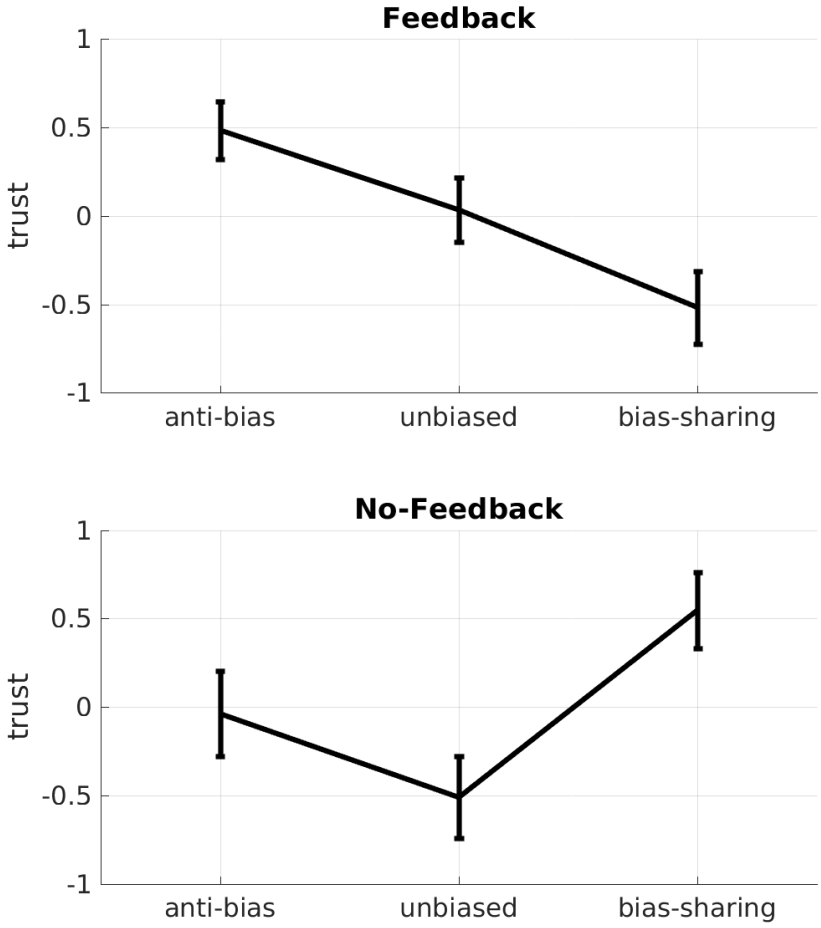


Figure 4.5: The effect of adviser type on explicitly reported trust, separately for the Feedback and No-Feedback condition. Error bars represent s.e.m.

Feedback Study. Within groups analyses were carried out to understand differences among advisers present in each experimental group. As Figure 4.5 shows, when provided with feedback, participants seemed to show an effect of adviser type, suggesting that even with feedback they consistently trusted some advisers more than others. A one-way repeated measures ANOVA showed a significant effect of adviser type ($F(2, 48) = 4.90, p = .01, \eta^2_G = .16$). Pairwise comparisons showed that the bias-sharing adviser was significantly trusted less than the anti-bias adviser ($t(24) = 3.09, p = .004, d = 1.07, B_{H[0,1]} = 47.34$) when participants were prompted to

report their reliability estimates explicitly. No differences were however found between the anti-bias and the unbiased advisers ($t(24) = 1.60, p = .12, d = 0.51, B_{H[0,1]} = 1.67$) nor between the unbiased and the bias-sharing advisers ($t(24) = 1.56, p = .13, d = 0.56, B_{H[0,1]} = 1.83$).

No-Feedback Study. Similarly a repeated measures one-way ANOVA on trust measure revealed a significant difference among advisers ($F(2, 44) = 3.56, p = .03, \eta^2_G = .13$) confirming that when prompted to express their explicit reliability estimates participants tended to consistently trust some advisers more than others. Planned comparisons revealed that the bias-sharing adviser was trusted marginally more than the anti-bias adviser ($t(22) = 1.48, p = .07, d = 0.53$, one-tail, $B_{H[0,1]} = 1.74$) and significantly more than the unbiased adviser ($t(22) = 2.81, p = .005, d = 0.98$, one-tail, $B_{H[0,1]} = 22.42$). The anti-bias and the unbiased advisers did not significantly differ from each other ($p > .1, d = 0.41, B_{H[0,1]} = 1.13$).

Influence

Between-participants analyses. A mixed-design ANOVA with feedback as a between-participants factor and adviser type as a within-participant factor showed no significant effect for feedback ($F(1, 46) = 1.36, p = .24, \eta^2_G = .01$) nor adviser ($F(2, 92) = 1.12, p = .33, \eta^2_G = .009, \epsilon = 0.77$), but a significant interaction between the two ($F(2, 92) = 4.80, p = .01, \eta^2_G = .03, \epsilon = 0.77$), suggesting that the presence of feedback impacted upon the influence that the advice had on participant's opinions. Figure 4.6 shows the effect of feedback presence and adviser type on participants' opinion changes as quantified by the influence measure.

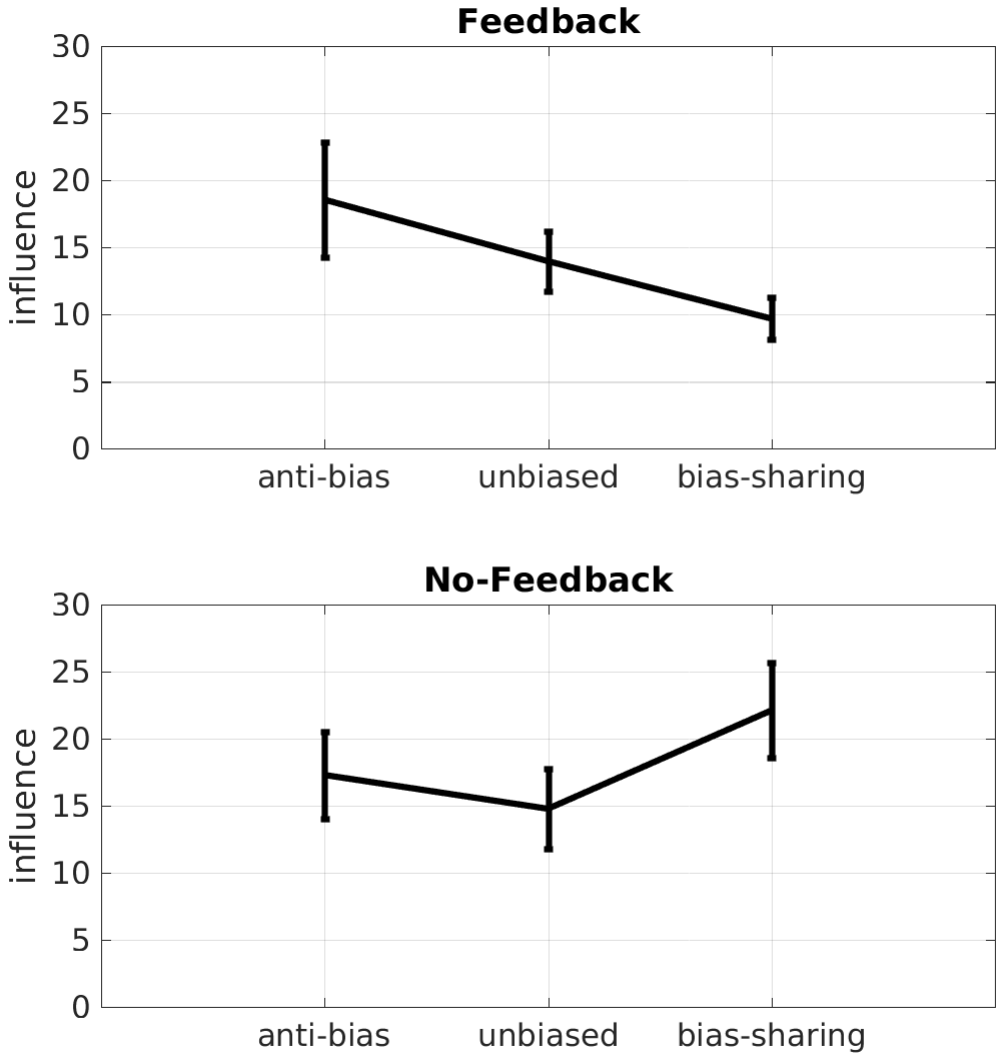


Figure 4.6: The figure shows the effect of adviser type on influence measure, divided by the two feedback groups. Error bars represent s.e.m.

Feedback Study. When looking at confidence change behaviour in the Feedback condition a one-way repeated measures ANOVA showed a marginally significant effect of adviser type ($F(2, 48) = 2.88, p = .06, \eta^2_G = .05$). A two-tails t-test showed that the anti-bias adviser was more influential than the bias-sharing one ($t(24) = 1.98, p = .05, d = 0.54, B_{H[0,5]} = 3.71$), that the bias-sharing adviser was significantly less influential than the unbiased adviser ($t(24) = 2.26, p = .03, d = 0.44, B_{H[0,5]} = 6.56$)

and no reliable difference was apparent between the anti-bias and unbiased advisers ($t(24) = 1.10, p = .28, d = 0.26, B_{H[0,5]} = 1.47$). Although the numerical difference was higher between the bias-sharing and anti-bias advisers compared to the bias-sharing and unbiased advisers, the former had a larger p-value, probably due to the high variability of the influence measure for the anti-bias adviser.

No-Feedback Study. A one-way ANOVA on the influence measure revealed a significant difference among advisers ($F(2, 44) = 3.25, p = .04, \eta^2_G = .03$). Planned comparisons showed that the bias-sharing adviser was significantly more influential than the unbiased adviser ($t(22) = 2.63, p = .007, d = 0.46$, one-tail, $B_{H[0,5]} = 13.62$) and marginally so than the anti-bias adviser ($t(22) = 1.46, p = .07, d = 0.29$, one-tail, $B_{H[0,5]} = 2.06$). No difference was found significant between the unbiased and the anti-bias advisers ($p > .1, d = 0.16, B_{H[0,5]} = 1.07$). Thus, these results seem to suggest that the bias-sharing adviser was more influential than the other two when trial-by-trial feedback was not available.

Thus, the presence of trial-by-trial feedback can change the perception of advice reliability. In particular, the bias-sharing adviser is trusted less when feedback is provided but more when feedback is absent. This was apparent both when looking at reported trust and influence measures.

Model

The following analyses were performed to understand what patterns of trust should be expected by a Bayesian model estimating the reliability of each adviser based on predetermined pieces of information. The three variants of the model each represent a simple mechanism by which an observer could arrive to a trust judgment θ about different sources of advice using objective feedback (*Accuracy* model), agreement rates (*Consensus* model) and metacognitive signals (*Confidence* model) respectively. Each

model variant was fitted to each participant's expressed confidence, experienced advice and, in the case of the *Accuracy* model, advice objective accuracy. The variants model what those participants should represent about advisers if they were only sensitive to objective accuracy (*Accuracy* model), simple agreement (*Consensus* model) or agreement graded by confidence (*Confidence* model). Differences in θ-values observed among the three variants thus show how same observed data can give rise to different trust judgments depending on what type of information an observer has access to. Data from the two Feedback conditions was pooled together, because model variants' estimation of θ is unaffected by the presence of feedback. Figure 4.7 shows the pattern of results that the three model variants produce on the pooled data from both feedback conditions. A reinforcement learning model is also described in Appendix A and applied to data from Experiment 3. It shows that the confidence-based strategy to infer advice reliability can also be applied to classical models of learning (e.g., delta-rule), with similar conclusions reached.

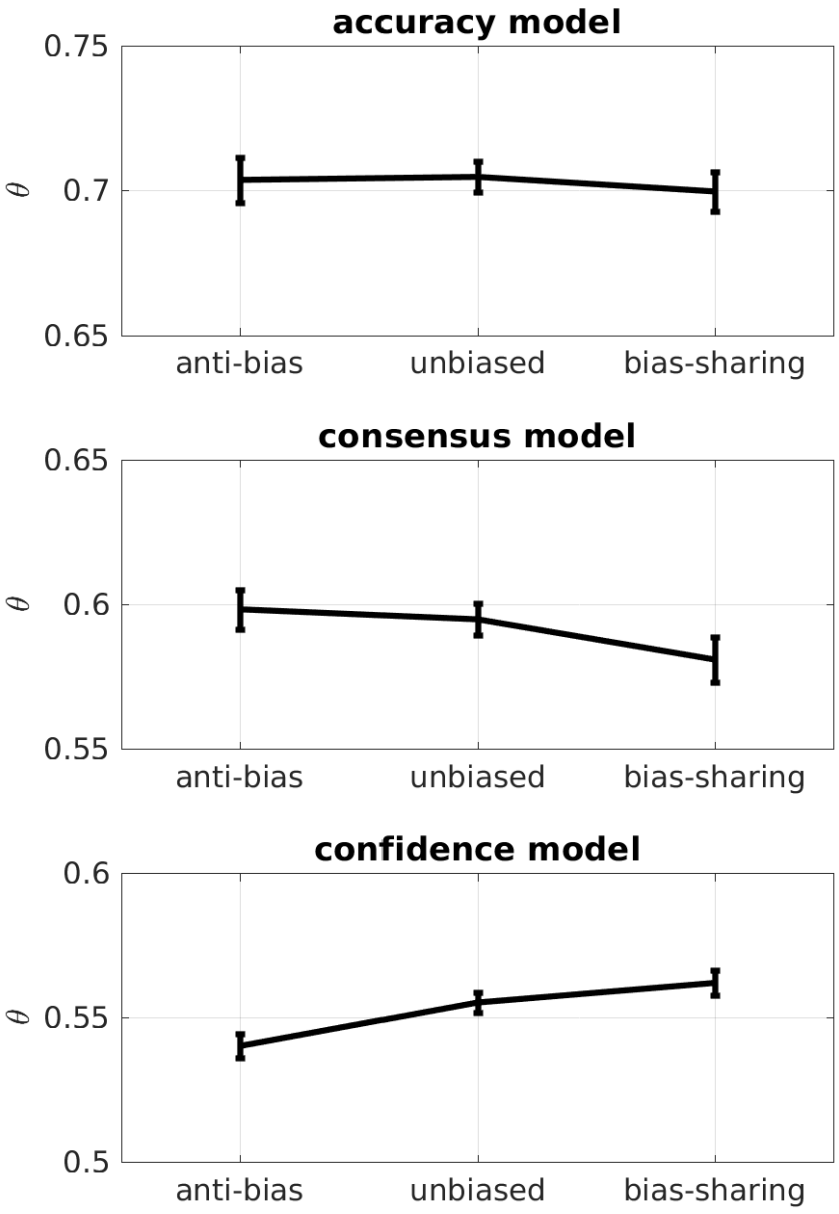


Figure 4.7: Average θ-values, representing the model's trust in different advisers, differ depending on what pieces of information the model has access to. Error bars represent s.e.m.

Accuracy Model.

The *Accuracy* variant is targeted to reproduce the Feedback condition, where participants had access to trial-by-trial feedback about their own and advice accuracy. A one-way repeated measures ANOVA showed no significant effect of adviser ($F < 1$) on last trial θ-values estimated by the *Accuracy* model, and no difference between the bias-sharing and the anti-bias adviser ($t < 1$), suggesting that the model could not discriminate among advisers even if it had access to objective trial-level feedback. This makes sense as all advisers were equal in terms of accuracy rates.

Consensus Model.

The *Consensus* variant is targeted to reproduce the No-Feedback condition if participants did not have access to metacognitive information to form advice reliability judgments. A one-way repeated measures ANOVA on last trial θ-values as estimated by the *Consensus* model showed no significant effect of adviser ($F(2, 94) = 1.70, p = .18, \eta^2_G = .02$). No significant difference was found between the bias-sharing adviser and the anti-bias adviser ($t(47) = -1.55, p = .12, d = 0.34, B_{H[0,0.05]} = 0.09$), suggesting that this model variant could not discriminate between advisers purely based on agreement rates. The presence of a numerical trend is likely due to chance variations in exact agreement counts for each participant, as this was not fixed *a priori*. Regardless, the direction of the trend is opposite to that observed in the No-Feedback condition of the experiment (i.e., less trust for bias-sharing adviser).

Confidence Model.

The *Confidence* variant is targeted to reproduce the No-Feedback condition if participants had access to metacognitive information to form advice reliability judgments. A one-way repeated measures ANOVA on last trial θ-values as estimated by the *Confidence* model showed a significant effect of adviser ($F(2, 94) = 7.95, p < .001, \eta^2_G =$

.10). A significant difference was found between the bias-sharing adviser and the anti-bias adviser ($t(47) = 3.54, p = .001, d = .74, B_{[0,.05]} = 114.64$) and between unbiased and anti-bias advisers ($t(47) = 2.99, p = .004, d = .57, B_{[0,.05]} = 18.16$), but not between unbiased and bias-sharing ($t(47) = 1.21, p = .22, d = .25, B_{[0,.05]} = 0.41$). These findings suggest that by accessing metacognitive signals (as provided by participants' confidence ratings) the model was able to discriminate among different advisers. Interestingly, the direction of the effects (i.e., greater trust for the bias-sharing adviser) was similar to what observed in the No-Feedback condition.

Experiment discussion

Experiment 3 showed that when advisers are matched in terms of accuracy and agreement with the participant's choices, differences in trust reports and measures of influence can still be observed. Importantly these differences are modulated by the presence of objective feedback. In Experiment 3 the presence of feedback partly reversed the pattern of trust ratings and influence measure that was observed when no feedback was available to participants. When objective feedback was available, people seemed to trust more and be more influenced by advisers who tended to disagree more often with them when they were sure and agree with them when they were unsure. At the same time they seemed to discount both in terms of trust and influence the advice received by an adviser who tended to agree when they were themselves sure and disagree more often when they were unsure. On the contrary, when objective feedback was removed, this pattern partly reversed and people seemed to trust and be more influenced by the adviser who shared their same bias. They also seemed to discount the adviser who did not share their same bias and even more so the unbiased adviser.

Participants in the Feedback group showed an effect of adviser type even though they received feedback indicating that all advisers were equal in terms of both agreement rate and accuracy. This should hint to the fact that participants were sensitive

to the informational value of the advice and not only to the advice reliability *per se*. Advice information value was here defined as the information gain that is observed in a Bayesian observer after receiving the advice (Figure 4.4). It quantifies the intuition that larger confidence updates follow more informative events. Participants trusted more and were more influenced by the anti-bias adviser who was in turn the most informative one (i.e., producing larger nominal updates). On the contrary they discounted advice coming from a bias-sharing adviser, which is consistent with this adviser's poor informational value (i.e., low IG_e).

In the No-Feedback condition, although the experimental manipulation placed the unbiased adviser in between the bias-sharing and anti-bias advisers, people perceived the unbiased adviser as the least trustworthy and they were less likely to listen to her advice. This fact is even more puzzling when considering that the bias-sharing adviser and the unbiased adviser were very similar in terms of advice information gain (and both significantly less informative than the anti-bias adviser). However these two profiles ended up being at the two opposites of the trust spectrum when looking at the behavioural data. Even the *Confidence* model, which ranked advisers opposite to their information gains, still ranked the unbiased adviser in between the other two.

A possible explanation for the observed pattern is that participants did not simply evaluate the correctness of the advice when estimating advice reliability, but instead compared the variability of the advice across and within confidence bins. Weiss and Shanteau (2003) suggested that when objective feedback cannot be easily retrieved from the environment, the best way to estimate expertise (and thus judgment accuracy) is by dividing the decision maker's judgments variance across different classes with the variance of judgments within each class. The authors describe a measure - the Cochran-Weiss-Shanteau (CWS) index - that quantifies the intuition that experts (e.g., doctors) should show high variance in categorising items belonging to different classes (e.g. making diversified diagnoses to different patients) but low variance in

categorising items belonging to the same class (e.g. making same diagnoses to individuals belonging to a homogeneous cohort of patients). The authors show that the CWS index of expertise reliably tracks the expertise of different decision-makers across several domains.

When applied to the current experiment, participants might implicitly treat trials falling within different confidence bins or percentiles as different classes of judgments. The CWS index in this scenario would thus be zero for the unbiased adviser as this adviser exhibits no variance in the probability of agreement across different confidence classes. On the contrary the other two advisers would get a non-zero CWS value as they both show some variability in their agreement with the participant across different confidence classes. However this explanation alone cannot fully account for the results found in the experiment because bias-sharing and anti-bias advisers seemed, consistently across the measures considered, trusted differently. It is thus likely that participants must have engaged to some extent in a combination of strategies to produce an estimation of the reliability of a source. The combination of multiple strategies might have in turn caused the complexity observed in the pattern emerged.

General Discussion

Much of the work in psychology in the last decades has focused on heuristics and biases affecting human judgments (Gigerenzer, 2008; Tversky & Kahneman, 1974). Given its limited capacities, the brain must often adapt to use approximations or "shortcuts" to find good-enough solutions to otherwise intractable problems. Heuristics are simple solutions that work most of the time but lead to characteristic errors (Tversky & Kahneman, 1983). In the three experiments presented so far, the intractable problem that our participants face is estimating reliability without an objective standard against which to judge performance and advice. A viable solution is to use an internal sense of reliability. This solution works well when agreement correlates with accuracy, as in the case where initial judgment and advice are independent (Experiment

1). Here trust and influence patterns were little affected by the presence or absence of an objective standard. But the process goes astray where this independence is broken, as in our studies where advice was contingent on participants' judgment (Experiments 2 and 3). Adverse consequences showed up in terms of distinct patterns of trust and influence with and without an objective standard.

In the current experiments, participants valued the informativeness of the advice when objective feedback was immediately available (e.g., as suggested by their trust in the anti-bias adviser). People are able to use outcome signals (like rewards and feedback) to track cues reliability according to associative learning mechanisms (Behrens et al., 2008; Sutton & Barto, 1998). When feedback is removed however, they tend to paradoxically rely on the least informative advice (e.g., trusting the bias-sharing adviser). An intriguing hypothesis is that when feedback is removed participants are only left with their own initial probabilistic estimate. Any attempt to use this estimate to attribute feedback to their advisers is cursed by the error characterising their own original judgment. As an example, imagine you are sure that you are drinking a very expensive wine at a blind wine tasting event. You will likely judge anybody who says otherwise as incompetent in wine matters. In the (not so unlikely) event that you are mistaken, you could keep thinking that a potentially brilliant sommelier is only a novice, with potentially bad consequences for you.

Speculatively, the fact that when lacking objective feedback people prefer advisers who share their same bias (hence who are more likely to agree) might be at the heart of echo-chamber effects often described in network science (Jasny et al., 2015; Sunstein, 2001), whereby individuals with similar characteristics tend to form clusters that are impenetrable to external information. When exposed to competing opinions on social media, like-minded people belonging to the same cluster tend to prefer, consume and share within-cluster information more than between-cluster information (Bessi, 2016; Del Vicario et al., 2016). One hypothesis is that situations

where feedback is difficult to obtain (e.g., socio-political matters) will foster formation of chambers with polarised views (Bessi et al., 2016). This follows from the fact that, once feedback is removed, people will judge reliability based on their own convictions (i.e., bias). The presence of bi-directional information channels between agents (as opposed to judge-adviser systems) will attract people who share the same bias together, increasing the polarisation of their views (i.e., confidence) as well as producing phenomena of *groupthink* (Janis, 1972; Turner & Pratkanis, 1998). To test this hypothesis, agent-based modelling could be used to show how the manipulation of feedback availability influences network structure and information flow. This approach could link the cognitive mechanisms described here to larger-scale dynamics. Agent-based models are useful tools to study phenomena of emergence from simple interacting parts. Manipulating agents' access to metacognitive signals could shed light on how the cognitive structure of agents can affect the structure of a network (Epstein, 2013). Additional circularity might be introduced if agents are given choice over the information they sample, for example if reliability estimates were used in selecting incoming information (Aiello et al., 2012; Denrell, 2005).

Irrespective of the behavioural results observed, the modelling analysis shows that a model endowed with access to its own metacognitive signals can discriminate between advisers while models who have only access to objective feedback or agreement rate fail to do so. The result suggests that metacognitive signals like confidence judgments can be not only useful for internal uncertainty monitoring (Yeung & Summerfield, 2012), cognitive control (Botvinick et al., 2001) and social coordination (Bahrami et al., 2010), but also to evaluate the reliability of external information sources. The same conclusion was also reached by analysis of the behaviour of a reinforcement learning model, which was appropriately modified to estimate advice reliability in the three scenarios (feedback, no-feedback without metacognitive access, no-feedback with metacognitive access) and applied to Experiment 3 data (Appendix

A). The present findings demonstrate the potential bidirectionality of this inference process: confidence is not only the end-product of the information flow that goes from the external stimulus to a perceptual inference, but it can feed back to help make inferences about external events. Confidence represents a probabilistic estimation that once formed can be used to infer the state of variables that did not directly generate it.

The current study differs from existing work in group decision making (Bahrami et al., 2010; Sorkin et al., 2001). While previous works have tried to show how confidence judgments from two individuals can be used to combine estimates of an external stimulus, the current work shows that internal confidence about the state of the external stimulus can be used also to estimate the reliability of a social partner over and beyond what can be inferred from objective external cues. It used a traditional Judge-Adviser System paradigm (Bonaccio & Dalal, 2006; Sniezek & Buckley, 1989; Yaniv & Kleinberger, 2000), but it tries to go beyond it by finely controlling for important variables such as the information carried by the advice, the subjective perceived difficulty of the task and the presence of objective feedback. The study tried to bring together the Judge-Adviser System tradition with current cognitive theories of confidence (Pouget et al., 2016).

Conclusions

Experiment 3 concludes my first series of experiments. These studies aimed at showing that in a social context confidence is a valuable attribute of someone's judgment. It helps others discriminate when you are more likely to be correct (and thus value your contribution) but also helps a decision-maker to make consistent judgments irrespective of feedback availability. However they also show how these cognitive mechanisms can backfire when judgments from different observers are not independent.

The next chapter will describe a new line of work using an original design. The Judge-Adviser System paradigm used so far is a useful tool to isolate specific effects underlying trust formation and opinion change. However results are often not easy to generalise given the rigid staged structure of the interaction. The paradigm described in the next chapter was designed with these limitations in mind. It aims at understanding how judge-adviser systems relate to ecological real-life-like interactions, characterised by recursive information flow and non-linear dynamics.

5

Beyond Picture Partners

'The medium is the message.'

– Marshall McLuhan

Chapter Abstract

Social decision making in daily life is characterised by the interplay of dynamic agents, yet many studies in social psychology have investigated situations where information flows only in one direction, from a static social partner to a decision maker. This chapter introduces a novel social decision-making paradigm and presents results from a first study. In this paradigm, two people are brought together to the lab and are asked to perform a sequence of perceptual judgments. On each trial, each individual is asked to report their initial confidence on a binary choice. As soon as both individuals input their independent answers information is communicated between the two in one of the two following ways: in the Non-Interactive (NI) condition, each individual sees the initial independent judgment of their partner. In the Interactive (I) condition each individual sees the current opinion of their partner. Importantly participants are incentivised to keep their confidence reports up to date for the entire time window of the social part. This design gives us a continuous measure of confidence change in the two conditions. Results show that: 1) the majority of confidence adjustments follow a step function characterised by an abrupt change followed by smaller adjustments

around an equilibrium; 2) Interaction is characterised by increased recursivity; 3) Trial-level confidence is used to arbitrate conflict although deviating from Bayesian norm; 4) Interaction leads to a magnification of confidence change in agreement and a reduction of confidence change in disagreement. 5) No effect of condition on final accuracy is observed.

Experiment 4

Introduction

The Judge-Adviser System paradigm used in Experiments 1-3 and much prior work (Bonaccio & Dalal, 2006; Sniezek & Van Swol, 2001; Yaniv & Kleinberger, 2000) has many methodological advantages, including control over the information provided by the advice to participants, and distinct staged moments when the decision is made, advice is presented and confidence updated. However it is also characterised by lack of ecological validity that might limit the ability to generalise the results found in a lab setting to more realistic scenarios. Advice taking and social exchanges rarely happen in a vacuum and the stages of decision, advice and update often take place all at the same time. Judge-Adviser systems are characterised by a unidirectional flow of information from the adviser to the participant and by a one-step process where information cannot reverberate back from the judge to the adviser.

Aside from a few examples in which this characterisation is realistic (e.g., receiving recommendations from a consulting company or exchanging opinions by e-mail), most belief updates take place face-to-face, a situation characterised by a bi-directional flow of information between two agents. Here, the line between judge and adviser blurs and both partners affect each other's opinions without a clear distinction of cause and effect. Many researchers have started advocating for a more realistic approach to study social interaction, where social behaviour is studied in active and interacting agents rather than passive observers of social pictures (Boorman et al., 2013; Edelson et al., 2011; Schilbach et al., 2013). The idea that social phenomena cannot be disentangled from social contexts, where interaction happens on a continuous time scale rather than in discrete time steps, raises the possibility that opinion change follows a non-linear dynamic. For example some researchers (Mahmoodi et al., 2013) have found phenomena of confidence escalation when a dyad members performing a task together are allowed face-to-face confrontation. However, little is known about

the direct comparison between continuous and discrete social exchange. The present series of experiments was carried out to fill this gap in the literature. It investigates whether differences between the two modalities - static vs. dynamic interaction - arise even when the information available to participants to make a decision is kept constant.

To investigate these questions, we adapted the dot counting task to a new social scenario, characterised by two real people performing the task in parallel. Participants perform each trial alone, providing their choice and confidence judgment at once using the same input scale used in Experiment 3. After participants provide their independent responses, we allow them to exchange their views over the course of a limited time window, lasting few seconds. All information exchanges happen through computer interface, allowing exact control over relevant variables. During the social time window, participants' confidence changes are monitored in real-time so to track their moment-by-moment bi-directional influence. Each participant is thus both judge and adviser, allowing us to monitor confidence changes and influence in a system characterised by two-way information exchange (Schilbach et al., 2013).

Importantly, however, we compare conditions characterised by static information exchange - where participants can only see their partner's initial judgment - with conditions characterised by dynamic information exchange - where participants can see their partner's moment-by-moment opinion change and thus how they are effecting their partner's views. By keeping decision-relevant information constant in the two conditions it is possible to isolate the contrast between static and dynamically evolving advice. The study is exploratory in nature and aims at showing that opinion change and judgment confidence is not only affected by participants' traits and decision-relevant information, but also by the way information is shared and collectively transformed between advisers. Given that communication happens entirely

through the computer interface, the paradigm allows us to precisely know what information each participant has access to at each time point during the experiment, in my opinion advancing previous studies where interaction took place through verbal communication (Bahrami et al., 2010; Fusaroli et al., 2012). Although the experimental setting in which interaction takes place is far being a realistic social interaction, I believe the paradigm captures the gist of the dynamics that it tries to reproduce.

Based on recent proposals in the literature (Auvray et al., 2009; Dumas et al., 2014; Mattout, 2012; Schilbach et al., 2013), the key prediction is that changing the direction and dynamics of the information exchange (bi-directional vs. one-way; recursive vs. static) will affect the final judgments that participants will converge to, after the social part. Notice that the perceptual evidence accumulated during the private phase by the two members is the only information needed to successfully accomplish the task: the amount of perceptual evidence present in the dyad is the same across the two conditions. The only feature that is manipulated is how that information is allowed to flow from one person to the other, in other words the means of communication. If interaction did not have any effect, we would expect the two conditions not to differ. If on the contrary the format the information is communicated is important as well as the information itself, then we would expect trials in different conditions to differ.

An opinion space to understand social phenomena. This new paradigm however raises a methodological problem. The difficulty of studying more than a single person in a social situation lies in the fact that the distinction between cause and effect is impossible to draw (Watzlawick, Bavelas, & Jackson, 1967). A similar problem is faced when studying many other complex dynamic systems. The brain itself can be regarded as such, with neurons and brain areas being entangled together in a non-stationary interaction. To obviate this issue, neuroscientists have resorted to multivariate statistics (Haynes & Rees, 2005; Kamitani & Tong, 2005). Compared to

univariate statistics (like a *t-test*) where activity distributions are compared along one axis, multivariate statistics tries to separate distributions along a geometric manifold where each factor is considered at the same time with the others (Cortes & Vapnik, 1995). The consensus view that emerged in the last decade is that information is distributed across units of a network instead of being encoded by specific components (McClelland & Rogers, 2003; K. Smith, 2013), although disagreement still exists regarding the mechanisms of its neural implementation (Averbeck, Latham, & Pouget, 2006). Representations can then be encoded by the dynamic interaction of individual units (King, Pescetelli, & Dehaene, 2016; Stokes, 2015). The interaction of N individual units can be thought of as movements along a N-dimensional space, where each dimension represents the level of activation of the unit, say its firing rate. Information from different stimuli can be kept separate as long as those stimuli fall into different regions of the multidimensional space formed by each unit's activation (King & Dehaene, 2014; McClelland & Rogers, 2003; Stokes, 2015). Units have typically been neurons (Averbeck et al., 2006), neural networks' hidden units (McClelland & Rogers, 2003), voxels (Haynes & Rees, 2005; Kamitani & Tong, 2005) and scalp sensors (King et al., 2016). It is here suggested that given the similarities with these other dynamic complex models, similar statistical tools should be employed also to investigate real-time social interactions.

In its simplest form two people (N=2) discuss the likelihood of the state of a binary variable v=[0,1]. At each time point t the dyad's state can be represented as a single point along a phase space M (point A in Figure 5.1), whose dimensions each represents the estimated likelihood $p_i(v = 1|data)$ of the variable's value for each individual i. The likelihood that each member i assigns to the variable at each time point $p_i^t(v = 1|data)$ will change as people exchange information and affect each other's opinion in real-time. Changes of mind, decreases in confidence and confidence escalations should thus be thought of as movements along this "opinion space" (Figure 5.1).

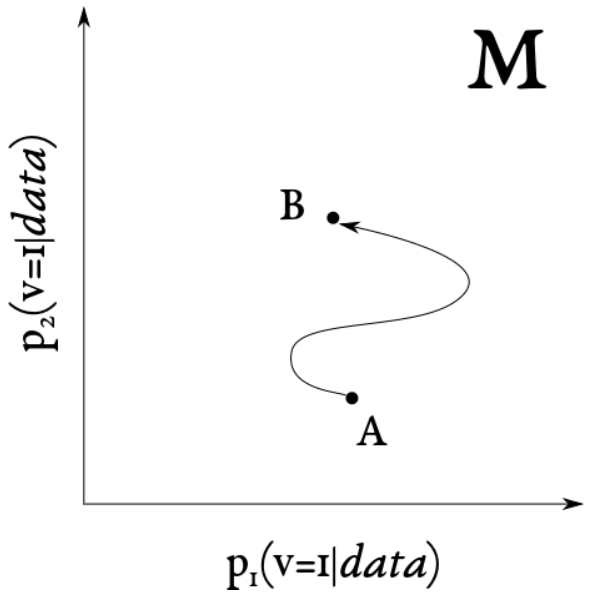


Figure 5.1: Opinion space for binary response variable with N = 2.

Conceiving social interaction as movements along an opinion space reduces the risks of studying oversimplified social situations and allows to analyse two real agents as a single multivariate system. This in turn overcomes the need to separate a receiver (e.g. a judge) of information from a sender (e.g. an adviser) and the need for a staged social exchange.

Methods

Participants. Twenty-four dyads (N=48, 37 females, age: M= 22.52, STD=3.16) were recruited in pairs for money or course credits, through local announcements and the University of Oxford volunteers platform. The invitation informed potential volunteers to bring a friend of the same gender. This was done to avoid confounds due to gender differences in the use of confidence scales and it represents standard practice in the literature (Buchan, Croson, & Solnick, 2008; Mahmoodi et al., 2015). All dyads responded positively to the call, apart for one whose members were gender mixed due to unforeseen circumstances. The study was approved by local ethical committee. All participants gave informed consent before taking part to the study.

Paradigm. Participants sat on the two sides of a desk divided by a wooden occluder (Figure 5.2), each given a separate LCD monitor, keyboard and mouse. All devices were controlled by the same computer, Dell OptiPlex 9020. Together they worked through repeated trials of the same dot-count perceptual decision task as in Experiments 1-3, but now receiving information from a real-human partner rather than a virtual adviser.

All trials consisted of two parts: a private decision followed by a social exchange. On each trial the private decision started with the brief presentation (160 ms) of two delimited boxes, on the left and on the right of a central fixation cross, containing dots arranged randomly. Each dyad member indicated their independent response by mouse-click on a semi-continuous post-decision wagering scale (Persaud et al., 2007), ranging from "100% sure LEFT" to "100% sure RIGHT", with the middle level removed to force participants commit to one or other decision. The scale had fifty levels per side. Participants were informed that each level of the scale corresponded to one token, which was given to them if the answer was correct and taken away from their budget if the answer provided was wrong. Unbeknownst to participants each token was worth 0.01 £. The discounted cumulative earnings at the end of the experiment were given to participants, rounded by 2.50 £.

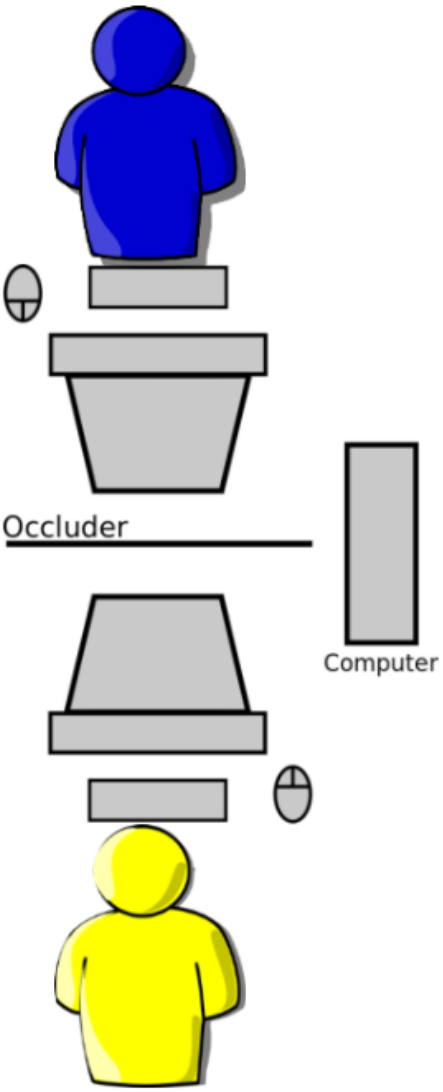


Figure 5.2: Participants sit on opposite sides of a wooden occluder and use one monitor, one keyboard and one mouse each, controlled by the same computer.

The member who responded first waited until the second had input a response as well. As soon as both members confirmed their answer by pressing space bar, the social exchange part started, where each dyad's member was informed about their partner's opinion. At this point confidence changes were recorded continuously. In contrast to the standard Judge-Adviser System paradigm (Bonaccio & Dalal, 2006), where confidence updates happen in discrete steps, here we aimed at recording con-

fidence on a continuous temporal scale, using a continuous post-decision wagering. The mouse x-position along the scale was recorded every 200 ms and each data point so collected was treated as an individual post-decisional bet, contributing to the total amount of tokens participants were supposed to maximise. This was done to incentivise participants to update their cursor position along the scale as soon as their internal confidence changed. No clicking nor confirmation were required during the social part to facilitate a reliable and continuous tracking of confidence change. This part expired after five seconds (26 data points). At this point feedback was provided to both members about the tokens earned by each member and a new trial began.

A staircase procedure was applied to both participants, so that independent of their individual sensitivity to the task, both experienced an equal amount of correct and error trials. Thus, different dot displays were presented to the two dyad members on each trial. Importantly, however, the box with most dots was the same for the two participants on any given trial. This ensured that social information coming from the other person carried meaningful information.

Manipulation. Our manipulation involved only the social exchange part. Two conditions were designed and alternated across blocks. In the Non-Interactive (NI) condition, the choice and confidence level selected by each dyad's member in the private phase appeared on their partner's scale as a static coloured cursor. Dyad members were at this point asked to continuously monitor and update their confidence by moving their mouse along the scale. In the Interactive (I) condition, the social exchange part started exactly as in the Non-Interactive one, with each dyad's member's cursor appearing on their partner's scale. However, and for the whole duration of the social part (five seconds), if a member updated their confidence, this would instantly appear also on their partner's scale and vice-versa. This led to a situation where participants were not only informed of their partner's original opinions but also

how those opinions were changing in real-time as a function of their own changes of mind (Figure 5.3).

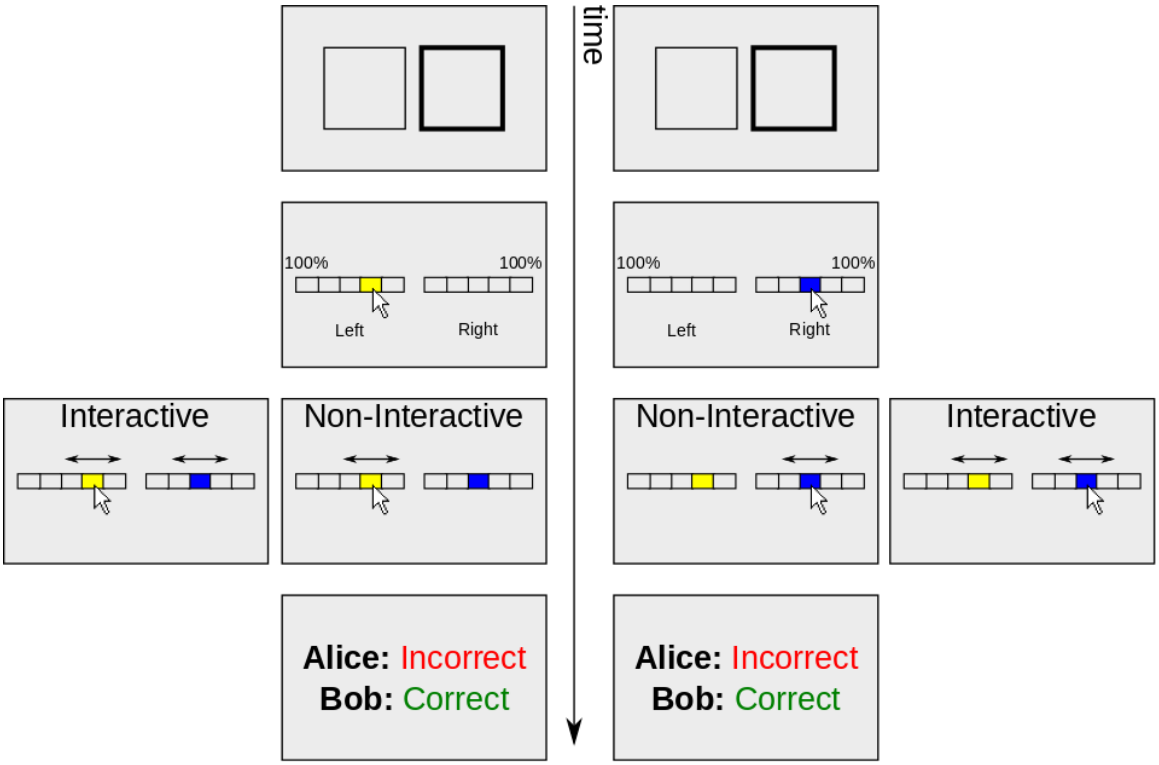


Figure 5.3: After playing the dot task alone participants see on their own scale their partner's initial opinion (NI condition) or their partner's current cursor position (I condition). They receive feedback on their accuracy and earnings at the end of the trial. Bidirectional black arrows represent real-time movement along the scale. The scale used in the actual task had 50 levels per interval.

The experiment included three blocks of practice of 10 trials each (practice with the perceptual task, practice with the Non-Interactive condition, practice with the Interactive condition) and 14 experimental blocks of 25 trials each. Each experimental block contained 2 null trials, which were private decision only trials. In null trials, earnings were calculated from the confidence expressed during the private phase only. Null trials were introduced so that participants were motivated to report their confidence accurately also during the private decision. Normal trials' earnings were computed instead from the social part.

Analyses were performed to assess how social exchange (interactive or non-interactive)

affected dependent variables of interest, mainly confidence, accuracy and confidence-to-accuracy calibration. We thus looked at accuracy change, confidence change and calibration change from pre-to-post social information, separately for the two conditions.

Results

A note on continuous update. The use of continuous post-decision wagering gave us a time series of confidence positions for both dyad members on every trial over the five seconds time window of each trial's social part. Visual inspection of the data showed that the usual pattern of confidence update over time was characterised by a single rapid transition from starting confidence level to final confidence. Further adjustments were rare and remained around this final equilibrium point. The number of transitions was formally defined as the number of times during a trial's social window that the following conditions applied simultaneously: $[C(t) \neq C(t-1)]AND[C(t-2) = C(t-1)]$, where $C(t)$ is the confidence recorded at time sample t. The number of update transitions within a trial significantly differed between conditions, but the effect size was small (average number of transitions: Non-Interactive = 1.11, Interactive = 1.18, $t(47) = 2.95, p = .004, d = 0.12$). The result indicates that marginally (but consistently) more updates happened during the Interactive condition.

To test for differences between conditions in the updating dynamics, we computed the average absolute difference between each confidence point and the next one ($|C(t) - C(t-1)|$) over the course of the five seconds time window of the social part. Figure 5.4 shows the average absolute difference between consecutive time points (200 ms apart) in the two conditions, indicating the stability of the update (zero corresponding to stationary confidence). A major transition takes place around second 1 of the social time window. Notice that the gradual decrease observed after the first second does not necessarily indicate progressively smaller update sizes but

could derive from averaging different trials where equilibrium was reached at different time points of the social window. The inset shows the difference between the Interactive and Non-Interactive condition within participants over time. Points above the reference line indicate larger updates in Interactive condition, points below indicate larger updates for Non-Interactive trials. Shaded grey areas show point-wise significant differences between the two time series, suggesting that confidence updates in the Interactive condition might have lasted for longer. However the difference did not survive a cluster-based permutation t-test to account for multiple comparison problems.

As different individuals might show different confidence change peaks and skewness, we fitted a Poisson distribution to individual data for each condition separately and performed second-order statistics on the estimated λ, a parameter controlling the rightward skewness of the distribution. Greater λs indicate that confidence changes took longer to reach equilibrium. A one-tail t-test showed that the Interactive condition was characterised by larger λ values (Non-Interactive = 8.16, Interactive = 8.48, $t(47) = 2.19, p = .03, d = 0.18$), indicating greater rightward skewness and thus longer times to reach stationary confidence. Although different distributions could have been chosen instead (e.g., ex-Gaussian, χ^2, beta etc.), the Poisson distribution was preferred due to the single free parameter to be fit and because it is usually used to describe counts of observations.

Thus, although some evidence seems to suggest that the Interactive condition led to longer updates, the differences were modest and the effect size small. As the average trial consisted of a single transition, analyses on confidence reported below are performed on last confidence points registered on each trial, unless explicitly specified.

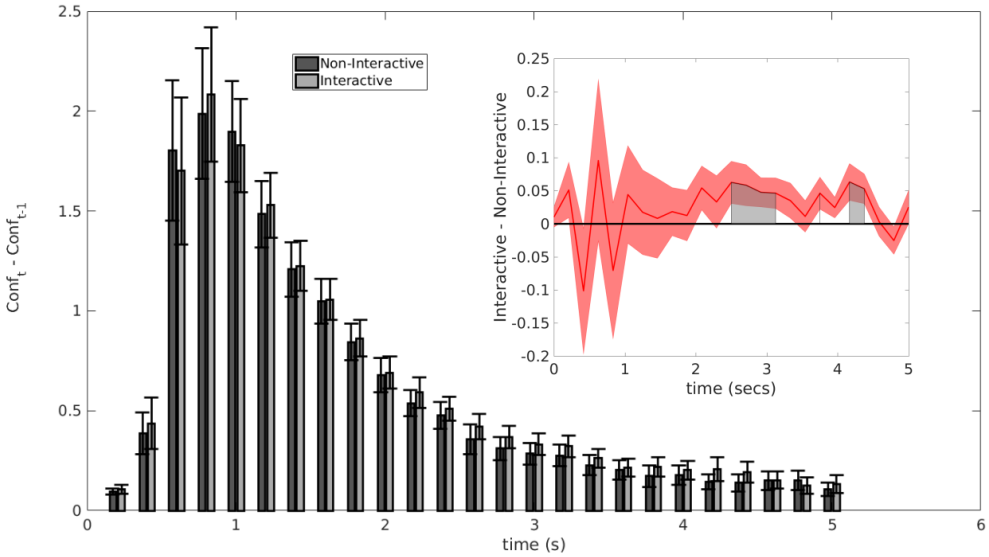


Figure 5.4: Main plot: average absolute difference between two consecutive confidence data points over the course of the 5 seconds social window. A large update is performed within the first two seconds and then reaches a confidence equilibrium. Inset: average difference between conditions. The absolute difference between one data point and the next one was compared between conditions. Grey areas, indicating pairwise significant differences, did not survive to a cluster-based permutation t-test.

Confidence distributions. Figure 5.5 shows the pre-social information confidence distributions of each participant tested. Many distributions were highly clustered towards high confidence. It may be possible that low risk aversion characterised our participants thus making them bolder than their actual confidence. In tasks with average accuracy greater than chance the optimal post-decision wagering strategy is always to bet the maximum allowed. Lack of loss aversion thus can potentially make interpretation of wagers difficult, as participants might have learnt to bet high irrespective of internal confidence. As addressed in the Discussion however, there is reason to believe that results were not affected by the extreme use of the confidence scale. Interestingly, distributions of members of the same dyad seemed more similar to each other than distributions belonging to members of different dyads. This observation inspired Experiment 7, described in Chapter 7.

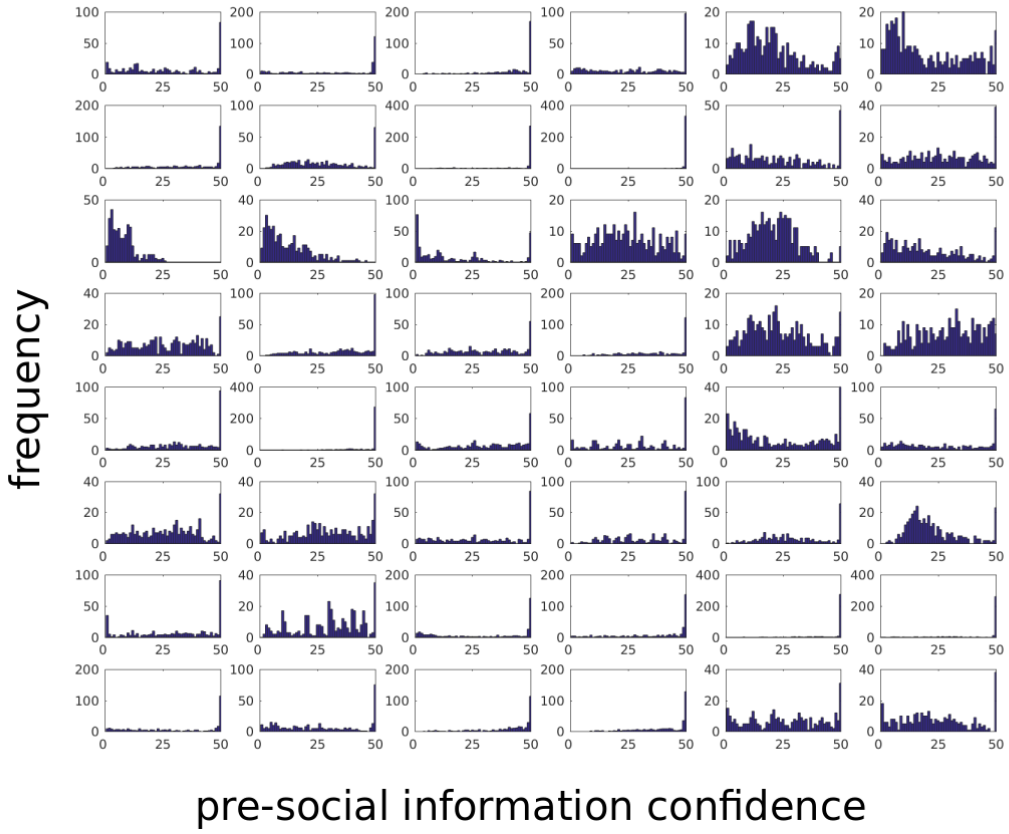


Figure 5.5: Pre-social information confidence distributions for each participant tested. The first dyad members are shown on the first row, first and second plots. The second dyad is shown on the first row, third and fourth plots, and so on.

Confidence changes. The influence that social information has on each dyad member can be quantified as the distance between their post-social information confidence and pre-social information confidence: $\delta_C = C_{post} - C_{pre}$. As shown in Experiments 1-3, δ_C is expected to be positive in agreement trials and negative in disagreement trials. Given that participants' first judgments were independent and staircased to about 70% accuracy, we expected an agreement rate of 0.58 (i.e., $0.7^2 + 0.3^2$). Close to this nominal value, the observed agreement rate was 0.60 ± 0.03. A 2x2 ANOVA on confidence change (consensus x condition) showed significant main effects of consensus ($F(1, 47) = 150.26, p < .001, \eta^2_G = .7$) and condition ($F(1, 47) = 9.40, p < .01, \eta^2_G = .005$), but no significant interaction ($F < 1$).

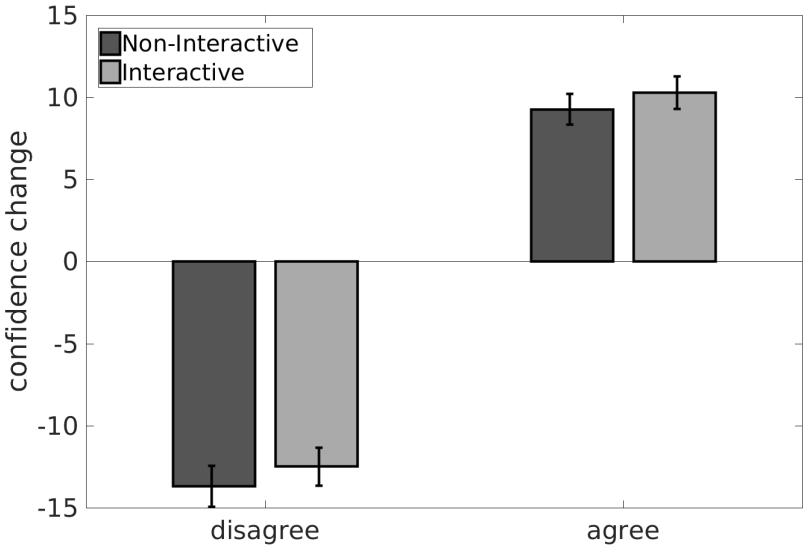


Figure 5.6: The effect of condition on the magnitude of confidence change. Larger changes are observed in the Non-Interactive condition on disagreement trials and in the Interactive condition on agreement trials. Interaction generates greater updates in agreement and smaller updates in disagreement. Error bars represent s.e.m.

As Figure 5.6 shows, δ_C was negative in disagreement and positive in agreement, indicating that, as one would expect, participants tended to increase their confidence when their opinions agreed and decrease their confidence when they disagreed. Interaction had opposite effects in agreement and disagreement: participants increased their confidence more when in agreement ($t(47) = 2.69, p = .009$) but decreased their confidence less when in disagreement ($t(47) = 2.08, p = .04$), compared to a Non-Interaction case. Thus, the net effect of interaction was on average to make confidence changes more positive, regardless of whether the dyad members initially agreed or disagreed. Comparing average post-social information confidence, however, we did not find a significant difference between the Interactive and Non-Interactive condition (Non-Interactive $= 34.66 \pm 8.48$, Interactive $= 34.87 \pm 8.88, t < 1$). Instead, confidence differences between the conditions were more evident in the private decision phase, being somewhat greater in the Non-Interactive condition (mean $= 30.56 \pm 8.99$) than in the Interactive condition (mean $= 30.05 \pm 8.93$), although the difference was very

small and only marginally significant ($t(47) = 1.82, p = .07$). Figure 5.7, representing density plots of the pre- and post-social information confidence distributions across conditions, indicates that perhaps Interaction was characterised by fewer extreme pre-social information confidence ratings. The difference was however weak and it is not believed to have confounded our findings.

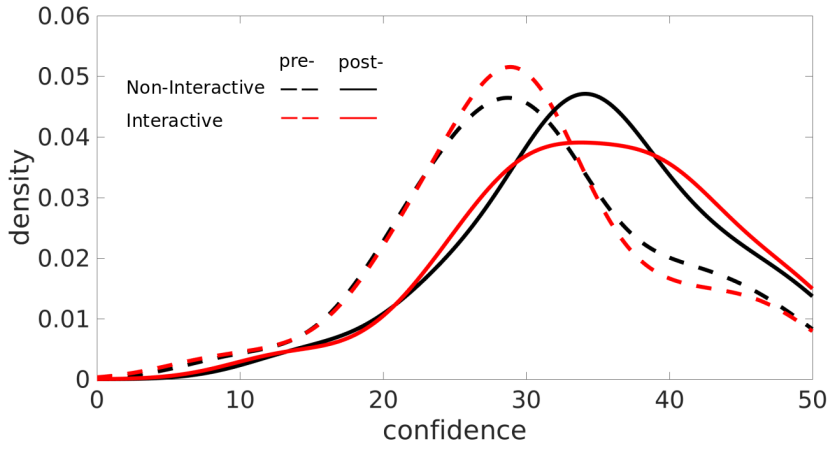


Figure 5.7: Confidence distribution density plot before and after social information, divided by condition (average across participants).

Asymmetry in confidence changes. Figure 5.8 shows the frequency distribution of confidence changes divided by condition and consensus (agreement vs. disagreement). It can be observed that, again, agreement tended to make participants more confident and disagreement tended to make them less confident. All distributions peaked around zero, which was by far the most common change (note that the y-axis in Figure 5.8 uses a square root scale), suggesting that very often participants ignored social information received. It can also be noticed that on some proportion of disagreement trials participants increased their confidence and on some proportion of agreement trials they decreased it. This is an interesting result if we consider that, from a Bayesian perspective, disagreement with an independent observer should always lead to reduction of confidence and agreement should always lead to an increase of confidence (if it is assumed that the partner performs better than chance). Indeed

disagreement (no matter how weak) means that the other person carries disconfirming evidence for the participant's own opinion, a fact that should always lead the participant to become more uncertain. Conversely, agreement (no matter how weak) carries confirmatory evidence and should result in greater confidence[1].

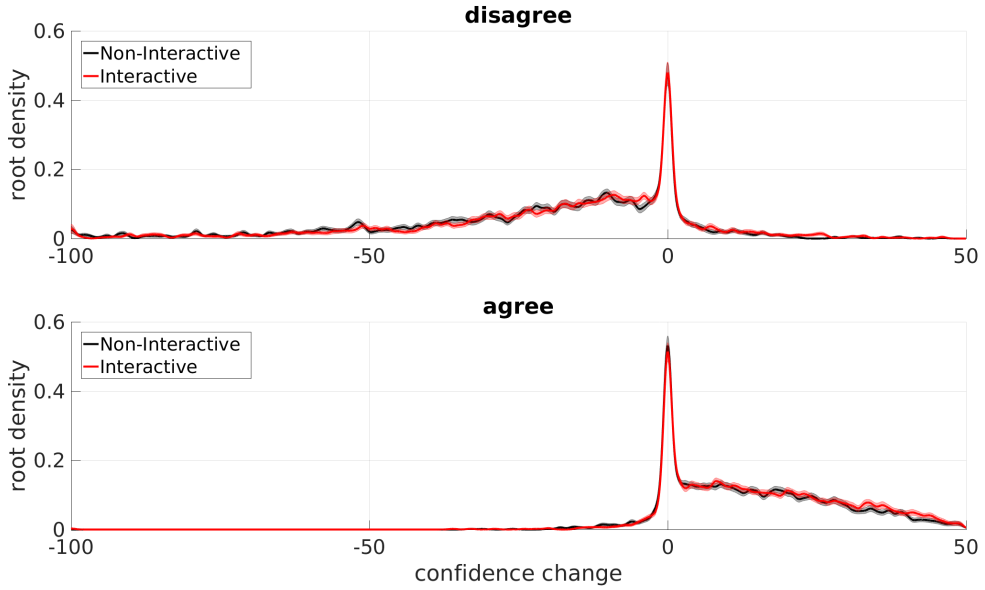


Figure 5.8: Density plot of confidence change divided by condition and consensus. It can be noticed that overall agreement trials correspond to increases of confidence and disagreement to decreases in confidence. However some confidence changes in the wrong direction are also observed.

To explore these behaviours more formally, a two-way repeated measures ANOVA was computed on the probability of an irrational confidence change with factors condition and consensus (agreement vs. disagreement). Irrational confidence changes were defined as decreases in confidence in cases of agreement and increases in confidence in cases of disagreement. To avoid including trials where increases/decreases in confidence were simply due to involuntary cursor movements ("trembling hand"),

[1]Consider the case where observer A indicates LEFT as the correct answer with an initial (prior) confidence level of P(LEFT), then learns of low confidence agreement from a partner (which we can term "adviceL"). In this case, the posterior belief from a Bayesian perspective is given by $P(LEFT|adviceL) = P(LEFT) * [P(adviceL|LEFT)/P(adviceL|LEFT)P(LEFT) + P(adviceL|RIGHT)P(RIGHT)]$. According to this equation, the posterior belief $P(LEFT|advice)$ will exceed the prior $P(LEFT)$ whenever social information has an above-chance likelihood of being correct (i.e., $P(adviceL|LEFT) > P(adviceL|RIGHT)$).

we considered "change" a shift bigger than 5 confidence points in the unexpected direction. Results were robust across other cutoffs greater than zero. Results show a significant effect of consensus ($F(1, 47) = 7.88, p = .007, \eta^2_G = .07$) but not of condition ($F(1, 47) = 2.56, p = .11, \eta^2_G = .001$) and a significant interaction between the two ($F(1, 47) = 9.90, p = .002, \eta^2_G = .005$). These significant effects reflect the fact that irrational changes were more frequent in disagreement than agreement (0.018 vs. 0.004 of trials) and that the Interactive condition produced more "irrational increase" trials than the Non-Interactive condition (0.020 vs. 0.015 of trials, $t(47) = 2.59, p = .01$), as well as fewer irrational decreases (0.003 vs. 0.005 of trials, $t(47) = 1.98, p = .05$).

Presumably irrational decreases happen when a partner agrees but is much less confident than the participant, which might lead the participant to think that they ought not to have been so confident in the first place. Many opinion aggregation strategies described in the literature (Bang et al., 2014; Larrick & Soll, 2006; Pescetelli et al., 2016) - like averaging - can explain irrational decreases, but not irrational increases. We were thus particularly interested in irrational increases in confidence after disagreement, which occurred more frequently than irrational decreases and are difficult to reconcile with any obvious confidence-update strategy. This irrationality could occur through non-linear dynamics introduced by the possibility of interaction.

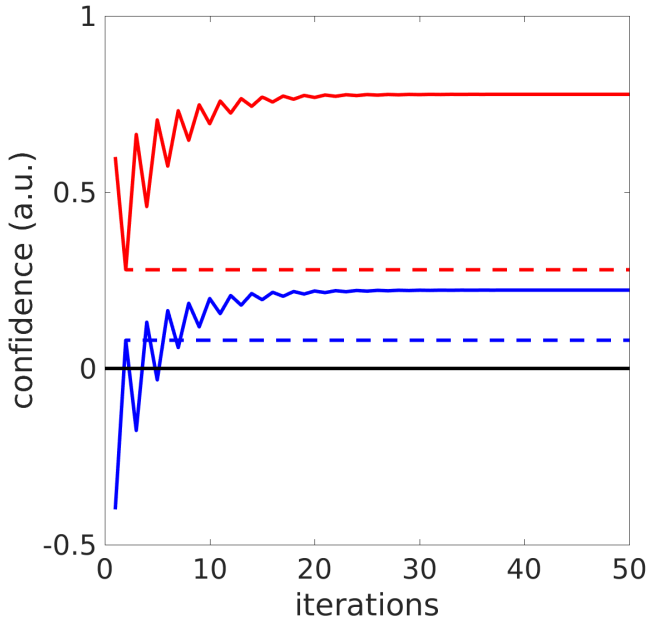


Figure 5.9: The figure shows a simple toy model emulating how confidence could increase after disagreement in interactive condition. Two individuals, red and blue, starts at time 1 with different levels of confidence on opposite decisions (i.e. disagreement). In a Non-Interactive condition the two individuals update their initial opinion with their discounted partner's opinion (dashed lines). In Interaction on the contrary the same update rule is applied on every iteration until equilibrium is reached (solid lines). This latter condition leads to irrational increases (e.g, red line ends more confident than at starting point).

To see how, consider the following simple simulation that tried to mimic belief update in the two conditions. Consider an example trial where a participant (P_{max}) starts off on a confidence level of $C^{pre}_{max} = 0.6(a.u.)$ while his/her partner (P_{min}) weakly disagrees ($C^{pre}_{min} = -0.4$). Imagine now that both participants use a simple update strategy, namely summing their own initial confidence with their partner's discounted confidence (Pescetelli et al., 2016) (here: discount factor = 0.80). In a situation where no interaction is allowed participants can only use their partner's initial opinion to get to a new confidence, thus reaching levels of $C^{post}_{max} = 0.28$ and $C^{post}_{min} = 0.08$. Imagine now an interactive scenario where the participant has access at each iteration to his/her partner's *current* confidence level and uses it to recursively updates his/her initial confidence. Figure 5.9 shows that this simple strategy leads to an oscillatory

update that stabilises for P_{max} on a higher confidence than initially held. The effect can be explained by the fact that as soon as P_{min} crosses the decision boundary 0, disagreement turns into agreement, thus supporting P_{max}'s initial opinion, instead of providing contradictory evidence. To test for recursive dynamics in our behavioural data we counted, for each condition, the average number of times the direction of the update (i.e., stationary/increase/decrease) changed in the update window. Formally: $(r_t - r_{t-1}) \neq 0, r = sign(C_t - C_{t-1})$. Changes of update direction were significantly more frequent in Interactive than Non-Interactive condition (Non-Interactive $= 2.41 \pm 1.26$, Interactive $= 2.55 \pm 1.27, t(47) = 2.62, p = .01, d = 0.11$), supporting the intuition behind our toy model. Besides, it must be observed that such a process could in principle happen implicitly before or during participants' cursor's movement without being overtly oscillatory. This simulation offers a proof of concept rather than an exhaustive description of the mechanisms underlying irrational increases. The key point is simply to show how irrational increases in confidence could arise through interaction. Irrationality here comes from the fact that a partner's update (which reflects the participant's own influence) is then incorporated into the participant's updated belief as if it were independent evidence. That being said, the simulation predicts that only dominant participants should increase their confidence in cases of disagreement. We re-tested the frequency of irrational increases observed in dominant and dominated participants across the two conditions (where dominance is defined trial-wise according to expressed confidence) and found that indeed only dominant participants showed a significantly higher rate in interaction than in no-interaction (0.012 vs. 0.008 of disagreement trials, $t(47) = 3.29, p = .001, d = .21$), but dominated participants did not (0.007 vs. 0.007 of disagreement trials, $p > .8$).

A multivariate opinion space to represent dyad states. As both participants could independently vary their confidence at the same time along the confidence scale, confidence between dyad members was orthogonal to each other. Thus, any

analysis that averages across trials without taking into account this complexity risks missing effects that are observable only on specific subsets of trials. To explore this complexity in the data, we plotted dependent variables of interest, like confidence changes, on a 2-dimensional "opinion space" (Pescetelli et al., 2016), representing them on a multivariate surface (see Figure 5.10 below for an example). The x-axis of the opinion space represents the confidence of the more confident or "dominant" member (P_{max}) *on any given trial*, thus ranging from 1 (minimum confidence level allowed) to 50 (maximum confidence level allowed). The y-axis of the opinion space represents the confidence of the less confident or "dominated" member (P_{min}) *on a given trial* and relatively to their more confident partner. As P_{min} can either agree or disagree with P_{max}, y-axis ranges from -50, disagree with maximum confidence, to 50, agree with maximum confidence. This creates a grid of possible social situations where the dyad's state - namely both members' choices and confidence levels - is fully represented by a set of coordinates. This notation also gets rid of redundancy of the side of the choice (LEFT vs. RIGHT) and maintain only the reciprocal distance of opinions along the scale. Unless specified otherwise, the (x,y) coordinates represent the dyad's state before social information is exchanged, while the colour (z-axis) can represent any dependent variable of interest (e.g., the change in confidence from pre- to post-social information of either dyad member).

The space can be divided in two areas, depending on initial consensus. The top half represents situations where the dyad members agreed on the same interval before the social exchange period. The bottom half on the contrary represents situations where they initially disagreed. The bottom right and top right corners represent situations of maximum conflict (-50, 50) and maximum agreement (50, 50) respectively. Point (0,0) on the contrary represents a nominal situation of maximum uncertainty, although a wager of 0 was not allowed in this experiment. The domain of existence of the opinion space does not include, by definition, points where the x-coordinate is smaller than the y-coordinate. Points outside the domain of existence are represented as the lowest

value of the color map in all figures below. The opinion space is simply a method to visualise the data of interest - e.g., the confidence changes explored above - to give insights into possible effects that do not appear in the averaged data. I will provide a qualitative description of the graphs, but do not apply any statistical testing, also because subdividing trials into many cells (each individual set of coordinates) results into sparse or uneven split of trials.

To further understand how confidence changes were affected by the initial configuration of members' opinions, we plotted median confidence change δ_C using the multivariate representation (Figure 5.10). Trials within each condition and dyad were linearly interpolated, due to data sparseness, using MATLAB "scatteredInterpolant" function. The figure shows median confidence change across dyads in Non-Interactive (panels A,D) and Interactive (panels B,E) conditions. Confidence changes were divided by trials when the participant held the dominant (A-B) or dominated (D-E) opinion. The multivariate plot can also be used to plot contrasts between conditions, as shown in panels C and F for the dominant and dominated case respectively. The contrast plot represents the difference between the Interactive and Non-Interactive conditions. In both dominant (panel C) and dominated (panel F) trials, warmer colours are observed on grid positions corresponding to situations of uncertain agreement (points x on the graphs). The warmer colour represents the fact that in these situations interaction led to greater confidence increases than in the non-interactive case in both members. As it can be seen from the colourmap, confidence change magnitude in the Interaction condition (panel B,E) under uncertain agreement is around 20-30 confidence points, indicating that in these trials participants tended to converge to the maximum confidence level. A real-time animation of confidence change can be found at `https://niccolopescetelli.com/confidence-change-in-opinion-space/`.

Another point of interest in both contrast plots can be observed in the disagreement half of the opinion space (points y, panels C and F). These were trials when the

dominant member was highly confident and the dominated member weakly disagreed. The warmer colour in panel C represents the fact that in these trials the dominant member reduced their uncertainty less compared to non-interactive information sharing. We have seen above how interactive condition reduced the disagreement effect on confidence change. The representation in opinion space helps to further pinpoint which trials and which dyad member this effect was mostly driven by. Similarly, confidence changes in dominated trials showed reduced confidence decreases in interaction, as indicated by the negative values on the contrast plot (panel F, point y).

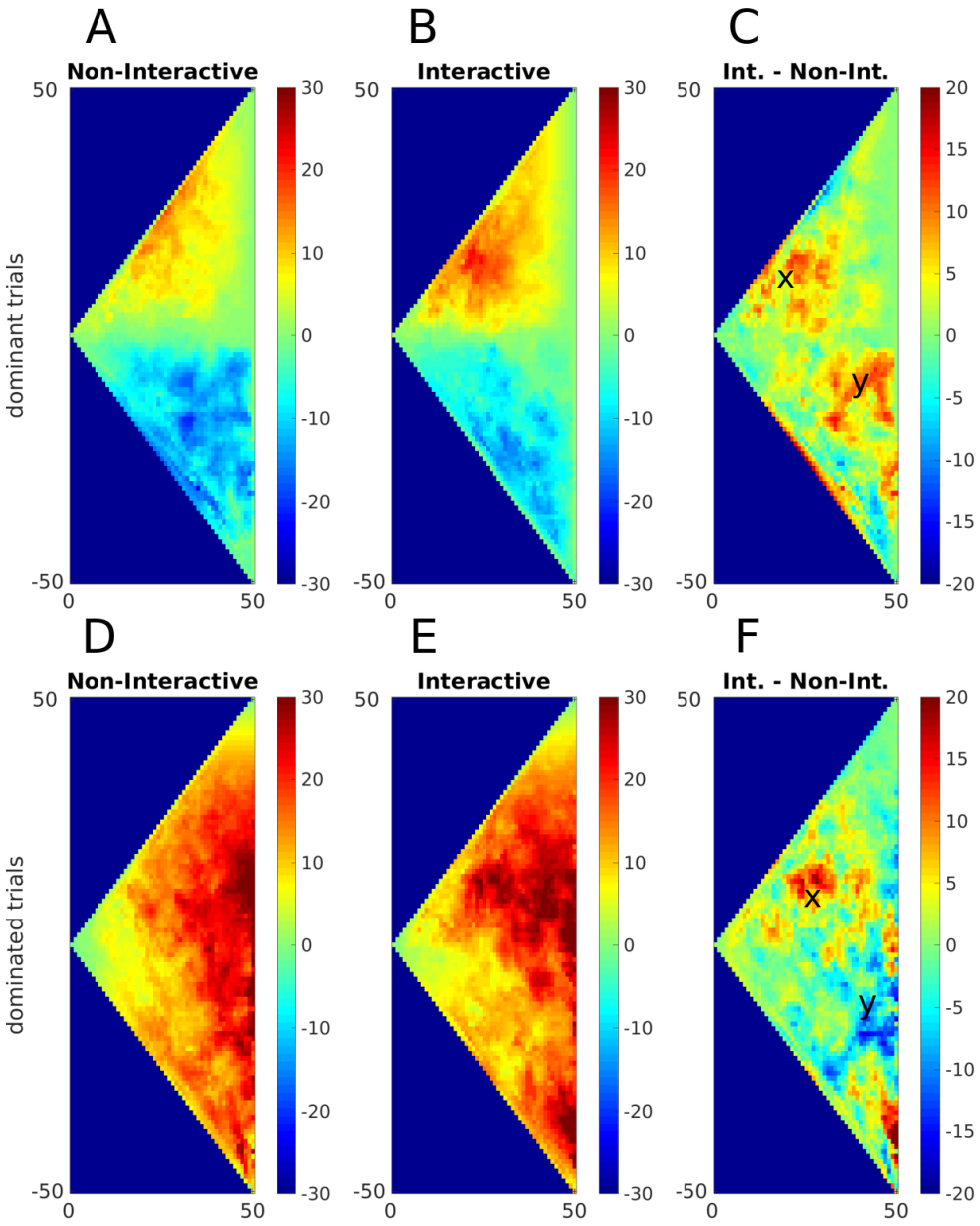


Figure 5.10: Confidence change observed for every social situation, according to the confidence of the more (P_{max}) and less (P_{min}) confident participant. (x,y) coordinates represent the pre-social information dyad's state, while the colour (z-axis), represents confidence change. Panels A-C represent confidence change in dominant trials in Non-Interactive condition, Interactive condition and their contrast respectively. Panels D-F represent confidence change in dominated trials in Non-Interactive condition, Interactive condition and their contrast respectively.

Coupling of confidence changes in interaction. The results presented so far indicate how confidence changes were affected by the partner's opinions, but say nothing

about how the two varied together in each trial. To investigate this issue, we correlated the absolute confidence changes of members of the same dyad. Absolute confidence change was defined as the absolute difference between final post-social decision confidence and pre-social decision confidence, with zero representing situations where the participant did not move along the scale after knowing their partner's opinion. I correlated across trials absolute confidence changes of members belonging to the same dyad. Figure 5.11 shows the average Pearson's r coefficient when calculated separately for the factorial combination of consensus (agreement vs. disagreement) and condition. A 2x2 ANOVA (consensus x condition) on Pearson's r coefficients, with dyad as the random effect (1 dyad removed for a missing cell), showed a main effect of consensus ($F(1, 22) = 20.93, p < .001$) but not of condition ($F(1, 22) = 1.71, p = .20$), and a significant interaction between the two ($F(1, 22) = 38.39, p < .001$). Not surprisingly, when dyad members could not see each other's confidence changes (Non-Interactive condition), one member's confidence changes were uncorrelated from their partner's. In the Interactive condition on the contrary, the two changes became coupled: In agreement trials, the correlation was positive, indicating that the more one member increased their confidence the more their partner also increased their confidence; in disagreement, the correlation was negative indicating that the more one member decreased their confidence in their initial decision the less their partner decreased their confidence. Pairwise contrasts showed that, compared to Non-interactive condition, in the Interactive condition correlation coefficients were significantly greater for agreement ($t(22) = 4.89, p < .001, d = 1.20$) and marginally smaller for disagreement ($t(22) = -2.02, p = .05, d = -0.52$).

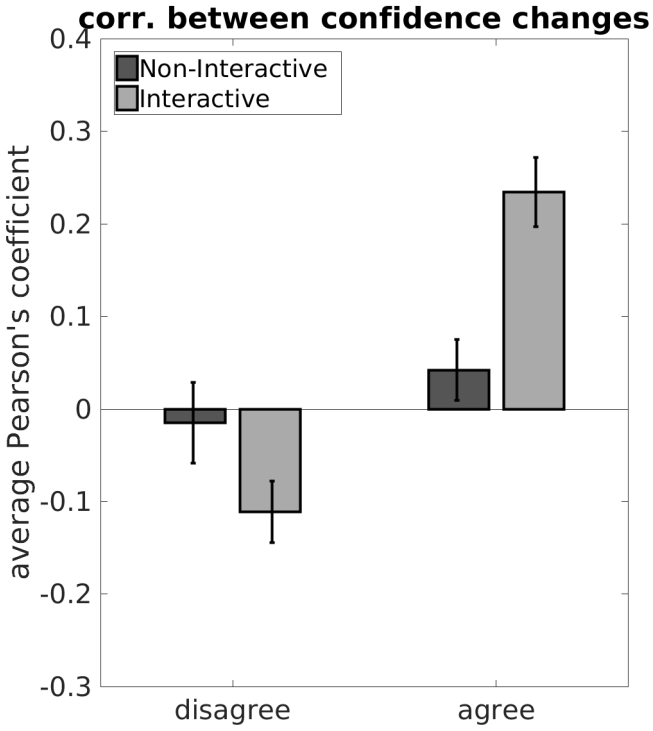


Figure 5.11: Effect of condition on the correlation between absolute confidence changes of the two participants across trials. Average Pearson's correlation coefficient is plotted as a function of consensus and condition.

This coupling of confidence change in interaction hints to the fact that participants made use of their partner's confidence change when updating their own confidence, as opposed to basing their change on their partner's independent (i.e., initial) opinion. A nested linear mixed-effects model on trial-by-trial absolute confidence change was run to test this hypothesis. The model was defined as follows:

$$\begin{aligned} |\delta_C^s| \sim C_{pre}^s * C_{pre}^p * Condition * |\delta_C^p| * Consensus + \\ (1 + C_{pre}^s + C_{pre}^p + Condition + |\delta_C^p| + consensus|dyad) + \\ (1|subject : dyad) \end{aligned} \tag{5.1}$$

where the dependent variable $|\delta_C^s|$ is the absolute confidence change observed on a particular dyad member (*self*). Predictors included the participant's initial confidence C_{pre}^s as well as their partner's C_{pre}^p, the condition (Non-Interactive vs. Interactive),

the relationship between their initial views (agreement vs. disagreement), absolute confidence change observed in the partner $|\delta_C^p|$, and all interaction terms. The introduction of the term $|\delta_C^p|$, i.e. the absolute partner's confidence change, is crucial in testing our hypothesis. Indeed, a rational agent would only use task-relevant information (perceptual evidence) to update its confidence. It would not use partner's confidence change, given that this is non-independent and biased by the agent's opinion itself. Confidence and confidence change were normalised within participants; condition and consensus were declared categorical predictors. Random intercepts and slopes for each dyad were modelled for each main effect. Participants were declared nested within dyads. Main effects are shown in table below:

	Estimate	SE	tStat	DF	p
Intercept	0.2264	0.0494	4.5739	15424	<.001***
Interaction	-0.0512	0.0297	-1.7185	15424	.08
Agreement	-0.3118	0.0713	-4.3716	15424	<.001***
$\|\delta_C^p\|$	-0.0287	0.0180	-1.5896	15424	.11
C_{pre}^s	-0.0251	0.0424	-0.5911	15424	.55
C_{pre}^p	0.2953	0.0321	9.1809	15424	<.001***

Table 5.1: Fixed main effects of a linear mixed-effect multilevel model run on trial-by-trial absolute confidence update. Main predictors are (a) condition: Non-Interaction (reference), Interaction; (b) consensus: Disagreement (reference), Agreement; (c) partner's absolute confidence change ($|\delta_C^p|$); (d) personal initial confidence (C_{pre}^s); (e) partner's initial confidence (C_{pre}^p).

The interaction term between condition and consensus ($\beta = 0.14, SE = 0.03, p < .001$) confirmed the opposite effects condition had on confidence change in agreement and disagreement trials. Importantly, this interaction was positively modulated by partner's absolute confidence change ($\beta = 0.29, SE = 0.03, p < .001$), suggesting that the more a participant's partner was willing to change their initial confidence the greater the participant's changes were in agreement trials and the lower they were in disagreement trials, confirming our hypothesis that participants made use of non-independent information, thus generating the judgments correlations found in

Figure 5.11. The same condition by consensus interaction was negatively modulated by participant's initial confidence ($\beta = -0.08, SE = 0.03, p = .007$) suggesting that the stronger the confidence initially held the less Interaction differed from a Non-Interactive baseline. The opposite relation was true for the participant's partner's initial confidence ($\beta = 0.09, SE = 0.03, p = .001$) indicating that, in Interactive compared to Non-Interactive condition, greater partner's initial confidence predicted, after agreement, greater participant's confidence increases. In disagreement, on the contrary, the same factor predicted smaller confidence decreases.

Implications for accuracy. One hypothesis brought forward in the literature on collective intelligence is that social interaction hampers the collective wisdom by breaking the independence of the individual judgments (Lorenz et al., 2011). The traditional interpretation of wisdom of crowds effects (Galton, 1907) is the Noise Cancelling Hypothesis, which explains the accuracy improvement seen in opinion aggregates as a statistical phenomenon of noise reduction following averaging of independent samples (here the private initial opinions). This hypothesis predicts that breaking the independence between measures should have negative effects on accuracy, as errors get correlated instead of averaging out. In their experiment Lorenz et al. (2011) showed that simple exposure to others' opinions had damaging effects on performance. If this interpretation is correct, in the present study we should observe that (1) simple exposure to another person's opinion negatively affects performance; (2) the effect of social exposure is even more damaging on performance in the Interactive condition, as this condition affects the independence of the individual estimates more than the Non-Interactive one (Figure 5.11).

A 2-way ANOVA on accuracy with factors condition (Non-Interactive vs. Interactive) and decision type (pre- vs. post-social information) showed a significant effect of decision type ($F(1, 47) = 47.00, p < .001, \eta^2_G = .16$) but no significant effect of condition ($F < 1$) nor significant interaction ($F(1, 47) = 0.91, p = .34, \eta^2_G = .001$).

Results show that social information had a beneficial effect on average accuracy (pre-social information accuracy = 0.72, post-social information accuracy = 0.75, $t(47) = 7.14, p < .001$). The finding indicates that, contrary to Lorenz et al. (2011), exposure to another person's opinion did not reduce accuracy. Furthermore, Interaction did not reduce accuracy improvement compared to Non-Interaction, indicating that increased dependence between confidence updates (as indicated in the analyses of Figure 5.11) had no significant damaging effect on accuracy either.

However, condition might have affected people's opinions without making them changing their initial answer (Bonaccio & Dalal, 2006). We thus tried to define a more nuanced definition of accuracy improvement as confidence changes toward or away the correct end of the scale: $\delta_{acc_G} = (C_{post} - C_{pre})^{Acc} * (C_{pre} - C_{post})^{1-Acc}$, where Acc can be either 1 or 0. We tested whether participants improved their initial judgment although keeping their initial views. The same 2x2 ANOVA did not show any effect of condition on graded accuracy improvement δ_{acc_G} nor interaction ($F < 1$), thus rejecting the idea that condition had an influence on accuracy.

Calibration change. Results so far show that interaction increased participants final confidence without affecting their final accuracy. We thus investigated whether interaction affected the calibration of confidence relatively to objective accuracy, defined here as the type II A_{ROC} (Fleming & Lau, 2014). The same two-way ANOVA used for choice accuracy was run on type II A_{ROC}. Results show a significant effect of decision time ($F(1, 47) = 89.58, p < .001, \eta^2_G = .25$), indicating calibration improvement from pre- to post-social information phase (0.60 vs. 0.66), but no effect of condition nor interaction between the two ($F < 1$), indicating that overall condition did not impact on calibration improvement.

Comparing human behaviour with Bayesian optimality

In this section, I compare human confidence updates with a normative framework obtained from Bayes theorem. If we assume that participants' expressed confidence linearly maps onto probability scale, we can aggregate participants' independent judgments using Bayes rule. We assume for simplicity that participants are perfectly calibrated, that is their expressed confidence (expressed in probability) has a 1:1 relation with the probability of a correct response on a given trial. We know that this is a simplifying assumption, as studies have demonstrated that people greatly vary in their confidence calibration (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Song et al., 2011). Nevertheless, variability in partner's confidence judgments should be used by the participant to inform his/her own judgment as it carries crucial information about the partner's trial-by-trial probability correct.

Participants' confidence (range: -50 = "100% Sure LEFT" to 50 = "100% Sure RIGHT") were passed through a linear transformation that mapped them onto a probability scale and prevented values 0 and 1 to avoid impossible solutions (range: 0.01 = "100% Sure LEFT" to 0.99 = "100% Sure RIGHT"). The probabilities P_s and P_p so obtained - representing dyad members' independent priors (cf. Harris, Hahn, Madsen, & Hsu, 2016)- can now be integrated as:

$$post_{norm} = \frac{P_s P_p}{P_s P_p + \neg P_s \neg P_p} \tag{5.2}$$

where $\neg P_s$ and $\neg P_p$ are $1 - P_s$ and $1 - P_p$ respectively, representing the subjective probability of an error. The posterior confidence so obtained (in probability scale) represents the post-social information confidence held by a normative opinion aggregation method.

Humans show overconfidence compared to Bayes

The normative confidence change δ_{norm} was then compared to the empirical confidence change δ_{emp} observed in the data. This produced an error term that quantifies the discrepancy between normative and empirical change: $Err = \delta_{emp} - \delta_{norm}$. Positive error values indicate the participant's post-social information confidence was greater than the normative model. Correspondingly, negative values indicate that participants' post-social information confidence was too low. The results, shown in Figure 5.12, show that on average participants were more confident than the normative posterior after a disagreement and less confident than the normative posterior after an agreement. The magnitude of the discrepancy was larger in disagreement trials. A 2x2 ANOVA on the Err quantity revealed a significant effect of consensus ($F(1, 47) = 68.37, p < .001, \eta^2_G = 0.46$) and condition ($F(1, 47) = 4.97, p = .03, \eta^2_G = .002$) but no significant interaction ($F < 1$), indicating that deviations were more on the positive side (overconfidence) for Interactive trials than for Non-Interactive ones, and for disagreement than for agreement. When compared to zero (normative posterior), in disagreement both conditions resulted in overconfidence ($t(47) > 8.34, p < .001$). In agreement trials, the Non-Interactive condition was significantly below zero - indicating underconfidence ($t(47) = -3.06, p = .003$) - but the Interactive condition was not ($p > .1$), suggesting that interaction made agreeing participants' confidence update indistinguishable from the normative update. The finding relates back to previous results indicating that people often stick with their original confidence rating, and that they are more confident in the Interactive condition than Non-Interactive condition.

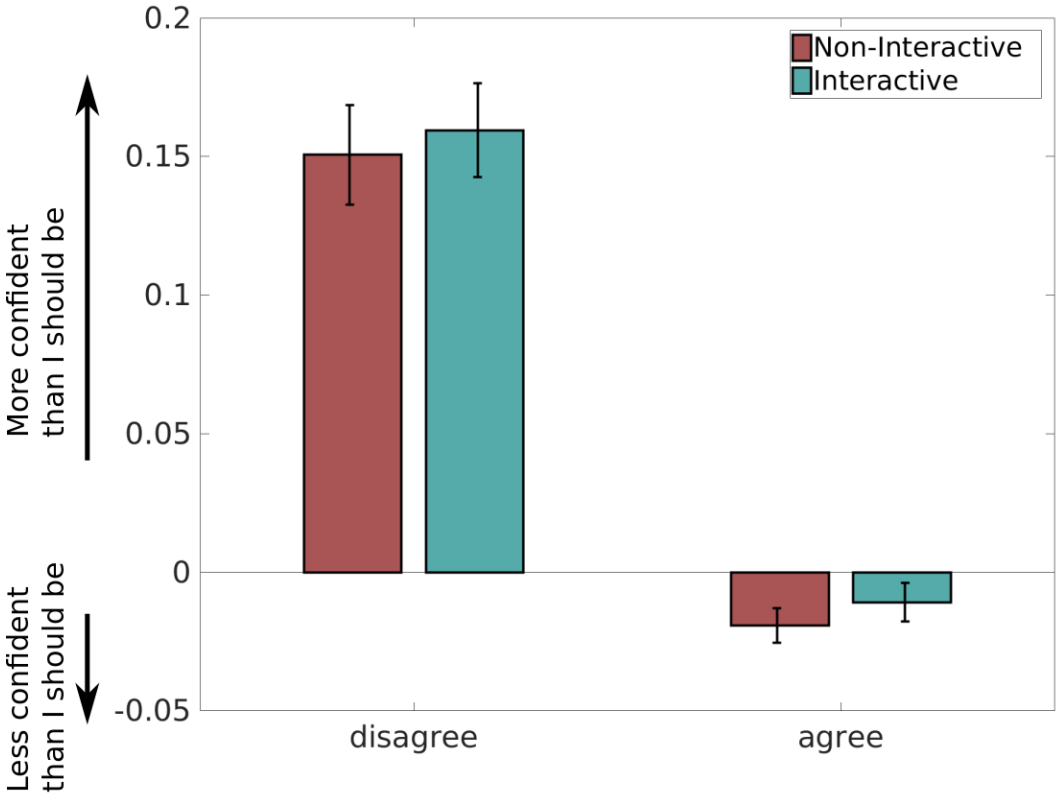


Figure 5.12: How confidence changes observed in the data compare with confidence changes expected according to a normative Bayesian update.

Root mean squared errors between the normative and empirical posteriors are shown in opinion space in Figure 5.13. The plot aims to give a qualitative description of the normative behaviour compared to the humans. The graph shows that normative optimality more accurately represents the dominant opinion's shifts rather than the dominated opinion ones, as indicated by the magnitude and extension of warmer areas in the latter (indicating larger errors). The representation along the opinion space allows us to understand in which trials participants depart from a Bayesian strategy the most. Differences between dominant and dominated trials are due to the fact that, as seen previously, individuals often stick with their initial judgment. This conservativeness, however, plays against individuals in dominated disagreement trials characterised by high conflict, as the normative framework prescribes here to follow the more confident view.

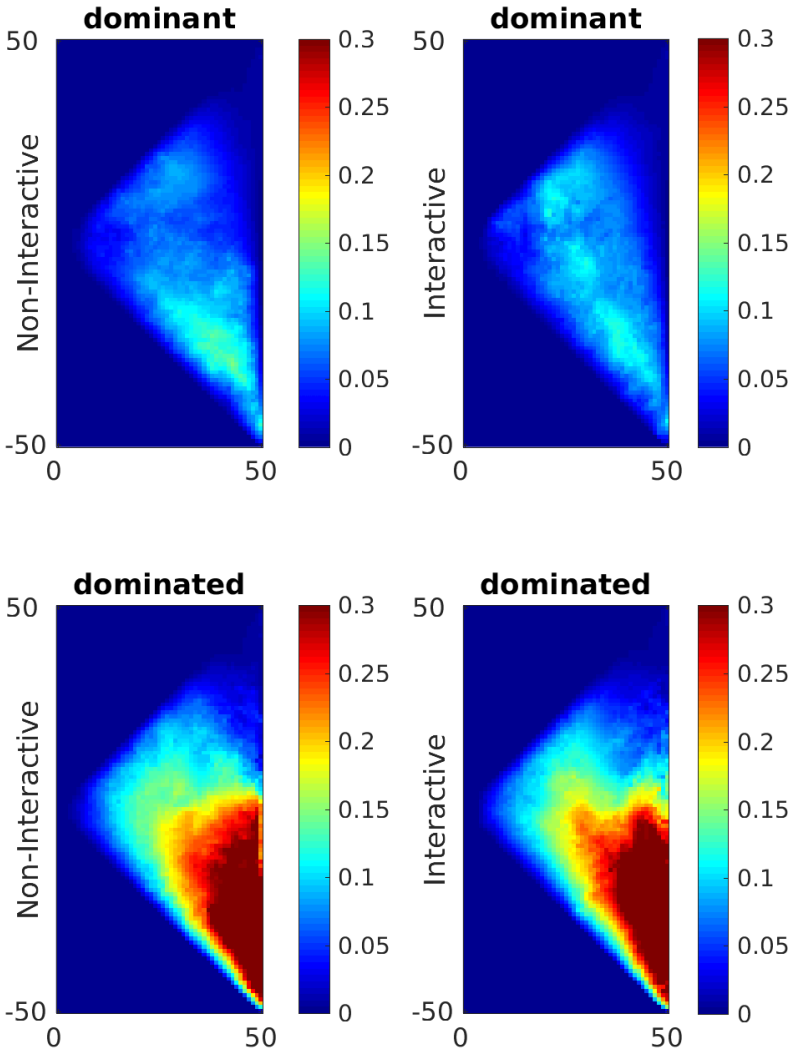


Figure 5.13: Residuals (RMSE) between Bayesian posterior confidence and human posterior confidence. The representation along the opinion space allows to understand in which trials participants depart from a Bayesian strategy the most.

Inferring social information perception with inverse Bayes

To understand what strategy human participants are adopting when receiving social information we reversed Bayes equation to infer participants' perception of the social information they were receiving. The standard Bayes equation is used by the optimal observer to infer the predicted posterior confidence given a prior confidence level P_s

and a partner's opinion P_p (equation 5.2). However, by solving the equation in P_p (i.e., the likelihood term), we can infer the *perceived* partner's confidence $\hat{P}_p$, given the participant's prior P_s and posterior confidence $post_s$:

$$\hat{P}_p = \frac{post_s(P_s - 1)}{2P_s post_s - P_s - post_s}; \tag{5.3}$$

In other words, we are asking: What perceived evidence can justify the observed participant's final confidence, given his/her initially stated confidence? Figure 5.14 shows, across all trials and all participants (pooled data), *perceived* partners' support (y-axis) as a function of partners' *stated* support (x-axis) and prior subjective confidence (colour), with 1 corresponding to social information that maximally agrees with one's own opinion and 0 to social information that maximally disagrees. A perfectly unbiased participant would have all points along the $y = x$ line, thus using their partner's social information as the partner themselves is stating it should be used.

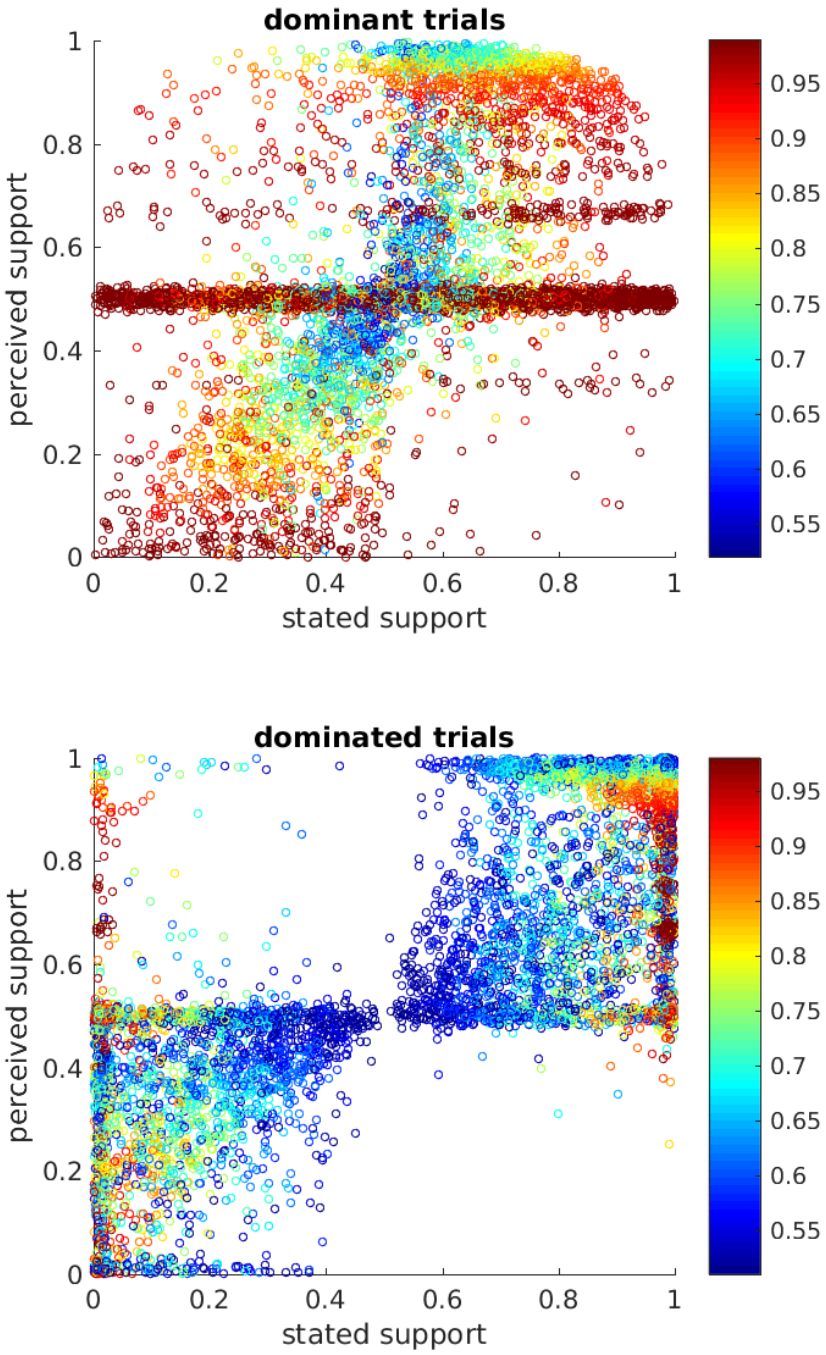


Figure 5.14: By reversing Bayes rule, the figure shows how much a partner's opinion is perceived to support one's own opinion compared to the partner's objectively expressed support. Colour represents participant's prior confidence.

It can be seen however that great variability exists across trials. As evidenced already in the confidence change distribution (Figure 5.8), many trials fall along the the $y = 0.5$ line, which means that in many trials people completely ignored the social information received. For agreement and disagreement trials separately and for each pre-social information confidence level, we counted the number of times participants ignored social information (i.e., the perceived evidence of a given participant fell into the [0.45 0.55] bin) and divided it by the total number of observations. As expected, the rate of ignoring social information was the highest when the participant started from a high pre-social information confidence (Figure 5.15). However, it did not seem to scale monotonically with initial confidence as a Bayesian normative framework would suggest. The probability of ignoring social information was also very high in disagreement trials when the participant started from the lowest possible confidence, thus when social information was maximally useful. A linear model with a quadratic term for pre-social information confidence bin was fitted to the data: $p(Ignore) = 1 + C_{pre} + C_{pre}^2$. Results showed that in disagreement both the linear ($b = -0.25, SE = 0.05, t = -4.70, p = .04$) and quadratic term ($b = 0.04, SE = 0.009, t = 5.19, p = .03$) were significant, explaining 87% of the variance ($AdjustedR^2$). In agreement trials on the contrary neither the linear ($b = -0.3, SE = 0.15, t = -1.93, p = .19$) nor the quadratic ($b = 0.06, SE = 0.02, t = 2.45, p = .13$) terms reached significance, indicating that in disagreement but not in agreement a quadratic term could describe the pattern of ignoring advice. A word of caution in interpreting the results must however be made. The analysis is purely exploratory as the pattern was not initially predicted by our theory. Post-hoc analyses may risk increasing the rate of false positives. However, exploring the data - particularly when phenomena are poorly understood - has the potential of creating new hypotheses and design future experiments.

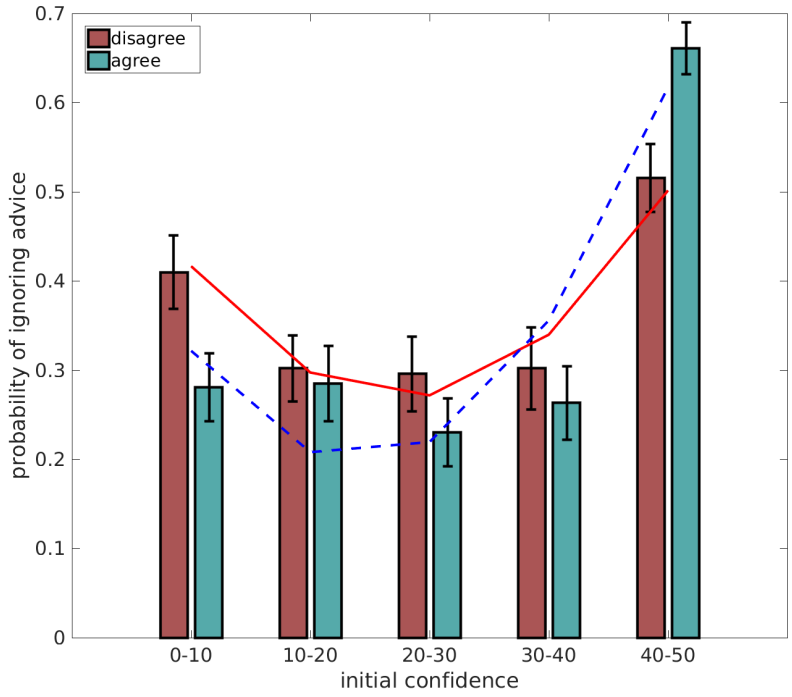


Figure 5.15: The bar plot shows the probability of ignoring social information as a function of initial confidence. Error bars represent s.e.m. Ignoring social information is defined as the number of trials perceived evidence was between 0.45 and 0.55 divided by the number of total observations in each condition. A quadratic relation between pre-social information confidence and probability of ignoring social information exists in disagreement but not in agreement trials.

We plotted the distributions of the partner's stated and perceived evidence for every participant, divided by trials in which the participant held the dominant vs. the dominated view. Comparing the two distributions provides insights into how empirical evidence (i.e., partner's stated support for the participant's opinion) gets distorted when perceived by the participant (i.e., actual use of partner's social information). We removed trials in the highest pre-social information confidence bin to avoid analysing trials when ignoring social information was actually driven by initial confidence. Results (Figure 5.16) show a stark dissociation between a partner's stated support and the participant's perceived support. Because social information given to the dominant member is, by definition, given with relatively low

confidence, the distribution of stated social information peaks around 0.5. Similarly, social information given to the dominated member necessarily peaks at high confidence levels - indeed the peaks were at the highest possible confidence levels of 0 and 1. Importantly however the perceived evidence for *both* members followed a trimodal distribution, peaking at 0.5 (neutral social information), 1 (strong evidence in favour), and to a lesser extent 0 (strong evidence against), suggesting a non-linear transformation of the received social information. In other words, a relatively homogeneous distribution of evidence (supporting subjective view) is converted into a semi-categorical distribution. To quantify this intuition I fitted on the agreement portion of the graph ($x > 0.5$) a linear model with a quadratic term for evidence bin: $frequency = 1 + evidence + evidence^2$. Results showed that the quadratic terms did not reach significance for partner's stated evidence (dominant: $b = 0.50, SE = 0.21$; dominated: $b = 0.22, SE = 0.46$), but were significant for the participant's perceived evidence (dominant: $b = 2.39, SE = 0.57, t = 4.19, p = .02$; dominated: $b = 3.84, SE = 0.56, t = 6.75, p = .006$), explaining 80% and 90% of the variance for dominant and dominated trials respectively ($AdjustedR^2$). It was further noticed that the dominant partner seems to overinterpret weak agreement, as indicated by the peak around 1; the dominated partner seems to discount strongly disagreeing social information, as indicated by lower blue bars compared to red bars, for evidence below 0.30.

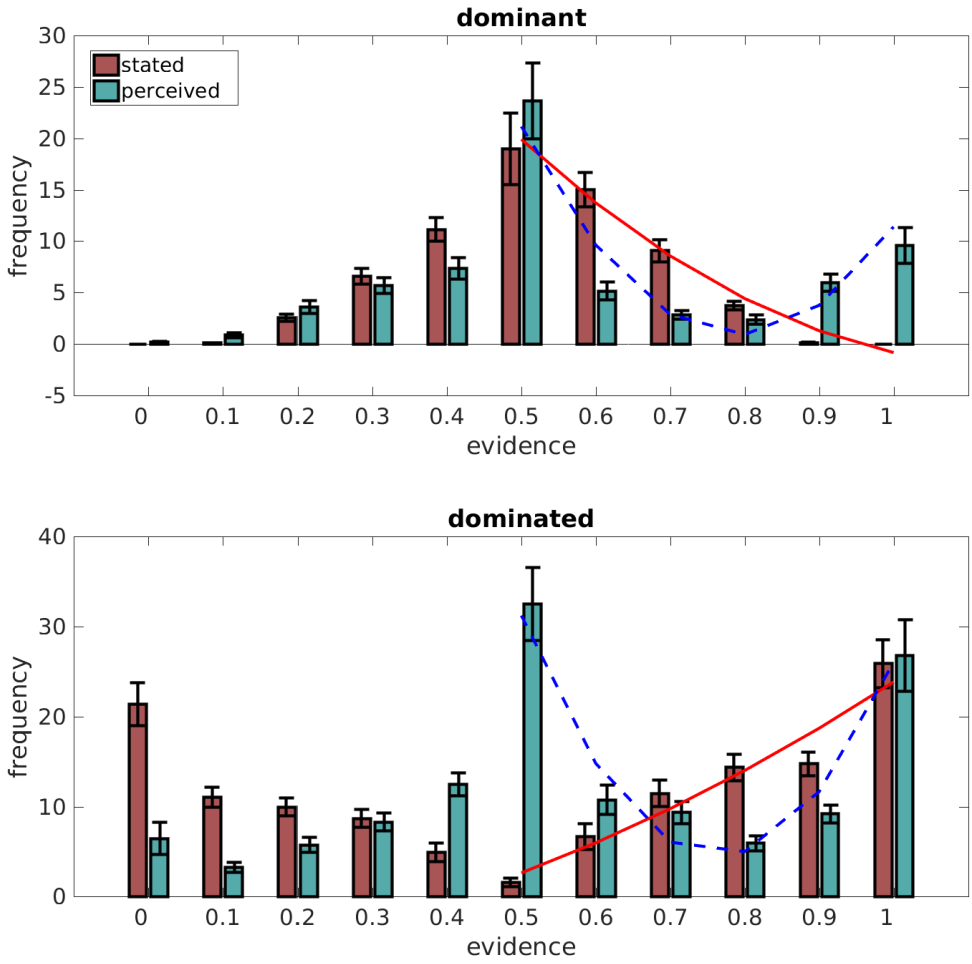


Figure 5.16: Contrast between partner's stated confidence in support for participant's independent view and participant's perceived support of partner's opinion, as inferred using inverted Bayes.

Egocentric and confirmation biases

Participants tended to ignore advice most of the time. This behaviour resulted in perceived evidence clustering around 50% (neutral evidence) and participants being close to rational optimality in agreement, but less so in disagreement (Figure 5.12 and 5.13). Egocentric bias is a phenomenon consistently observed in Judge-Adviser System studies consisting of applying different weights to own and others' opinion (Bonaccio & Dalal, 2006; Yaniv & Kleinberger, 2000). Our design also allows us to

look at conditions under which this bias is particularly strong.

To quantify the extent of egocentric bias I fitted a linear model on perceived evidence with only predictor partner's stated evidence (0=confident disagree; 1=confident agree). Regressions were fitted for each participant, dominance type and for agreement and disagreement separately. Regression lines were anchored at (0.5,0.5), so to obtain a bilinear transfer function from stated to perceived evidence. This extra degree of freedom allowed us to have different discounting factors for agreement and disagreement trials. Fitted coefficients α (i.e., slope in agreement trials) and β (i.e., slope in disagreement trials) represent the discount factors that participants apply to their partners' stated opinion, in agreement and disagreement trials respectively. A 3-way ANOVA on discounting factors with factors dominance, condition and consensus showed an effect of consensus ($F(1, 47) = 10.97, p = .001, \eta^2_G = 0.035$). This effect indicates that contradictory social information (i.e., disagreement) was discounted more than supporting evidence (i.e., agreement), a phenomenon known in literature as confirmation bias (Nickerson, 1998). No significant main effects of condition or dominance were found ($F < 1$) nor a significant interaction between the two ($F < 1$). Significant interactions between consensus and condition ($F(1, 47) = 10.05, p = .002, \eta^2_G = 0.003$) and consensus and dominance ($F(1, 47) = 19.79, p < .001, \eta^2_G = 0.03$) were found, indicating that Interactive condition tended to increase discounting in disagreement and to decrease it in agreement. The effect can be explained by the increased agreement effect and decreased disagreement effect observed in the Interaction condition (Figure 5.6). Figure 5.17 shows discounting factors divided by dominance, condition and consensus. Lower values indicate greater discounting (e.g., 0.5 means that the supporting evidence objectively provided by social information is perceived as halved). It can be seen that all values are below 1, indicating discounting of partner's views.

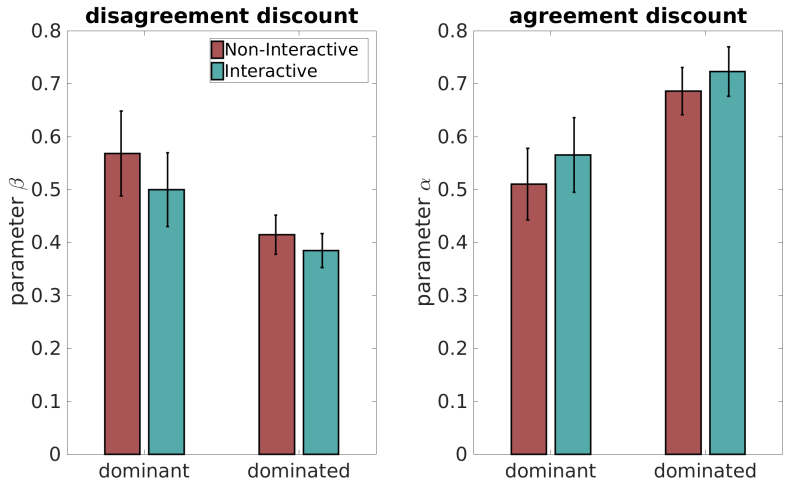


Figure 5.17: The bar plot above shows the discounting factors β and α, representing the discount that partner's social information, received on disagreement and agreement trials respectively.

Human discrepancy from the rational observer (Figure 5.13) was recomputed taking into account social information discounting and re-plotted in opinion space for dominant and dominated trials and for interactive and non-interactive conditions (Figure 5.18). As expected this formulation better describes participants' behaviour, compared to Figure 5.13.

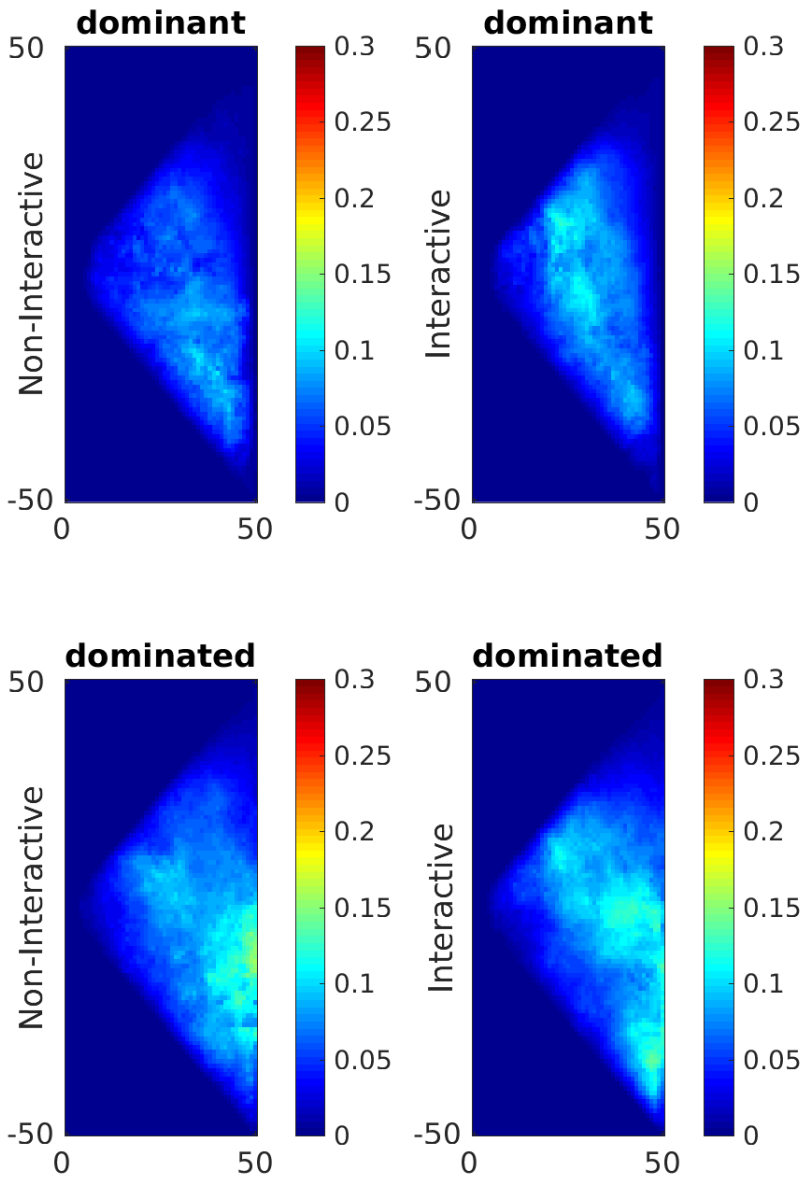


Figure 5.18: The graphs above represents the discrepancy between the nominal Bayesian update and the empirical human update, after taking into account egocentric and confirmation biases. Results are divided by dominant and non-dominant trials and by Interactive and Non-Interactive conditions.

Discussion

Experiment 4 aimed at comparing social exchanges characterised by one-step communication with social exchanges characterised by recursive interaction. Compared to traditional Judge-Adviser System experiments (Bonaccio & Dalal, 2006; Sniezek & Buckley, 1989), real-life social situations are characterised by real-time bi-directional dynamics (Schilbach et al., 2013). The experiment aimed to explore how a social system composed of people exchanging opinions is affected by the modality of social information communication over and above the content of communication itself.

The effect of condition on confidence change highlights the point that in conditions with equal information - in both the Interactive and Non-Interactive conditions, dyad members viewed the stimulus for 160 ms only - different opinion aggregation strategies can be observed depending on the communication modality between partners. Real-time dynamical interaction produced higher confidence changes in agreement and smaller confidence changes in disagreement. The reason why this happened can be explained by the breaking of independence between members' opinions. Confidence changes of the two participants became correlated during Interaction compared to a Non-Interactive baseline. In agreement, greater changes in one member corresponded to greater changes in the other. In disagreement, greater changes in one member corresponded to smaller changes in the other. The reason why these two effects coexist was made clearer by the trial-by-trial mixed-effects hierarchical model. The mixed-effects model showed that the opposite effect that condition had on absolute confidence change in agreement and disagreement trials was positively modulated by a member's partner's absolute confidence change. This suggests that different pieces of information were used to update one's own initial confidence in the two conditions. In Non-Interactive trials participants could make use only of their own and their partner's initial confidence. In Interactive trials on the contrary, participants had also access to the other person's confidence change. The significant interaction

found in the mixed-effect model between consensus, condition and partner's absolute change suggests that participants indeed made use of partner's absolute change to update their opinion, without taking into account that it was biased (and indeed generated) by their own judgment. As seen already in previous chapters, people seem to use a host of cues to infer the probability of an event (in our task the correct box), with little or no adjustment for systematic biases. Another potential cue that participants might have been exposed to in Interaction is the responsiveness of the partner's update. Although reaction times could not be defined during the social part given the continuous update window, estimation of the speed to update was estimated by fitting a sigmoid curve to the confidence update (see Appendix B for details). Results seem to indicate that partner's confidence change was a stronger predictor of the participant's change than partner's responsiveness was.

Phenomena of belief escalation have already been observed in regards to confidence (De Martino et al., 2013; Mahmoodi et al., 2013). These studies found that reciprocal social interaction makes people's judgments more risky and extreme, a result reminiscent of earlier findings on group polarisation (Myers & Lamm, 1976). The present study shows that a crucial cause of confidence escalation is the use of redundant information: participants should only use each other's independent opinions to arrive at a final decision because this is the only information that carries task-relevant value. However they also (incorrectly) use how much their own opinion is affecting their partners'. This creates dependencies that can potentially create non-linear dynamics in opinion aggregation. A side effect of such non-linear dynamics is, for example, an asymmetric distribution of irrational confidence changes, in agreement and disagreement trials. In particular, confidence increases after disagreement were more frequent than confidence decreases after agreement, suggesting that partner's changes of mind could sometimes be perceived as supporting evidence for one's views.

Plotting confidence changes in opinion space aided to distinguish which trial subsets the two conditions differed the most. Greater differences were observed in agree-

ment trials characterised by uncertainty and disagreement trials characterised by great imbalance between opinions' strengths.

Accuracy did not differ between the two conditions. Interaction did not affect confidence calibration overall, probably because confidence increases in agreement and decreases in disagreement happened both after correct and error judgments. According to a Noise Cancelling hypothesis of collective intelligence (Galton, 1907; Lorenz et al., 2011), social exchange should always lead to worse outcomes, because dependencies between individual measures increase noise correlation. According to a Confidence Sharing hypothesis however (Bahrami et al., 2010; Navajas, Niella, Garbulsky, Bahrami, & Sigman, 2017), interaction should have beneficial effects because confidence helps the group to wisely arbitrate between contrasting opinions. Based on this literature it was thus expected that accuracy in our experiment should have improved given that in both conditions participants shared their confidence levels. This is exactly what was observed: overall accuracy improved after social exchange. However, we also expected to observe smaller accuracy improvements in Interaction compared to Non-Interactive situations. This is because the dependency observed in interaction between confidence changes was expected to pollute participants task-relevant signal. However we did not observe any difference in accuracy improvements between the two conditions. Accuracy was operationalised both as choice accuracy ([0 1]) and as confidence in the correct answer to be sure to detect even sub-threshold improvements (Bonaccio & Dalal, 2006), however neither of these measures was affected by condition. A possible explanation is that although confidence increased when members agreed on the correct trial, it also increased when they agreed on the incorrect trial. Similarly, it decreased, unspecific to accuracy, following disagreement. Thus both accuracy improvement measures remained relatively unaffected by the presence of interaction, suggesting that dyad members seemed driven entirely by

their reciprocal confidences (and updates thereof) without any specificity to underlying objective accuracies or stimulus presentation. Confidence sharing thus seems to be a powerful strategy that benefits groups (Navajas et al., 2017) as well as individuals and that remains robust to phenomena of error cascades.

Presumably, however, the independence of partners' judgments protected accuracy from systematic errors and contributed to the absence of an effect of condition on performance. Arguably, if we had introduced a systematic bias across members' judgments, as it was the case in previous experiments, interaction might have led to greater errors than non-interaction.

The use of a Bayesian normative framework helped us better describe the mechanisms underlying confidence updates observed in participants. The model showed that although a normative probability update can describe well situations of agreement among participants, this is not the case when disagreement emerges between members. Further analyses showed that, particularly in disagreement, members of a dyad tended to discount each other's social information remaining excessively anchored to their initial opinions. Partner's opinion was heavily discounted ($\sim$ 80% to $\sim$ 60%) suggesting existence for an egocentric bias Yaniv and Kleinberger (2000). Moreover discounting was asymmetrical for agreement and disagreement, suggesting a confirmation bias where supporting evidence is treated differently from disconfirmatory one (Nickerson, 1998). Notice that, contrarily to Bahrami et al. (2010) who required participants to arrive at a joint decision, participants in the current experiment did not have to agree on one interval to complete the trial. This allowed to expand the original results from Bahrami et al. (2010) and to understand opinion change in situations when opinions do not get aggregated and participants "agree to disagree". No effect of condition was found on discounting parameters, as condition exerted opposite effects in agreement and disagreement trials.

An interesting result of the normative framework was shown by comparing empirical support for one's view received through social information (in probability terms) and inferred perceived support. By applying inverted Bayes inference, results evidenced that social information got distorted when being received by a participants. Contrary to empirical supporting evidence (i.e., partner's stated confidence for the participant's view), perceived evidence (i.e., use of social information) seemed to follow a bi- or tri-modal distribution, with peaks around 50%, corresponding to neutral social information, 100%, corresponding to maximally supporting social information, and to a lesser extent 0%, corresponding to maximally disconfirming social information. In other words, an originally homogeneous distribution of social information is converted into a semi-categorical distribution. An intriguing hypothesis is that people are solving a categorical inference problem. Instead of using continuous social information as it is provided by their social partners, participants are classifying each trial as "partner is wrong" vs. "partner is correct", and once this categorization is performed social information is used accordingly: the peak around 50% would then represent trials that the participant classified as "partner is wrong"; the peak around 100% represents agreement trials classified as "partner is correct"; and the 0% peak represents the much less numerous disagreement trials that the participant classified as "partner is correct". If correct, this explanation would suggest that participants try to minimize situations of uncertainty (e.g. 0.25 or 0.75 evidence), thus maximizing the impact of social information on final confidence.

Risk aversion. Subjective pre-social information confidence distributions were skewed toward the high confidence end of the scale. This hints to the fact that the payoff structure of the task might have made participants risk-seeking. The interpretation of confidence results with post-decision wagering scales is easier if participants are loss averse (Clifford, Arabzadeh, & Harris, 2008). Aware of these issues before starting recruiting, we tested participants for loss aversion using the coin gamble method

described in De Martino, Camerer, and Adolphs (2010), which allows to compute risk-aversion by quantifying the point of subjective indifference between losses and gains. An index of zero indicates that the subjective positive affect of gaining 1 reward unit (e.g., a pound) equals the subjective negative affect of loosing 1 reward unit (Kahneman & Tversky, 1979). A t-test across participants showed that risk-aversion indexes were significantly greater than zero ($t(41) = 6.19, p < .001$), after excluding six participants due to missing data, thus rejecting the hypothesis that participants did not show loss-aversion.

Conclusions

The present study is an attempt to capture non-linear dynamics that are potentially present in realistic social information sharing . This study is important in setting the limits of traditional Judge-Adviser Systems paradigms in studying social phenomena. It shows that not only the task-relevant information provided by social partners is important but also the modality in which information is shared and transformed across individuals. Real-time interaction as seen in most daily social exchanges is recursive and dynamic in nature so any static paradigm risks to miss important aspects of the phenomena under consideration. We showed that having access to one's partner's change in opinion (as opposed to partner's opinion alone) generates phenomena of confidence escalation in agreement trials, and reduced confidence reduction in disagreement. One of the symptoms of such non-linear dynamics is that irrational increases in confidence were observed during disagreement, contrary to what predicted by most models of opinion aggregation. The use of confidence sharing however allows both members to reach better performance than alone, irrespectively of condition.

In the following Chapter, the same paradigm is extended with two aims: (a) to replicate the results found in the current experiment, and (b) to make sure that the escalation effects found were not simply due to forgetting one's initial opinion.

We therefore included confidence anchors to remind each participant where their own confidence (Experiment 5) and their partner's confidence (Experiment 6) initially started from.

6

Is it memory or interaction?

'The most interesting information comes from children, for they tell all they know and then stop.'

– Mark Twain

Chapter Abstract

Two experiments are conducted to extend the same paradigm introduced in Experiment 4. The confidence escalation observed in Experiment 4 could in principle be due to the fact that in the Interactive condition participants updated their confidence multiple times because they failed to remember what confidence judgment they had initially given. Adding a confidence reminder to this condition should rule out this possibility and provide evidence for a genuine interactive effect. In Experiment 5 a "self reminder" condition (Interactive$_{self}$) is added to the existing conditions, in which the Interactive condition is enhanced with the presence of a static confidence anchor reminding participants of their private initial opinion. Experiment 6 tests the memory hypothesis one step further by comparing the baseline Interactive condition with two reminder conditions: (1) the interactive condition with self-reminder (Interactive$_{self}$), already described for Experiment 5 and (2) the interactive condition with other-reminder (Interactive$_{other}$), which reminds participants of *their partners*' initial opinions. Results of both experiments suggest no large effect of reminders,

refuting the memory explanation for the interactive effect. Moreover, in both experiments key results from Experiment 4 were robust to the introduction of a new incentive scheme for giving calibrated confidence judgments, suggesting that the effects found in Experiment 4 were not dependent on the specific confidence scale used.

Experiment 5

Introduction

Experiment 4 showed several differences emerging between two experimental conditions that were identical in terms of participants' access to decision-relevant information but that differed in terms of how that information was communicated between the two decision-makers. When two partners were allowed to interact dynamically and adjust their confidence changes in light of the other's updates, differences emerged primarily in the magnitude of confidence changes and the independence of confidence updates. The experiment found evidence for positive confidence escalation (Mahmoodi et al., 2013) when real-time interaction was allowed: In the agreement trials of the Interaction condition, confidence change magnitudes between two members of a same dyad were positively correlated, suggesting that the more a member increased his/her confidence the more their partner increased his/her own. In other words in the Interactive condition participants were not only using their partner's stated initial opinions but also their opinion change. These effects were not observed in the Non-Interactive condition.

A simulation presented in Chapter 5 (Figure 5.9) modelled interaction as the recursive integration of *current* partner's confidence and *initial* personal confidence. Another simple phenomenon however could be at the heart of the interaction effects found in Experiment 4: Perhaps participants in the interactive condition simply tended to forget what initial confidence judgment they had provided and were instead updating the *current* confidence of their partner with their own *current* confidence. Modifying the original simulation accordingly (not shown here) easily shows that this strategy quickly leads confidence of both participants to escalate towards the maximum confidence boundary on the side of the most confident initial opinion. The longer time taken in the Interactive condition for the confidence update to reach

a stable state (Figure 5.4) is also compatible with confidence being updated using partner's and own current confidence state. Longer times could be a by-product of participants updating their current confidence - or a distorted representation of their initial confidence - with the current opinion of their partner, because of the fact that they are engaged in multiple updating instead of a one-step single update.

To test whether the effect of interaction found in Experiment 4 was due to failures in remembering one's own initial confidence, a third experimental condition was created and compared to the previous two. In this new condition (called Interaction$_{self}$) a static reminder of one's own pre-social information confidence is presented on the scale along with the standard personal and partner's cursors typically presented during the Interactive condition. If the effects of interaction are only due to memory failures then the presence of a reminder should make those same effects disappear. Failure to reduce the interaction effects should be taken as evidence that differences between interactive and non-interactive conditions are not due to forgetfulness.

A worry from Experiment 4 was that people often used extreme values when rating their initial confidence, which was likely due to the post-decision wagering method to report confidence (Clifford et al., 2008). We thus introduced different instructions regarding the input of confidence ratings, incentivising confidence calibration over confidence magnitude. This gave us the opportunity to assess the robustness of key Experiment 4 effects with a different confidence scale.

Methods

Participants. Twenty-four dyads (14 female dyads, 1 mixed gender dyad) were tested. Mean age was 23.16±3.42. Participants were recruited online using the University volunteers platform and local advertisement websites. All participants gave informed consent before starting the experiment. The study was approved by local ethical committee.

Paradigm. The experiment comprised of 432 experimental trials divided in 18 blocks and 20 practice trials divided in 4 blocks. Practice blocks were designed to practice with the first-order task, the non-interactive condition, the interactive condition, the interactive plus reminder condition respectively. The methods were very similar to those used in Experiment 4, in regard of the dot-count task, trial stages and input modalities, with the following key differences. First, three conditions were defined by manipulating the access participants had to their partner's information: the two conditions already presented in Experiment 4 and a reminder condition. Conditions were varied within-participants across blocks (i.e. six blocks per condition). Participants experiences six identical modules, each comprising the three different conditions into three separate blocks. The order of the three conditions within a module was randomised across dyads. Second, the social part window was reduced to 4 seconds (21 data points), given that most updates in Experiment 4 occurred within 2 seconds of the social part. Third, it was decided to change the incentive system used for Experiment 4 and the instructions given to participants to use the confidence scale. This modification was motivated by two main reasons. The first reason was to make participants' confidence distributions less extreme and more uniform across the scale. Although in Experiment 4 all participants showed some evidence of loss aversion, confidence judgments were skewed toward the high end of the scale, creating potential issues in detecting small confidence changes in this direction (i.e., confidence increases) due to ceiling effects. The second reason was to check whether the effects found in the previous experiment were robust to changes in the incentive system and thus in the use of the confidence scale. Failing to reproduce Experiment 4 results when changing the incentive system would be a strong indication that they were (at least partially) dependent on the specific instructions participants received. Details about how the new incentive scheme worked and about the instructions given to participants are described in the paragraph "Incentive scheme".

Manipulation. Three conditions were defined that affected only the social part of the trial. A Non-Interactive and an Interactive conditions were defined as in Experiment 4, which allowed us to see if those effects replicated. An Interactive plus self-reminder condition (Interactive$_{self}$) was constructed so that participants were shown a reminder of their own initial private confidence during interaction. The reminder was presented as a static gray shaded cursor.

Notice that in all conditions the social part started exactly with the same initial configuration of objects on the screen and cursors were presented in the same position as they were left at the end of the private part. Any difference among conditions must then be attributed to the specific communication channels that each condition entails, assuming equal initial conditions of the dyad state. Conditions alternated regularly over blocks (six repetitions each) and their order was shuffled across participants.

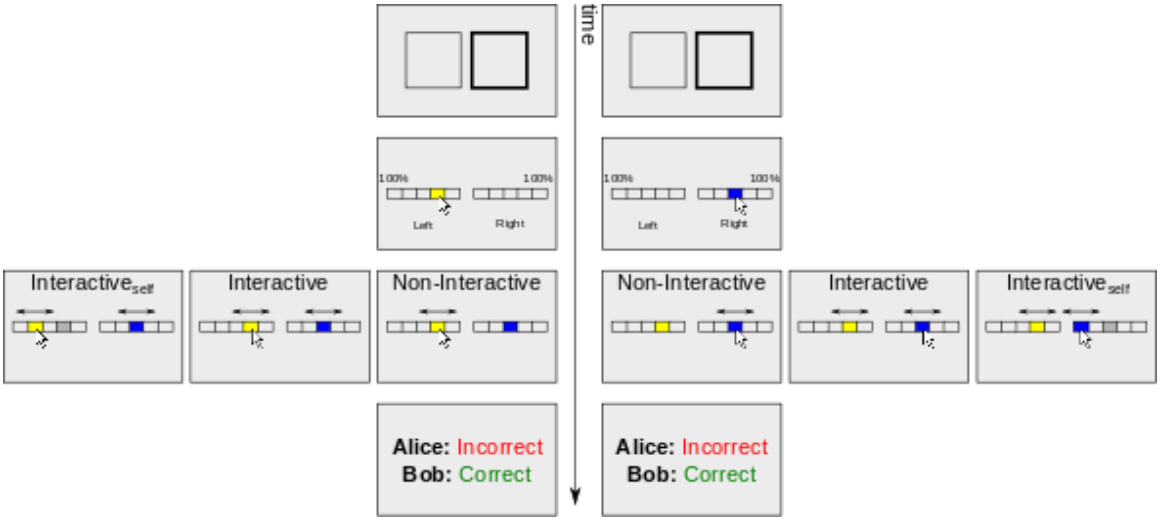


Figure 6.1: Experimental paradigm implemented in Experiment 5. Three conditions are explored and compared within-participants. During the Non-Interactive condition participants are shown the initial independent opinion of their partner. During the Interactive condition participants are shown the current opinion of their partner in real-time. During the Interactive plus self-reminder condition (Interactive$_{self}$) participants are shown the current real-time opinion of their partner and are at the same time reminded of their own original opinion as a shaded cursor on the scale. This manipulation makes sure that if participants update their initial confidence they are constantly reminded of where along the scale they started from. In all conditions participants have four seconds when they are asked to track their confidence state in real-time. The confidence scale that was actually used had 50 levels per interval.

Incentive scheme. In the current experiment participants were informed that their final reward would be inversely proportional to the average absolute deviation of their accuracy from the calibration line. The calibration line was defined by the line $y = x$, i.e. where confidence expressed in percentage points is identical to the probability of a correct response. Instructions stated: "We will average all trials when you were 60% confident and see if you were indeed 60% accurate. Then we'll see if you were 70% accurate on trials where you said you were 70% confident and so on. The higher the discrepancy the less you will get.". Importantly participants were told that during the social part this measure was computed on a moment-by-moment basis and that the best strategy to maximise their gains was thus to continuously update their confidence cursor based on their internal sense of confidence.

For this calculation, at the end of each block the confidence distribution of each participant was divided into 5 bins and the weighted average absolute distance between bin accuracy and bin center was taken as a calibration error:

$$Err = \frac{\sum_{b=1}^{5} |Acc_b - Conf_b| * N_b}{\sum_{b=1}^{5} N_b} \tag{6.1}$$

where N_b is the total number of data points recorded in each bin. Err was computed for pre-social information and post-social information separately and the two were averaged together so that an equal weight was given to private and social parts. Importantly the formulation above computes the calibration error on each data point collected - i.e. 1 for pre-social information and 21 for post-social information decisions. This ensures that the error during the social part is a weighted average among bins based on the time spent in each one.

Results

Continuous update. During the social part of each trial, the x-position of the cursor along the confidence scale was recorded every 200 ms, giving 21 confidence data

points over the course of 4 seconds. The absolute difference between a data point and the previous one can be used as a measure of the stability of the confidence updates over time, with smaller numbers indicating that participant's updates have stabilised. This update stability measure is shown in Figure 6.2 for the three different conditions separately. It can be seen that in all conditions the larger confidence update occurred around one second from the start of the social part. Both interactive conditions showed larger updates on average around this period, followed by longer times to reach an equilibrium as suggested by the larger right tail.

The right panel of Figure 6.2 shows the difference between the two interactive conditions (i.e. Interactive and Interactive$_{self}$) and the non-interactive baseline condition. The hypothesis that interaction leads to a sequence of dynamic updates and thus to longer times to converge was tested using a one-tail t-test. Coloured areas under the curve represent uncorrected significant differences ($\alpha = .05$). The differences did not however survive a cluster-based permutation t-test, indicating that the different conditions were not statistically different from one another.

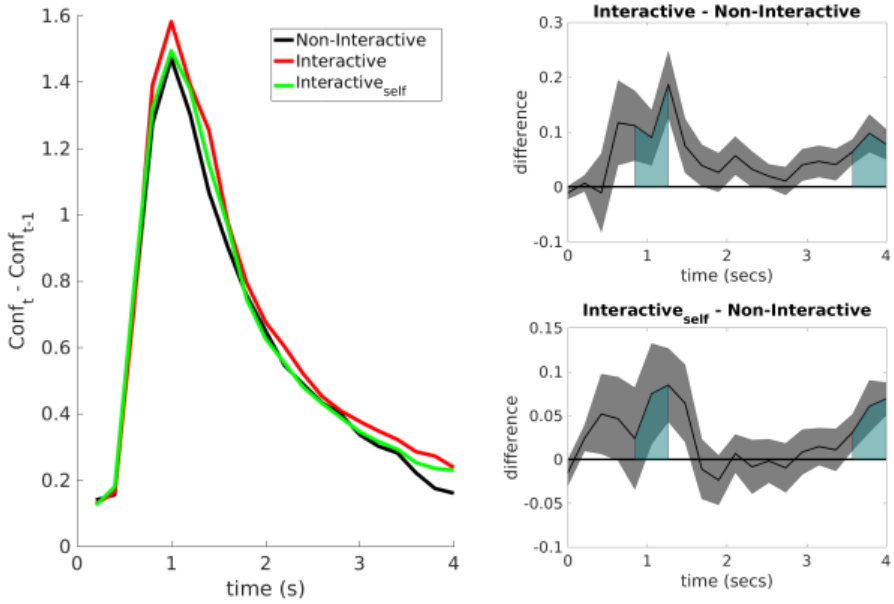


Figure 6.2: Average confidence difference between two consecutive data point recorded during the social window. The higher the difference the bigger the update. It can be observed that in all conditions the biggest updates are observed around the first second of the social part. Both interactive conditions show a larger update around the same time compared to the non-interactive baseline condition and a longer time to reach an equilibrium. The right panel shows the difference between each interactive condition and the non-interactive baseline condition. Areas under the curve represent uncorrected significant differences ($p < .05$). Error bars represent s.e.m.

Confidence distributions. Figure 6.3 shows the individual distributions of pre-social information confidence ratings (1: lowest rating; 50= highest rating). It can be seen that compared to Experiment 4 confidence distributions are on average less clustered towards the right, suggesting that the new incentive scheme was successful in making participants' use of the scale more uniform. It can also be seen that for some participants confidence clustered around the confidence landmarks provided. Although not ideal, this behaviour is unlikely to be a confound for subsequent analyses.

In Experiment 4, a marginal difference was found between the two conditions' average pre-social information confidence. We thus tested whether this difference

could also be found in Experiment 5. The three conditions did not differ in terms of the average pre-social information confidence of the participants ($F < 1$), suggesting that the difference observed in Experiment 4 was probably due to random noise.

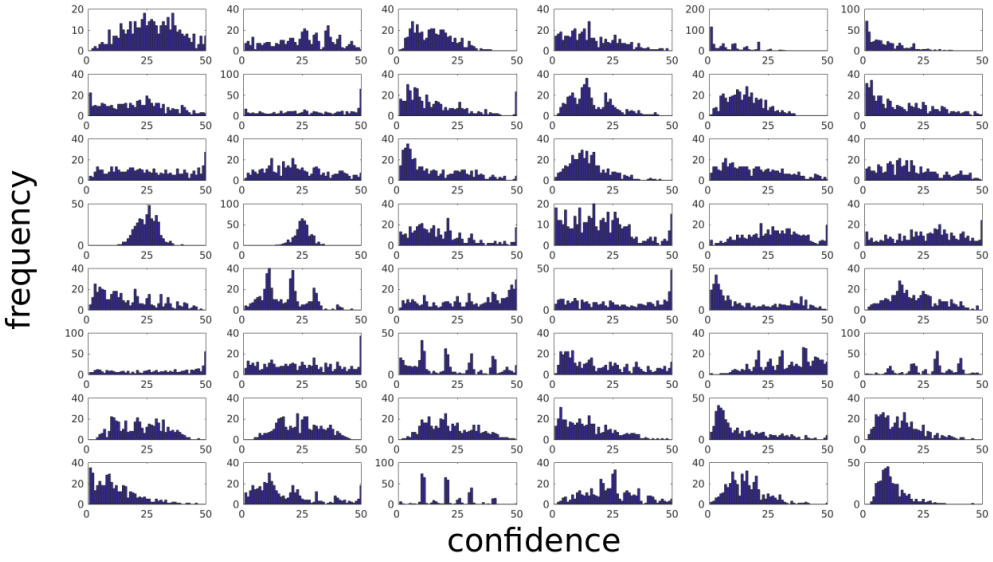


Figure 6.3: Confidence distributions of the 48 participants tested. Each row plots the participants of three different dyads. From top-left to bottom-right, plots one and two corresponds to the first dyad's participants, plots three and four to dyad two etc.

Confidence changes. Experiment 4 showed that interaction significantly increased confidence increases in agreement and decreased confidence decreases in disagreement. A measure of social information influence was defined as in Experiment 4 as the confidence change from pre- to post-social information: $\delta_C = C_{post} - C_{pre}$. Positive values represent confidence increases relatively to pre-social information confidence and negative numbers represent confidence decreases. A two-way repeated measures ANOVA on δ_C with factors consensus (agreement vs disagreement trials) and condition was run to replicate the results found in Experiment 4. It showed a significant effect of both consensus ($F(1, 47) = 158.03, p < .001, \eta^2_G = .68$) and condition ($F(2, 94) = 14.05, p < .001, \eta^2_G = .009$). The main effect of consensus, shown in Figure 6.4, unsurprisingly shows that participants tended to increase their confidence in

agreement and decrease it in disagreement. Importantly, however, confidence changes observed in the three conditions were significantly different when averaged across consensus: both interactive conditions showed more positive confidence changes than the non-interactive baseline condition ($t(47) > 4.10, p < .001, d > 0.35$), explaining the significant main effect of condition. Moreover a significant interaction between the two factors ($F(2, 94) = 5.19, p = .007, \eta^2_G = .002$) was found.

In agreement trials, confidence increased more in both Interaction conditions compared with the baseline Non-Interactive condition ($t(47) > 4.10, p < .001, d > 0.20$), replicating the results of Experiment 4. The two interaction conditions themselves differed ($t(47) = -2.16, p = .03, d = 0.08$). In disagreement, contrary to Experiment 4, confidence decreases were numerically but not significantly smaller between the Non-Interactive and Interactive conditions ($t(47) = 1.07, p > .2, d = 0.09$). Similarly they were non-significantly smaller in the Interactive$_{self}$ compared to Interactive condition ($t(47) = 1.58, p = .11, d = 0.10$). Confidence decreases differed however between the Non-Interactive and Interactive$_{self}$ conditions ($t(47) = 2.47, p = .01, d = 0.19$).

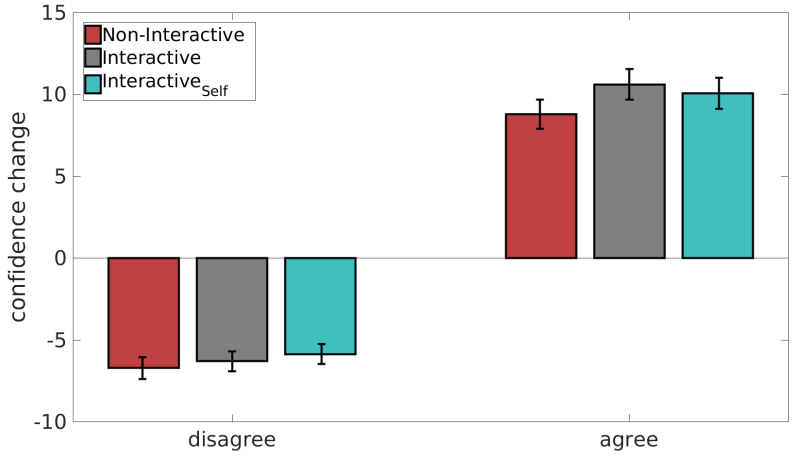


Figure 6.4: Confidence change from pre- to post-social information across conditions, separately for trials in which the dyad members disagreed vs. agreed. Positive numbers represent increases in confidence after social information exchange while negative numbers represent decreases.

Asymmetry in confidence increases. Figure 6.5 plots confidence change distributions for each condition, divided by consensus and averaged across participants. It can be clearly seen that no clear difference emerged among conditions and that all distributions peak at zero, suggesting that most frequent confidence update was to not update.

Right-tails in disagreement and left-tails in agreement represent irrational confidence changes. A two-way repeated measures ANOVA on the probability of an irrational change (corrected for total number of agreement and disagreement trials and trembling hand issues) showed a significant effect of consensus ($F(1, 47) = 16.98, p < .001, \eta^2_G = .12$) but not of condition ($F < 1$) and no significant interaction between the two ($F(2, 94) = 1.25, p = .28, \eta^2_G = .002$), suggesting that irrational increases were more frequent than irrational decreases (M±STD: irrational increases = 0.0166 ± 0.022 vs. irrational decreases = 0.0031 ± 0.004), but no consistent differences were found among conditions.

The results partly replicate what found in Experiment 4, suggesting that irrational changes are more frequent after disagreement than after agreement. Experiment 5 does not however replicate the finding that irrational increases were more frequent in the Interactive than the Non-Interactive condition, suggesting that perhaps this result was an effect of a different use of the confidence scale.

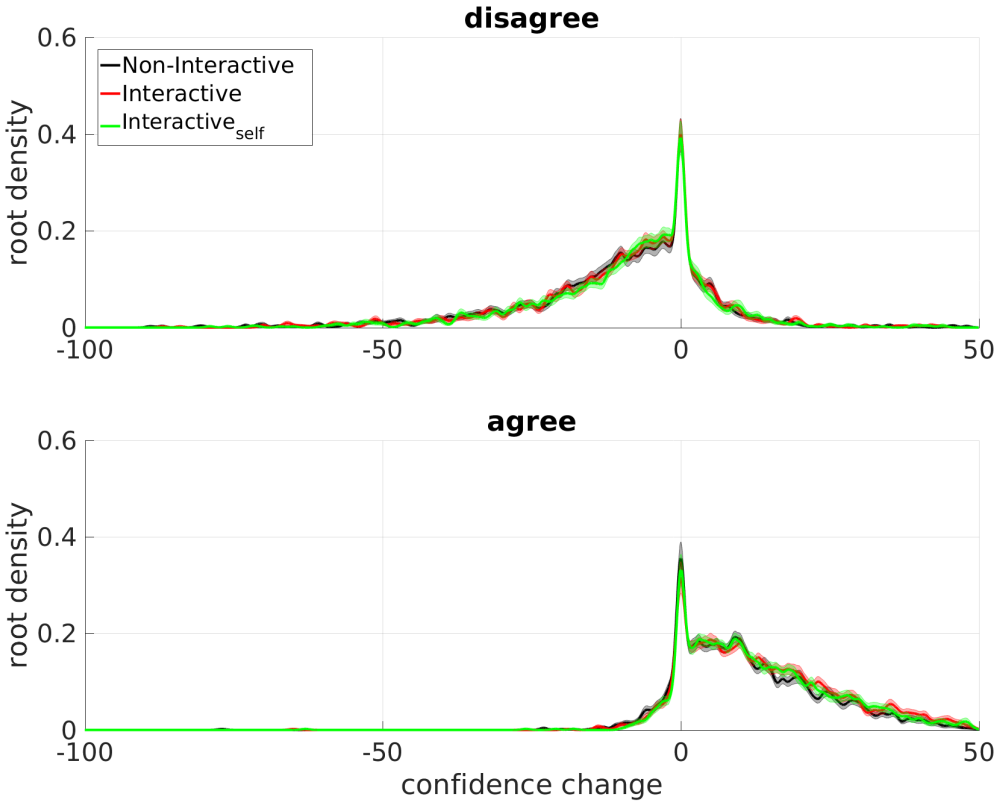


Figure 6.5: Confidence change distributions observed in the most confident participants divided by condition and consensus. Plots represent estimated probability density functions using a normal kernel method (bandwidth = 0.50). Error bars represent s.e.m.

Influence in opinion space. Visualising confidence changes along the opinion space can better represents participants' behaviour during the update window. Median confidence changes δ_C were plotted in opinion space to understand which subsets of trials (i.e. which initial conditions) showed larger confidence changes and which ones showed the strongest difference between experimental conditions. Confidence changes were plotted separately for dominant and dominated trials. Figure 6.6 shows the opinions surfaces so obtained, together with the contrast plots obtained by subtracting the Non-Interactive baseline from the Interactive and the Interactive$_{self}$ conditions (panels D-E and I-L). The contrast plots help us understand for which subsets of trials two conditions differed the most.

Two major areas of interest were identified in Experiment 4, one corresponding to weak agreement (participants are both unsure but happen to agree) and the other corresponding to unbalanced disagreement (one participant is very confident while the other weakly disagrees). In Experiment 5, similar areas of interest emerged. In both dominant and dominated trials, participants in interactive conditions showed larger confidence increases compared to a Non-Interactive baseline after weak agreement (warmer colours in correspondence of x-points of plots D,E,I,L). The magnitude of the increase in these areas, indicates that in interaction participants converged on high confidence agreement. A real-time animation of the density distribution of dyad states during the 4-second update window, as well as an animation of the contrast between conditions, can be found at `https://niccolopescetelli.com/confidence-change-in-opinion-space/`. The animated contrast plot shows that, although in the two conditions dyad states were similarly distributed along the opinion surface at the beginning of the update, more trials in the interactive conditions than in the Non-Interactive one gravitated towards point (50,50).

Imbalanced disagreement trials (labelled on the contrast plots by y) also seem to differ between interactive and non-interactive conditions, although the lower number of disagreement trials makes contrast plots in these regions more noisy. Again, points y warmer colours on plots D-E and points y colder colours on plots I-L, suggest that in dominant and dominated trials respectively, participants decreased their confidence less in interaction than in a Non-Interactive condition. Experiment 5 also showed that a third area of interest, labelled z on contrast plots, was balanced disagreement, a situation where the two participants disagree with equal high confidence. However, the fact that disagreement was less frequent than agreement combined with the less extreme use of the scale (Figure 6.3), makes data in this region sparse and potentially difficult to interpret.

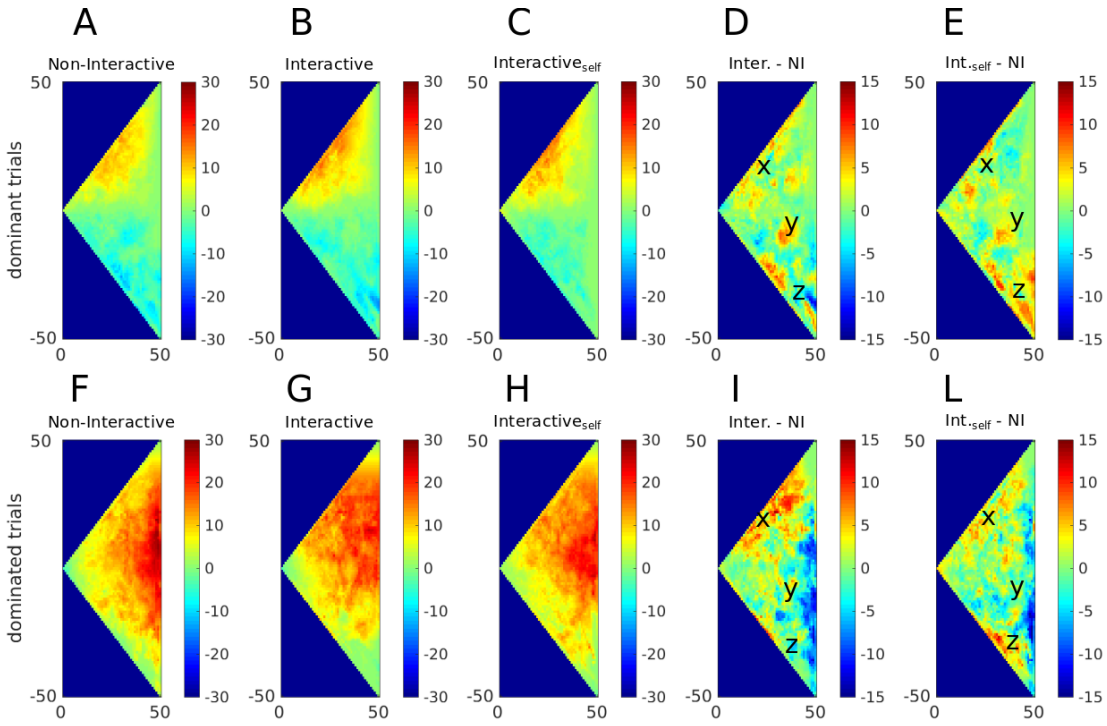


Figure 6.6: Median confidence change in opinion space divided by condition and dominance (first three columns). Warmer colours represent confidence changes in the direction of the dominant opinion, while colder colours represent confidence changes further away from it. The figure also shows contrast plots between the interactive conditions (Interactive and Interactive$_{self}$) and the Non-Interactive condition in the two rightmost columns.

The analyses above are primarily qualitative and the scatteredness of the data did not, unfortunately, allow more quantitative comparisons. They are however important for a number of reasons. First, they allow us to visualise confidence change in a multivariate way, without the need to average over trials that are not formally identical. Second, they allow us to understand which subsets of trials are similar across conditions and which ones are not, making it easier to determine what effects the manipulation has on behaviour. Third, they can inform subsequent analyses by restricting the trials of interest to trials that are likely to generate the effects observed.

Coupling of confidence changes in interaction. Experiment 4 showed that interaction produced positive correlation in dyad members confidence changes under

agreement and negative under disagreement. We thus tested whether the results replicated here. Figure 6.7 shows the average Pearson's r coefficient, divided by condition and consensus. Coefficients were entered into an ANOVA across dyads with factors condition and consensus. Results show that both condition ($F(2, 44) = 16.35, p < .001, \eta_G^2 = .12$) and consensus ($F(1, 22) = 13.16, p = .001, \eta_G^2 = .09$) had a significant effect on the correlation observed. The interaction between the two terms was also significant ($F(2, 44) = 27.01, p < .001, \eta_G^2 = .09$). No correlation was found in any of the three conditions in disagreement ($t(23) < .8, p > .4$). On the contrary in agreement both interactive conditions showed positive correlation coefficients ($t(23) > 4.5, p < .001$) while coefficients in the Non-Interactive condition were not significantly different from zero ($p > .1$). The results partly replicate results observed in Experiment 4. Similarly to Experiment 4, Experiment 5 indicated that confidence changes of members of the same dyad remained independent from each other in the Non-Interactive condition and interaction introduced positive correlations between confidence changes in agreement trials, with no difference found between interactive conditions ($p > .2$). The negative correlation found in disagreement trials in Experiment 4 between same dyad members was however not replicated in Experiment 5.

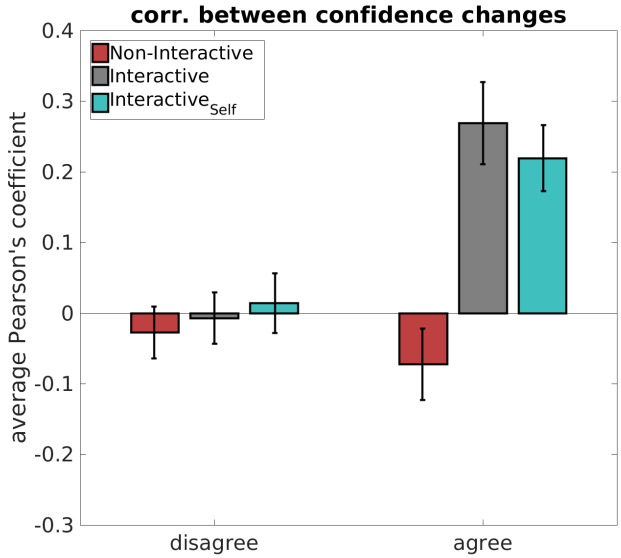


Figure 6.7: Coupling between absolute confidence updates of the two participants across different conditions and divided by consensus. It can be seen that in disagreement updates of one dyad member are not correlated with updates of the other member. In agreement on the contrary a positive correlation emerges as soon as participants are allowed to interact in real-time.

Experiment 4 showed, using a mixed-effects linear regression, that the coupling of confidence changes observed in the interactive conditions were likely explained by participants making use of partner's update magnitude to update their own confidence. The same mixed-effects model was applied to Experiment 5 data and largely replicated the main effects found there. The interaction terms between both interactive conditions and consensus were significantly above zero (Interaction: $\beta = 0.17, SE = 0.03, p < .001$; Interaction$_{self}$: $\beta = 0.20, SE = 0.03, p < .001$), indicating that during interaction agreement led to greater confidence updates compared to a Non-Interactive reference. Importantly, both terms were positively modulated by partner's absolute confidence change (Interaction: $\beta = 0.37, SE = 0.03, p < .001$; Interaction$_{self}$: $\beta = 0.26, SE = 0.03, p < .001$), replicating the finding that in interaction participants tended to make use of their partner's confidence changes to inform their own confidence updates.

Performance analysis. Experiment 4 had found a significant benefit of social exchange, but no significant differences between conditions according to different measures of performance, including choice accuracy, graded accuracy and confidence calibration. In Experiment 5, a two-way ANOVA on choice accuracy with factors condition and decision time (pre-social information vs. post-social information), showed a significant effect of decision time ($F(1, 47) = 101.66, p < .001, \eta^2_G = .17$), replicating the finding that choice accuracy significantly improved from pre- to post-social information phase (M: 0.71 vs. 0.74), but no significant difference of condition and no significant interaction ($F < 1$). These results contradict a Noise Cancelling hypothesis, which predicts that any increased dependence among judgments should hamper collective performance (Lorenz et al., 2011).

Similarly to Experiment 4, a graded measure of accuracy, i.e., confidence change towards the correct answer (Bonaccio & Dalal, 2006), was also considered to control for sub-threshold changes of mind. The same two-way ANOVA implemented for choice accuracy showed a significant effect of decision time ($F(1, 47) = 122.53, p < .001, \eta^2_G = .29$) and condition ($F(2, 94) = 3.26, p = .04, \eta^2_G = .01$), indicating greater confidence in the correct answer in the post-social information phase (M: 10.81 vs. 15.85) and larger values in both interactive conditions compared to Non-Interactive condition (means: Interactive = 12.71, Interactive$_{self}$ = 13.57, Interactive$_{other}$ = 13.72, $t(47) > 1.98, p < .06, d > 0.24$), confirming that interaction did not negatively affect graded accuracy. No significant difference was found between interactive conditions ($t < 1$), suggesting that the presence of a confidence reminder did not affect the effect of interaction. Finally the ANOVA also showed a significant interaction between decision time and condition ($F(2, 94) = 11.88, p < .001, \eta^2_G = .003$), indicating larger improvements for interactive conditions over Non-Interactive one ($t(47) > 4.01, p < .001, d > .26$).

Confidence calibration (measured as type II A_{ROC}) was one last measure of performance considered (Fleming et al., 2014). A two-way ANOVA showed an effect

of decision time ($F(1.47) = 94.08, p < .001, \eta_G^2 = .26$), indicating that calibration significantly improved after exchanging social information (0.56 vs. 0.62), but no effect of condition ($F < 1$), indicating that neither interaction or the presence of an anchor negatively affected calibration. A significant interaction term was also found ($F(2, 94) = 3.53, p = .03, \eta_G^2 = .02$), indicating differences in improvement across conditions. Pairwise comparisons showed that calibration improved significantly more in the Interactive compared to Non-Interactive condition ($t(47) = 2.53, p = .01, d = 0.38$). No significant difference in calibration improvement was found between interactive conditions nor between Interactive$_{self}$ and Non-Interactive condition ($p > .1$).

Overall, decision performance improved after social exchange and increased dependency between judgments through interaction did not hamper improvement, but instead, if anything, fostered it. The following section will explore deviations of participants' behaviour from a Bayesian normative framework.

Humans show overconfidence compared to Bayes. Confidence changes observed in empirical data were compared to changes predicted by a normative Bayesian opinion integration strategy with equal weights for self and other's opinion. Similarly to what found in Experiment 4, residuals between observed and predicted confidence changes (Figure 6.8) deviated from Bayes norm in disagreement more than agreement trials. A two-way ANOVA with factors consensus and condition showed a significant effect of consensus ($F(1, 47) = 74.08, p < .001, \eta_G^2 = .54$) and condition ($F(2, 94) = 9.50, p < .001, \eta_G^2 = .004$) and a significant interaction between the two factors ($F(2, 94) = 3.69, p = .02, \eta_G^2 = .001$). Both interactive conditions led to more positive deviations (suggesting greater post-social information confidence than predicted by the normative framework) than Non-Interactive condition ($t(47) > 3.4, p < .01, d = 0.30$). No difference was found between the two interactive conditions ($p > .8$). The interaction term indicates that the difference between agreement and disagreement residuals was larger for the Non-Interactive condition

compared to the Interactive condition ($t(47) = 2.64, p = .01, d = 0.11$) but not to the Interactive$_{self}$ condition ($p > .3$). Only a marginal difference was found between interactive conditions ($t(47) = 1.75, p = .08, d = 0.07$), suggesting that the addition of a confidence reminder little affected the Interactive condition.

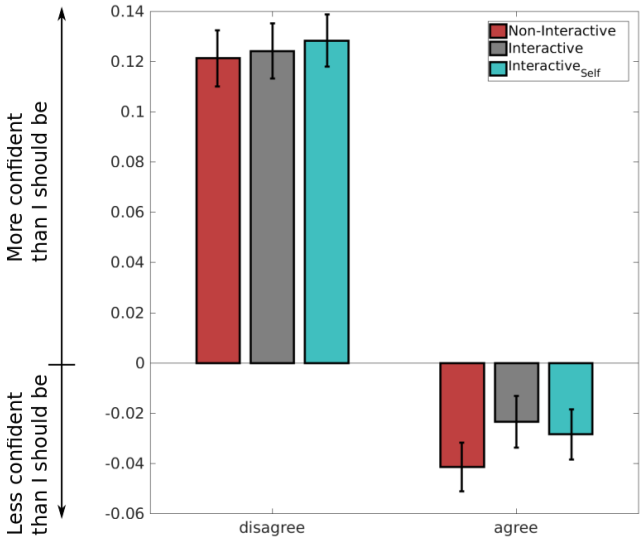


Figure 6.8: The figure shows how confidence changes observed in the data relate with the confidence changes expected by applying a normative Bayesian update rule. Participants showed a conservative bias thus decreased their confidence too little in disagreement and increased it too little in agreement trials.

Social information perception analysis. Experiment 4 showed that by applying inverse Bayes, we can compare the participants' perceived social support for their initial judgments with the objective supporting evidence provided by their partners, thus highlighting biases and distortions. Social information perception analysis was performed here to replicate the finding that an originally homogeneous distribution of objective supporting evidence is transformed into a tri-modal distribution of perceived evidence, with peaks on 0 (certain disagreeing evidence), 0.50 (neutral evidence) and 1 (certain supporting evidence). Figure 6.9 shows objective ("stated") and perceived distributions of social evidence for one's own opinion, divided according to trial dom-

inance. Pre-social information confidence greater than 40 were removed to avoid inconsistencies in the Bayes formula, due to high confidence trials.

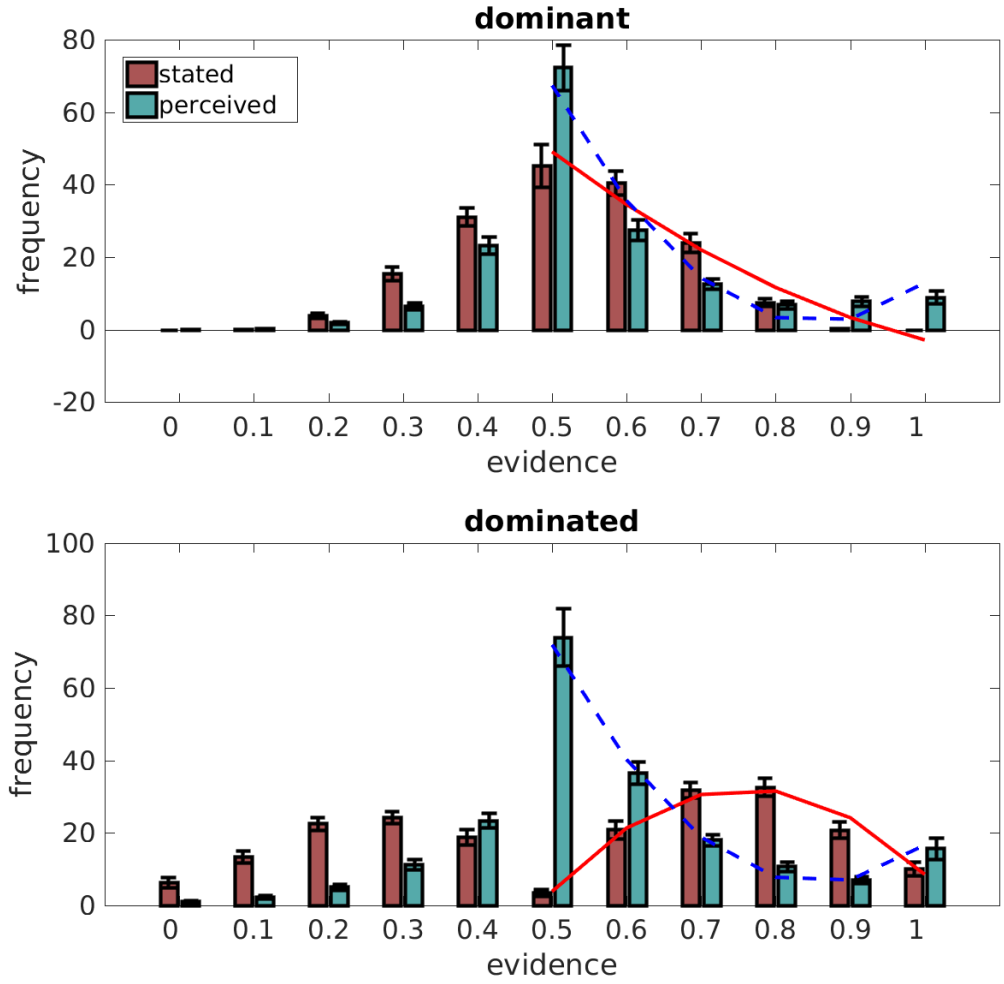


Figure 6.9: How much a partner's opinion is perceived to support one's own independent judgment, compared to objectively stated partner's supporting evidence. Differences between the two indicate cognitive distortions of social information.

Experiment 5 confirmed the presence of a cognitive distortion, whereby the objective evidence distribution is transformed into a distribution with peaks around 0.50 and 1. The peak around 0 (certain disagreeing evidence) was not observed in Experiment 5, probably due to the use of less extreme confidence judgments (Figure 6.3). Irrespective of dominance, the perceived evidence distributions seem to

follow a bimodal trend centred on 0.5 and 1, suggesting that participants tended to either ignore or completely follow received social information. To formally test for this intuition, a linear model with a linear and quadratic terms was fitted to agreement trials (x-axis greater than 0.50), as disagreement trials were less numerous: $frequency = \beta_0 + \beta_1 evidence + \beta_2 evidence^2$ (Table 6.1). When looking at the objective evidence received (Figure 6.9, red bars) the quadratic term was not significantly different from zero for dominant members (Dominant: $\beta_2 = 1.04, SE = 0.89, p = .32$) and negative for the dominated member (Dominated:$\beta_2 = -4.14, SE = 0.38, p = .001$). When looking at the perceived evidence, on the contrary, quadratic terms were significantly above zero for both dominant ($\beta_2 = 5.22, SE = 1.13, p = .01$) and dominated individuals ($\beta_2 = 5.16, SE = 0.50, p = .001$). The results suggest that, in both participants, the frequency of *objective* social information over evidence bins did not peak on 0.50 and 1, while the frequency of *perceived* evidence did, confirming the hypothesis that social information tended to be perceived as either neutral or fully supporting one's own views.

	Dominant Stated			
	Estimate	SE	tStat	p
β_0	65.78	9.74	6.75	.006
β_1	-17.72	6.37	-2.78	.06
β_2	1.04	0.89	1.17	.32
	Dominant Perceived			
	Estimate	SE	tStat	p
β_0	109.71	12.38	8.85	.003
β_1	-47.50	8.10	-5.86	.009
β_2	5.22	1.13	4.61	.01
	Dominated Stated			
	Estimate	SE	tStat	p
β_0	-21.99	4.19	-5.24	.01
β_1	29.95	2.74	10.90	.001
β_2	-4.14	0.38	-10.79	.001
	Dominated Perceived			
	Estimate	SE	tStat	p
β_0	113.94	5.49	20.74	.0002
β_1	-47.20	3.59	-13.13	.0009
β_2	5.16	0.50	10.27	.001

Table 6.1: Linear models with linear and quadratic terms for evidence bin. The model indicates cognitive distortions translating objective social information into perceived social information.

Egocentric and confirmation biases. Experiment 4 had shown that participants did not follow a pure Bayesian observer with equal weights for self and other's opinion. Often they seemed to ignore social information when it was most useful, namely in situations of low pre-social information confidence. Finally, there was evidence for different confidence update behaviour in agreement and disagreement trials.

Discounting parameters were fitted to the behavioural data of Experiment 5 to account for other's discounting (i.e., ego-centric bias (Yaniv, 2004b)). Discounting parameters were fitted separately for agreement (α) and disagreement (β) trials to account for asymmetric discounting in the two (i.e., confirmation bias (Nickerson,

1998)). A three-way ANOVA with factors dominance, condition and discounting parameter (discounting in agreement α vs. discounting in disagreement β) showed a significant effect for dominance ($F(1,47) = 15.10, p < .001, \eta_G^2 = .02$), condition ($F(2,94) = 4.32, p = .01, \eta_G^2 = 0.001$) and parameter type ($F(1,47) = 41.02, p < .001, \eta_G^2 = .17$), suggesting greater weights on partner's opinions in dominant compared to dominated trials (means: 0.67 vs. 0.49) and in agreement compared to disagreement trials (means: 0.81 vs. 0.35). The latter replicates the confirmation bias found in Experiment 4. Pairwise comparisons were performed to understand differences across conditions. Interactive condition showed less discounting overall compared to the Non-Interactive condition (0.60 vs. 0.56, $t(47) = 2.98, p = .004, d = 0.12$) and the Interactive$_{self}$ condition (0.60 vs 0.58, $t(47) = 2.08, p = .04, d = 0.07$). No significant difference was found between these two conditions ($p > .3$). Significant interactions were found between dominance and parameter-type ($F(1,47) = 9.99, p = .002, \eta_G^2 = .01$) and between condition and parameter-type ($F(2,94) = 12.96, p < .001, \eta_G^2 = .007$), but not between dominance and condition ($F < 1$). A marginal three-way interaction was also found ($F(2,94) = 2.82, p = .06, \eta_G^2 = .001$).

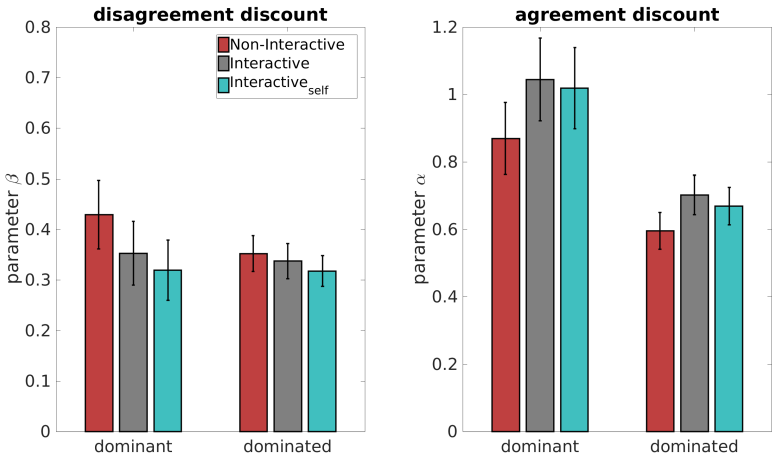


Figure 6.10: Fitted α and β representing the discounting factor for agreement and disagreement trials respectively.

Experiment Discussion

Experiment 5 was run with two main aims in mind: (1) reproduce results from Experiment 4 when using different incentive schemes; (2) test whether those same effects were produced by failure in remembering one's own initial judgment.

The results described above replicate the key findings of Experiment 4 and reproduce the overall pattern of differences between interactive and non-interactive conditions. In particular, interaction seems to significantly increment confidence increases observed from pre- to post-social information phase after agreement, regardless of the presence of a confidence reminder. Similarly to what observed in Experiment 4, the decrease in confidence observed after a disagreeing opinion is reduced in Experiment 5 in both interactive conditions, although not significantly in the Interactive condition. Differences in confidence change among conditions were once again driven by weak agreement trials and unbalanced disagreement trials. The findings also replicated the positive correlation emerging during real-time interaction between dyad members' confidence changes. Contrarily to Experiment 4 however, no negative correlation was found in disagreement trials, suggesting that during these trials dyad members' updates remained independent from each other irrespective of condition.

Accuracy improvements from pre- to post-social information were all positive and significantly different from zero. Conditions did not differ from each other suggesting that, notwithstanding the reduced independence of participants' judgments, performance improvements were robust. Interaction favoured greater improvements compared to the non-interactive baseline both in terms of graded accuracy (i.e., confidence change toward the correct choice) and confidence calibration.

The normative framework described for Experiment 4 was also applied here to show that people adopt qualitatively different strategies in agreement and disagreement, with greater weights put on partner's opinions in agreement trials. Confirming results found in Experiment 4, Experiment 5 provided further evidence that social

information perception differs from the objective social information received. In particular, participants tend to categorise received social information into strong evidence in favour of their initial opinion or null evidence.

Overall the experiment showed that most of the effects observed after the interaction manipulation are robust to changes in the use of the confidence scale, with few differences found in disagreement trials. Importantly, the introduction of a confidence reminder little affected the Interactive condition, suggesting that differences between interactive and non-interactive conditions were not simply due to memory failures of one's own initial confidence. The experiment thus offered a proof that interactive and non-interactive paradigms differ not only in terms of low-level characteristics but instead differences are intrinsic to the dynamics of how information is shared and manipulated across individuals.

Experiment 6 was carried out to test whether differences between interactive and non-interactive conditions were instead due to memory failures of one's *partner*'s initial confidence.

Experiment 6

Introduction

Experiment 4 showed differences in behaviour emerging from the manipulation of how social partners can exchange their independent pieces of information. The independence of confidence updates was affected by the presence of real-time interaction, suggesting that participants updated their confidence not only using the initial confidence of their partner but also their partner's updates. This strategy can generate phenomena of confidence escalation due to non-linearity of the interaction (Mahmoodi et al., 2013), whereby increases in confidence lead to further increases in confidence in a positive feedback cycle. Experiment 5 ruled out a simple explanation in terms of

participants forgetting their own initial confidence. The memory failure hypothesis was not sufficient to explain the effects found during interaction. However another explanation for confidence escalation is that participants forgot *their partner's* initial confidence and were thus incentivised to use, when available, their partner's current position as a proxy for it.

To test whether this explanation could explain the effects found in the Interactive condition a new condition was set out and compared with the Interaction and the Interaction$_{self}$ conditions. In this condition, called Interaction$_{other}$, the Interactive condition is enhanced by the presence of a static reminder about one's partner's initial confidence that remains on screen for the whole duration of the social exchange. If the memory explanation is correct we expect the effects of interaction to diminish when a reminder is presented. Failure in finding such results can be taken as evidence that the effects of interaction are not due to failures in memory.

Methods

Participants. 24 dyads (17 female dyads, 1 mixed gender) were recruited using University volunteers recruitment platform and local advertisement websites. Dyads were recruited by asking an interested volunteer to bring along a friend of the same gender. Participants (age=20.66±2.76) signed a consent form prior the beginning of the experiment. The study received ethical approval from the University ethical committee.

Paradigm. Participants performed 18 blocks of 24 trials each. Perceptual task, trial sequence and response modality were kept equal to previous experiments. The social window was kept to 4 seconds as in Experiment 5. Given that Experiment 5 was successful in making participants less extreme in their confidence ratings, in Experiment 6 it was decided to implement the same incentive scheme. The experiment

started with 4 practice blocks of 5 trials each, corresponding to practice with the perceptual task and with each condition separately. Performance was titrated to 70.7% accuracy using a 2-down 1-up procedure.

Manipulation. Three experimental conditions were implemented and alternated across blocks in six identical modules of three blocks each. The order of the three conditions within a module was randomly shuffled across dyads but remained identical within the same dyad. The first two conditions were the Interaction and Interaction$_{self}$ conditions, already described in Experiments 5. A third new condition, named Interaction$_{other}$, was implemented by adding to the Interactive condition a static cursor reminding the participant of their partner's initial confidence level. A colour code was used so to avoid confusion on what each cursor meant. Participant-related cursors were represented in white (active cursor) and grey (static reminder). Partner-related cursors were represented in bright colour (active cursor) and dark colour (static reminder).

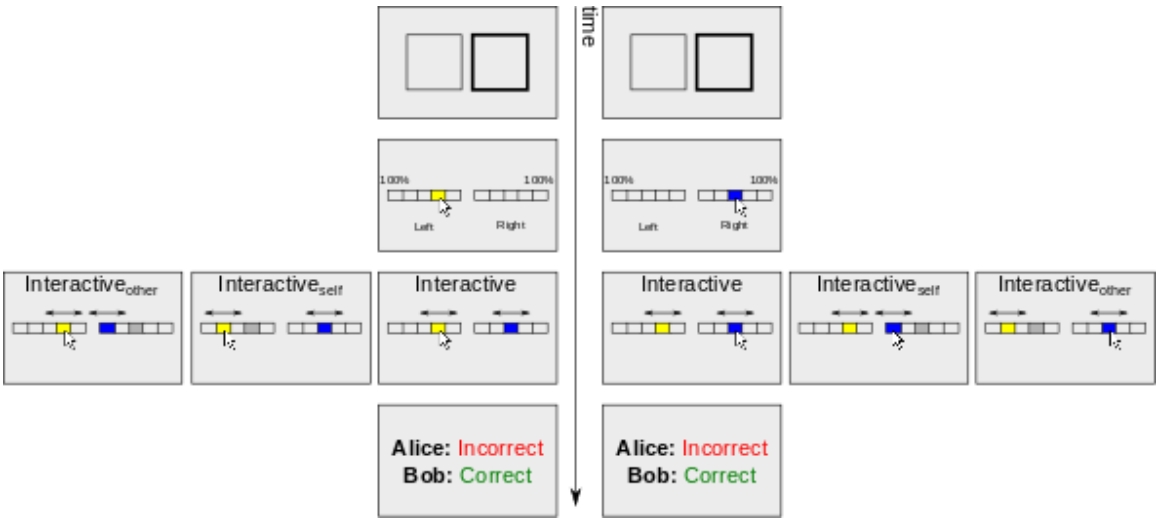


Figure 6.11: Experimental paradigm implemented in Experiment 6. During the Interactive condition participants are shown the current opinion of their partner in real-time. During the Interactive$_{self}$ participants are shown the current real-time opinion of their partner and are at the same time reminded of their own original opinion as a shaded cursor on the scale. During the Interactive$_{other}$ participants are shown the current real-time opinion of their partner and are at the same time reminded of their partner's original opinion as a shaded cursor on the scale. This manipulation makes sure that after a change in the configuration of the elements present on screen participants are reminded of where along the scale they started from or where their partner started from. In all conditions participants have four seconds when they are asked to update their own original confidence level using post-decisional information.

Results.

Continuous update. Similarly to Experiment 4 and 5, Figure 6.12 shows that a sharp confidence update occurred in all conditions around the first second of the social window and settled into an equilibrium by the end of it. Differences between anchor conditions and interactive baseline were tested for significance using both a point-wise two-tail t-test and a cluster-based permutation t-test to control for multiple comparisons problem (Figure 6.12, right panel). Results indicate no significant differences between conditions, indicating that the time used by dyads to reach their final decision was not affected by the presence of a confidence reminder.

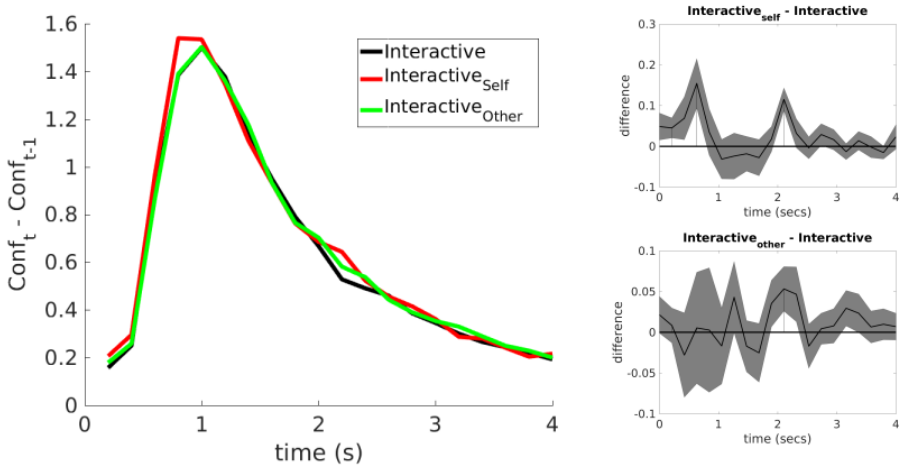


Figure 6.12: Difference in recorded confidence between two subsequent data points during the social window. The measure can be used to plot how quickly participants' updates converged to a final confidence level. Right panels: within-participants pointwise difference between anchor conditions and interactive baseline.

Confidence distributions. Figure 6.13 shows the pre-social information confidence distributions recorded from each participant. It can be seen that confidence distributions resembled more closely the distributions observed in Experiment 5 than in Experiment 4, as suggested by the absence of extreme high confidence values. Some participants show clusters in correspondence of scale landmarks. Experiment 4 showed marginal differences existing between conditions in pre-social information mean confidence. The result was not replicated in Experiment 5. Experiment 6 showed no significant difference among conditions in the mean pre-social information confidence ($F < 1$), suggesting that, on average, all interactive conditions started from similar levels along the scale.

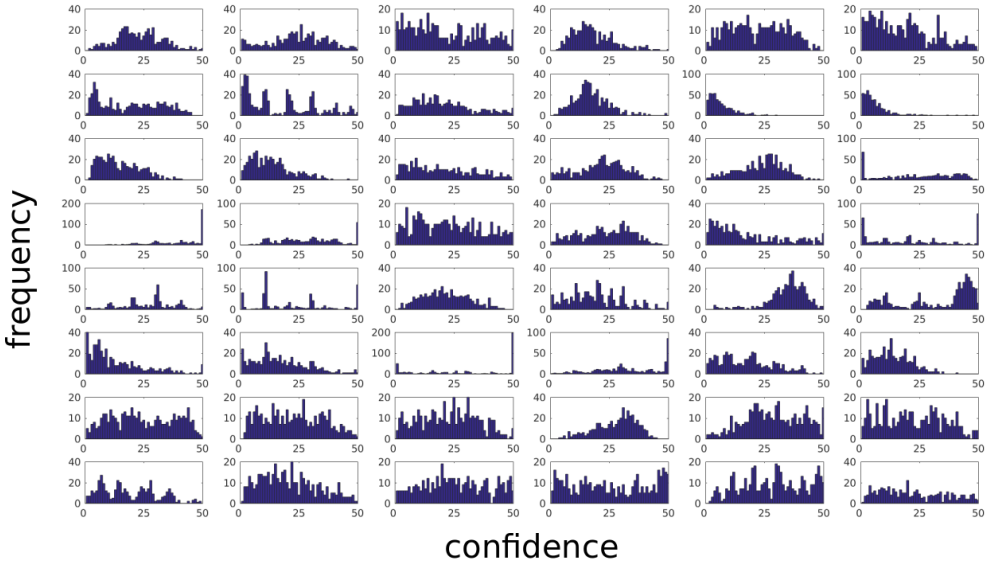


Figure 6.13: Pre-social information confidence distributions in Experiment 6. It can be seen that confidence distributions are more similar to distribution observed in Experiment 5 than Experiment 4, as indicated by the absence of extreme values (i.e., around 50). Each row represents confidence distributions of members of three dyads. First row: Dyad 1=first two columns; Dyad 2 = central two columns; Dyad 3 = last two columns; etc.

Confidence changes. Figure 6.14 shows the average signed confidence change observed after agreement or disagreement and divided by condition. A two-way repeated measures ANOVA on confidence change showed a significant effect of consensus ($F(1.47) = 246.77, p < .001, \eta^2_G = .75$) and only a marginal effect of condition ($F(2, 94) = 2.43, p = .09, \eta^2_G = .001$), suggesting that the introduction of pre-social information confidence reminders did not strongly affect the Interactive condition. Pairwise comparisons showed only a significant enhanced disagreement effect for the Interactive$_{other}$ condition compared to the Interactive baseline ($t(47) = 3.28, p = .001, d = 0.17$).

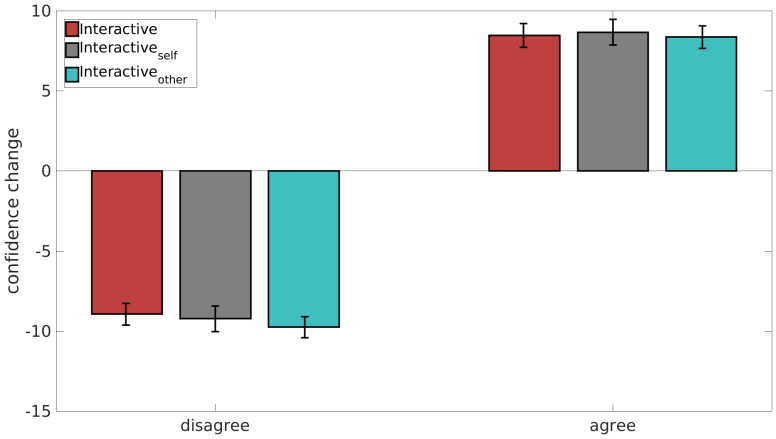


Figure 6.14: Average confidence change observed from pre- to post-social information and divided by consensus and condition. Increases in initial confidence are indicated by positive values while decreases of confidence are shown as negative values.

Asymmetry in confidence increases. The confidence change distributions of Experiment 6 are shown in Figure 6.15 as root density plots. As in previous experiments the most common confidence change was zero, suggesting that very often participants decided not to act upon social information. To test for asymmetries in irrational confidence changes, a two-way repeated measures ANOVA on the probability of an irrational confidence change was run. Results showed only a marginal effect of consensus ($F(1,47) = 3.08, p = .08, \eta^2_G = .02$) and no effect of condition ($F(2,94) = 1.03, p = .35, \eta^2_G = .001$) nor significant interaction ($F < 1$). Experiment 6 replicates the finding found in the previous two experiments that irrational changes were more frequent in disagreement than in agreement trials (M±STD: irrational increases $= 0.0124 \pm 0.018$; irrational decreases $= 0.0072 \pm 0.008$). Similarly to previous experiments no difference was found among conditions, suggesting that the presence of a confidence anchor did not affect the presence of irrational confidence changes in the baseline Interaction condition.

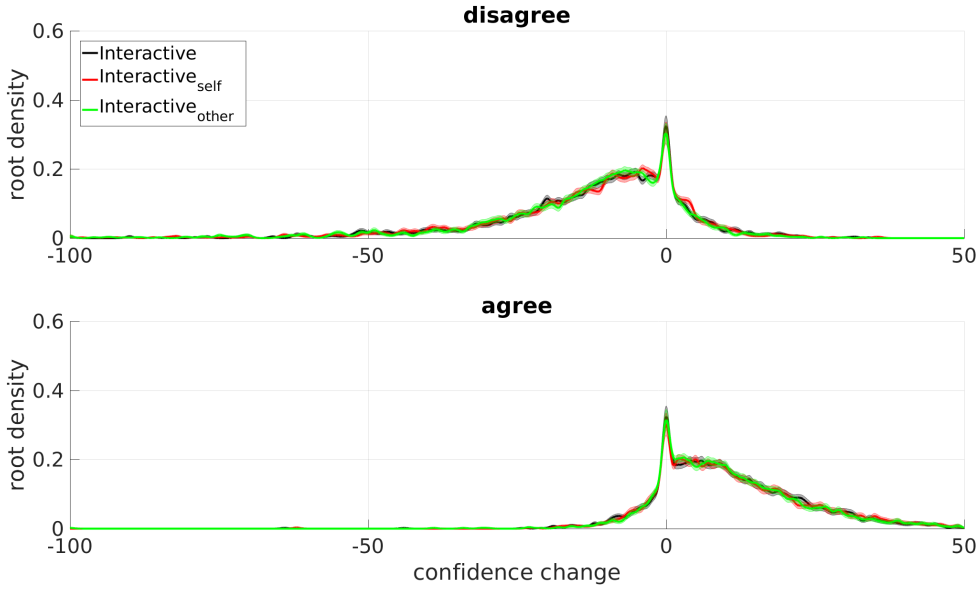


Figure 6.15: Root density distributions of confidence changes divided by condition and consensus. Density plots are obtained from Gaussian kernel function with bandwidth = 0.50.

Influence in opinion space. I then looked at confidence changes along the opinion surface. Figure 6.16 shows that the pattern of results is very similar to what obtained from the previous two experiments. Contrast plots (panels D,E,I,L) between anchor conditions and baseline interaction showed no difference in weak agreement areas (points x), indicating that the presence of confidence anchors did not alter median confidence change in these trials. Unbalanced disagreement (points y) showed no difference among conditions for dominant trials (plots D,E) but positive differences for dominated trials (plots I,L). The latter finding suggests that in these trials, dominated members seemed to be more swayed by dominant opinions in both anchor conditions compared to baseline Interaction, but dominant ones were not. Points z on dominant trials (panels D-E) seem to suggest that, compared to the Interactive condition, participants tended to decrease their confidence more in Interaction$_{self}$ (panel D) but not in Interaction$_{other}$ (panel E). Points z on dominated trials (panels I-L) suggest that, compared to the Interactive condition, participants tended to decrease their

confidence more in both Interaction$_{self}$ (panel I) and Interaction$_{other}$ (panel L), suggesting that being reminded of holding the dominated view made participants more inclined to decrease their confidence. A note of caution should however be raised in interpreting these findings due to the sparseness of data, particularly for extreme disagreement cases (points z).

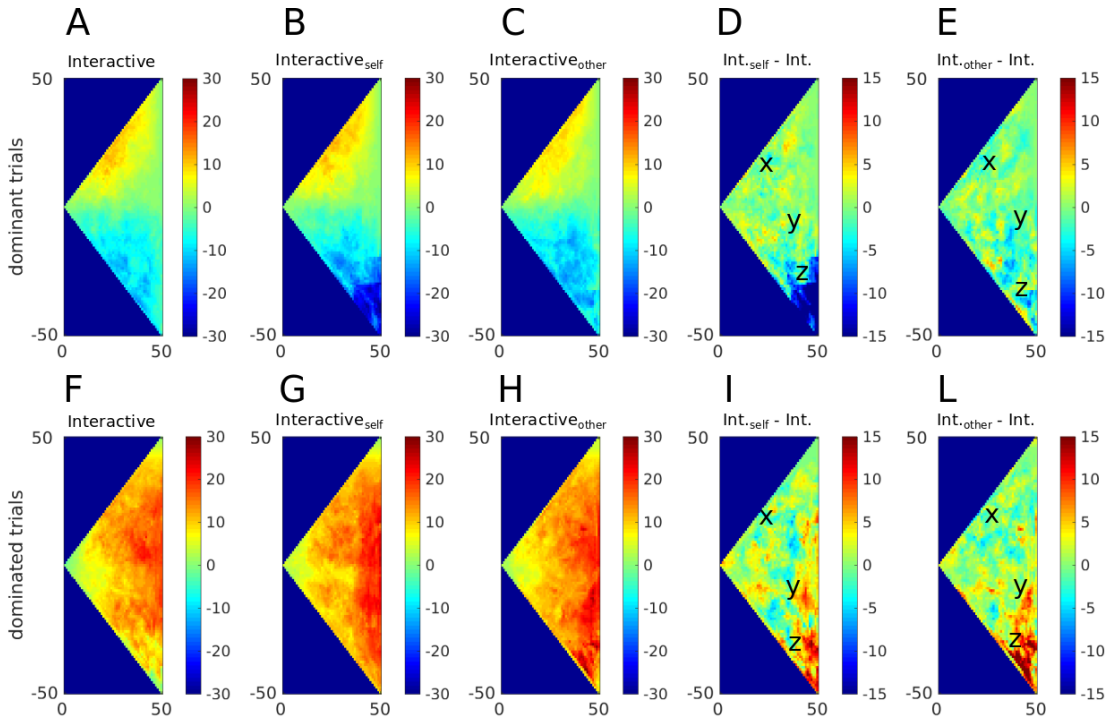


Figure 6.16: Median confidence change in opinion space for dominant and dominated trials and divided by condition (plots A-C and F-H). Plots D-E and I-L are contrast plots showing the difference between each reminder condition and baseline interaction.

Coupling of confidence changes in interaction. Pearson's correlation coefficients between confidence change magnitudes were compared across conditions and divided by agreement to test whether changes in one participant were correlated with changes in the other (Figure 6.17). Results of a repeated measure ANOVA on Pearson's coefficients showed that a significant effect of consensus was found ($F(1, 23) = 39.74, p < .001, \eta^2_G = .24$) but not of condition ($F < 1$) and no significant interaction between the two ($F < 1$). Contrary to Experiment 5 but similarly to

Experiment 4, in all conditions confidence change magnitudes in disagreement trials were marginally or significantly below zero (Interaction: $t(23) = -1.96, p = .06, d = -0.40$; $\text{Interaction}_{self}$: $t(23) = -1.87, p = .07, d = -0.38$; $\text{Interaction}_{other}$: $t(23) = -3.18, p = .004, d = -0.64$), indicating that interaction produced an inverse coupling also in disagreement, with little effect of reminders. Similarly, in agreement trials dyad members' confidence changes were positively correlated as indicated by the significantly positive correlation coefficients (Interaction: $t(23) = 3.40, p = .002, d = 0.69$; $\text{Interaction}_{self}$: $t(23) = 3.00, p = .006, d = 0.61$; $\text{Interaction}_{other}$: $t(23) = 2.97, p = .006, d = 0.60$). In conclusion, irrespective of reminder presence, interaction coupled together partners' confidence changes: greater confidence changes in one dyad member produced greater partner's confidence changes in agreement but lower partner's confidence changes in disagreement.

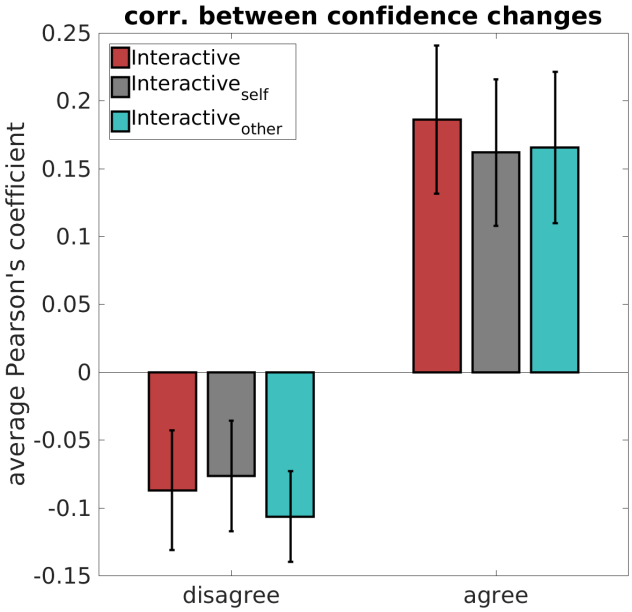


Figure 6.17: Coupling (as measured by Pearson's correlation coefficient r) between absolute confidence changes of members of the same dyad. Error bars represent s.e.m.

The mixed-effects linear regression described in Experiments 4 and 5, was run here to check whether the mediating role of partner's confidence change on subjective

confidence changes differed across conditions. Consensus positively interacted with partner's absolute confidence change ($\beta = 0.4318, SE = 0.0244, p < .001$), suggesting that the larger their partner's changes were the more participants tended to shift their confidence in agreement and less they tended to shift in disagreement. Importantly the effect was not modulated by either of the anchor conditions (Interactive$_{self}$: $\beta = 0.0230, SE = 0338, p = .49$; Interactive$_{other}$: $\beta = -0.0113, SE = 0.0339, p < .7$) suggesting that the introduction of confidence reminder did not affect the baseline Interactive condition in the extent to which partner's updates affected subjective updates.

Performance analysis. Both Experiments 4 and 5 showed that performance improved after social exchange and that interaction did not negatively affect the size of the improvement. Results were replicated in Experiment 6. A two-way ANOVA on accuracy with factors condition and decision time (pre-social information vs. post-social information) showed a significant effect of decision time ($F(1, 47) = 103.96, p < .001, \eta^2_G = .19$), indicating accuracy improvement due to social information exchange (M: 0.71 vs. 0.74). Importantly no effect of condition nor interaction were found (both $F < 1$), confirming that different conditions did not affect average accuracy or average accuracy improvement.

A two-way ANOVA on graded accuracy (defined as the confidence in the correct answer) with factors condition and decision time showed a significant effect of decision time ($F(1, 47) = 165.89, p < .001, \eta^2_G = .23$) indicating an improvement due to social information, but no main effect of condition or interaction ($F(1, 47) < 1.35, p > .2, \eta^2_G = 3e-4$), suggesting that confidence reminders did not affect graded accuracy.

To test for confidence calibration improvements, a two-way ANOVA with factors condition and decision time on type II A_{ROC} showed a significant effect of decision time ($F(1, 47) = 112.49, p < .001, \eta^2_G = .23$), indicating that calibration improved thanks to social information exchange (M=0.57 vs. 0.62), but not of condition

($F < 1$). A marginally significant interaction between the two terms was also found ($F(2, 94) = 2.95, p = .05, \eta^2_G = .007$). Pairwise comparisons showed that both the Interactive$_{self}$ ($t(47) = 2.23, p = .03$) and the Interactive$_{other}$ ($t(47) = 1.96, p = .05$) conditions produced significant greater calibration improvement over the Interactive baseline. The results suggest that, although not having any effect on accuracy, the presence of a confidence reminder (either own or partner's) helped participants to have a more accurate metacognitive evaluation, likely because of an increased access to independent estimates.

Humans show overconfidence compared to the Bayes. Compared to a simple Bayesian observer with equal weights of members' opinions, participants showed a similar pattern of results as the one observed in previous studies. Participants tended to be more confident than Bayes in disagreement and slightly less confident in agreement trials (6.18). A two-way ANOVA with factors consensus and condition was run on the residuals of human performance from normative predictions, with positive values indicating over-confidence compared to Bayes and negative values indicating under-confidence. Results show a significant effect of consensus ($F(1, 47) = 85.64, p < .001, \eta^2_G = .53$) but no significant effect of condition ($F < 1$). No significant interaction between the two terms was found ($F < 1$). Results thus confirmed that the addition of a confidence reminder did not affect the overall pattern observed in the Interactive condition.

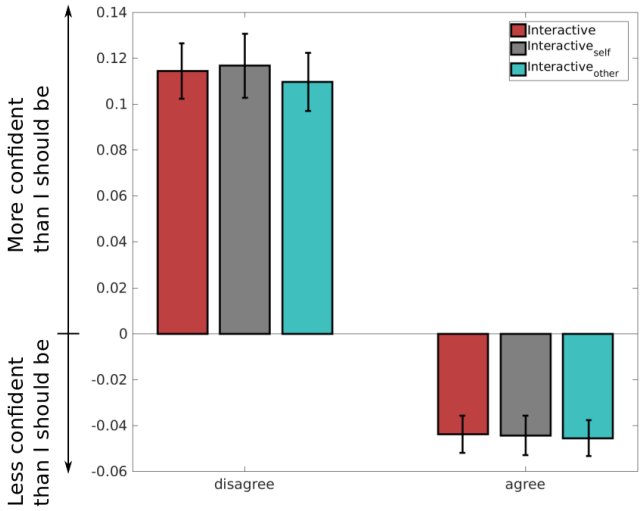


Figure 6.18: Residuals between human participants and a simple Bayesian model aggregating the two opinions using equal weights. Residuals represent over- or under-confidence compared to model's predictions.

Social information perception analysis. Objective evidence was compared to perceived evidence as described in Experiment 4. Figure 6.19 shows the two distributions divided by the dominance of the participant's view. As in previous experiments the graph shows a dissociation between the two distribution with the perceived distribution having higher peaks in correspondence of points 0.50 and 1, corresponding to neutral social information and strongly supporting social information respectively. Trials above confidence rating 40 were removed to avoid inconsistencies due to the use of prior probabilities equal to one.

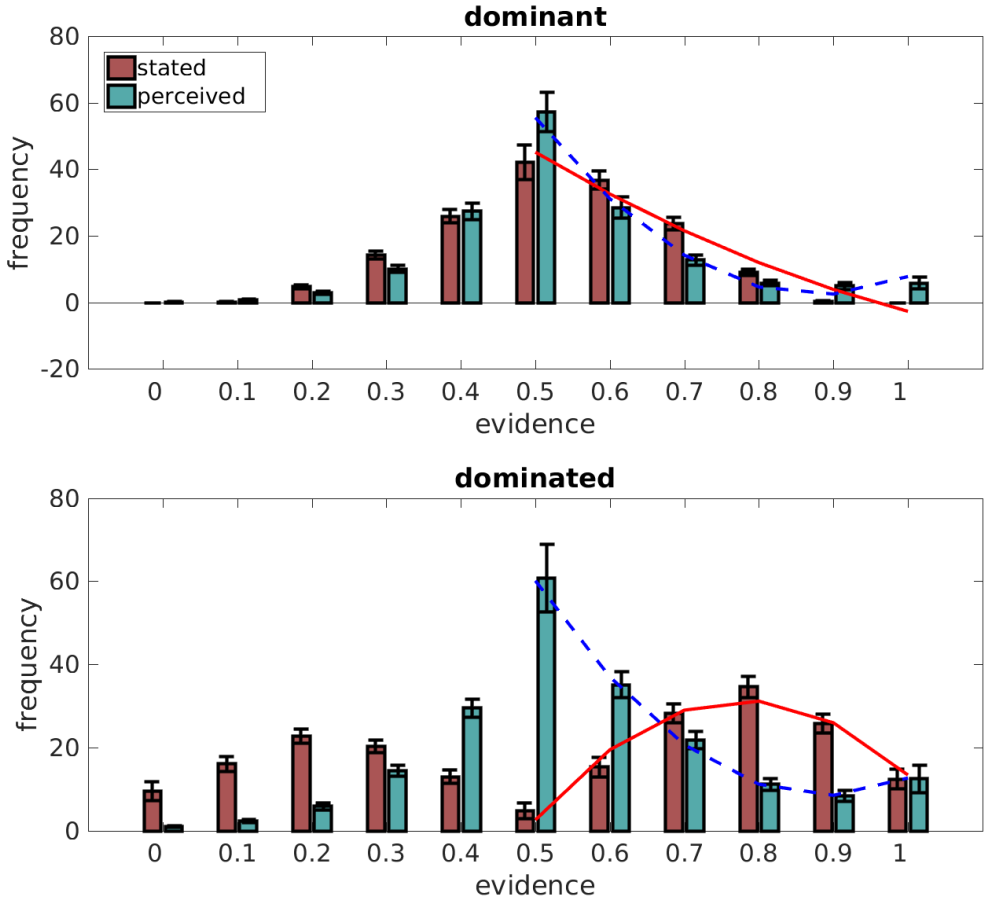


Figure 6.19: The figure compares the distribution of the supporting evidence provided by the partner's social information (objective evidence) with the evidence estimated to be perceived by the participant. The graphs show a dissociation between the two.

To statistically test for the presence of peaks on 0.50 and 1, a linear model that contained a linear and a quadratic term for evidence (x-axis) was fitted to the average frequency in each bin (y-axis): $frequency = \beta_0 + \beta_1 evidence + \beta_2 evidence^2$. Results are shown in Table 6.2. It can be seen that the quadratic term is positive and significant only for perceived evidence data (Dominant: $\beta_2 = 3.70, SE = 0.45, p = .003$; Dominated: $\beta_2 = 3.41, SE = 0.21, p = .0005$) suggesting that perceiving medium supporting evidence (e.g. bins 0.70-0.90) was less frequent that perceiving neutral evidence (bin 0.50) and confident supporting evidence (bins 0.90-1). Nevertheless, objective received evidence from partner's social information did not follow the same

positive quadratic distribution, suggesting that participants transformed social information before incorporating it into a confidence update.

	Dominant Stated			
	Estimate	SE	tStat	p
β_0	59.20	7.90	7.49	.004
β_1	-14.80	5.17	-2.86	.064
β_2	0.74	0.72	1.03	.37
	Dominant Perceived			
	Estimate	SE	tStat	p
β_0	87.33	5.02	17.38	.0004
β_1	-35.49	3.28	-10.79	.001
β_2	3.70	0.45	8.06	.003
	Dominated Stated			
	Estimate	SE	tStat	p
β_0	-21.63	6.18	-3.50	.03
β_1	27.91	4.04	6.90	.006
β_2	-3.67	0.56	-6.50	.007
	Dominated Perceived			
	Estimate	SE	tStat	p
β_0	90.07	2.35	38.24	.00003
β_1	-33.36	1.54	-21.65	.0002
β_2	3.41	0.21	15.83	.0005

Table 6.2: Linear models with quadratic terms on evidence bin. Bin frequency represent the dependent variable considered.

Egocentric and confirmation biases. A bilinear model was fitted on perceived social information evidence using objective evidence provided by the social information. Fitting lines were anchored on point (0.5,0.5) corresponding to neutral social information provided (nominal confidence rating equal to zero) and only slopes were allowed to vary. Slope estimates indicate the weight put on partner's opinions. Weights lower than 1 indicate egocentric bias (Yaniv & Kleinberger, 2000). Different weights were allowed for agreement (α) and disagreement (β) to account

for confirmation bias (Nickerson, 1998). A three-way ANOVA with factors dominance, condition and discounting factor type (α vs. β) showed a significant effect of dominance ($F(1, 47) = 28.34, p < .001, \eta^2_G = .04$) suggesting that greater discounting was operated by dominated individuals. A significant main effect of discounting factor ($F(1, 47) = 7.01, p = .01, \eta^2_G = .03$) indicated that discounting was significantly greater in disagreement than agreement ($t(47) = 2.64, p = .01, d = 0.43$), replicating the presence of a confirmation bias. Importantly, no main effect of condition was found ($F < 1$), suggesting that overall condition did not affect discounting magnitude. Condition was however found to significantly interact with discounting factor ($F(2, 94) = 3.68, p = .02, \eta^2_G = .002$). Pairwise comparisons showed that confirmation bias was less pronounced in the Interactive$_{other}$ condition compared to the other two ($t(47) < -2.12, p < .05, d < -0.17$). A marginal interaction between discounting factor and dominance ($F(1, 47) = 3.46, p = .06, \eta^2_G = .006$) and a marginal three-way interaction ($F(2, 94) = 2.47, p = .08, \eta^2_G = .0009$) were also found.

Experiment Discussion

Experiment 6 replicated all the key findings observed in the previous two experiments. Results show that the introduction of confidence reminders had moderate effects compared to the baseline Interactive condition. The presence of the other person's confidence reminder made participants decrease their confidence more in disagreement, thus ending on lower absolute confidence levels. Reminders did not seem to affect the independence of the confidence updates over and above what already observed in the Interactive condition. They did not produce differences in choice accuracy or accuracy improvements either. Only marginal differences in calibration improvements were found among conditions.

Comparisons with a Bayesian opinions integration strategy confirmed the observations made in the previous two experiments, suggesting that participants discounted social information received from partner. Participants tended to differently treat

agreeing and disagreeing evidence and asymmetrically discount the two. Furthermore, greater social information discounting was operated by participants holding the dominated opinion, probably due to a general tendency to discount/ignore social information that deviated from Bayes particularly in dominated trials.

Although some effects were observed by the introduction of the other member's confidence reminder, the experiment failed to provide evidence that the results observed in the Interactive condition were entirely due to a memory failure in remembering initial opinions (own or other's). The pattern of results observed in the Interactive condition was nearly unaltered, suggesting that even in the presence of a constant reminder anchoring participants to their initially expressed views phenomena of confidence escalation and updates coupling were observed. Thus it seems that confidence escalation and the correlations emerging in interaction between updates of members of a same dyad cannot be explained away by simple mechanisms specific to our paradigm. The results are so far in agreement with an explanation in terms of interaction modifying the dynamics of the information exchange between two decision makers. Interaction creates a situation where both participants can not only use the independent opinion of their partner to inform their post-decisional judgments but also how their partners react to the participants' opinion. When interaction was allowed seeing larger updates in their partners made participants' confidence change size increase in agreement and decrease in disagreement. The results add on a large body of evidence suggesting that confidence judgments are not only the product of a careful evaluation of decision relevant variables but often include several contingent cues that are not decision-relevant but flow into creating a unitary internal sense of confidence (Gigerenzer et al., 1991; Koriat, 2012a; Slegers, Brake, & Doherty, 2000). Interestingly interaction decreased the independence of the two members' judgments in all three experiments using the current paradigm. Contrary to a common interpretation of Wisdom-of-Crowds phenomena in terms of noise cancellation through

averaging of independent measures (Galton, 1907; Lorenz et al., 2011) it was repeatedly shown that increased dependence did not significantly affect accuracy nor accuracy improvement. This suggests that when people are allowed to share their confidence judgments instead of their choice preferences only, individual and dyadic performance can be extremely robust to failures (Bahrami et al., 2012b; Navajas et al., 2017).

Conclusions

In this Chapter I investigated the possibility that differences observed in Experiment 4 between Interaction and Non-Interaction were due to memory failures in the maintenance of initial confidence. If this was the case we would expect that the presence of a confidence reminder would improve any memory deficit thus reducing differences between interactive and non-interactive conditions. Experiment 5 showed that this was not the case when a reminder of one's own initial confidence was available for the whole duration of the social window. Experiment 6 tested this idea further by adding a condition where a reminder of the other person's initial opinion was made available. Again, the experiment failed to show differences between Interactive and reminder conditions, suggesting that the presence of an confidence anchor was not sufficient for interaction effects to disappear.

Overall, the two experiments are congruent with an alternative explanation suggesting that Experiment 4 key findings emerged because of fundamental different dynamics characterising interactive versus static social exchanges. The next chapter will use the interactive paradigm but explore further serendipitous results found in Experiments 4-6 indicating confidence alignment between participants confidence distributions.

7

Alignment of Confidence in Interaction

Chapter Abstract

A final study using the dyadic interaction paradigm is reported. A serendipitous finding of Experiment 4 suggested that participants within a dyad tended to align their confidence distribution over and above what would be expected by chance. This result was replicated in the Experiments 5 and 6 and reported in Appendix C. Experiment 7 tested an account of this confidence alignment phenomenon, that it emerges as a strategy to cope with inter-individual differences in the use of confidence scales. To test this hypothesis in the current experiment pre-advice confidence distributions were recorded before and after interaction took place. For the first part of the experiment participants were asked to perform the dot-count perceptual judgment on their own. During the second part, participants performed the task interactively as described in Experiments 4-6. Finally, during the last part of the experiment, participants performed again the task on their own. Results show that confidence alignment is low at the beginning of the experiment but quickly increases following interaction. The effects of alignment carry over during the final part. It is concluded that alignment is caused by social interaction. The effect is discussed within a broader range of literature on metacognition and its social value.

Confidence alignment in Experiment 4

The experiments reported in Chapters 4 and 5 investigated how interaction led to correlated changes in confidence. As well as showing predicted effects, these experiments revealed an unexpected effect of social information in terms of alignment of confidence distributions between members of a dyad. Simple visual inspection of the confidence distributions in Experiment 4 indicated that these were more similar for participants within dyads than across dyads, even though these distributions plot the initial "private" decision, before interaction on the trial. This is an interesting finding considering that, in joint tasks, people have been shown to be negatively impacted by differences in abilities (Bahrami et al., 2012b, 2010) and confidence use (Fusaroli et al., 2012; Pescetelli et al., 2016), or by equal weighting policies (Mahmoodi et al., 2015). Confidence matching can work as a heuristic to communicate uncertainty (Bang, Aitchison, et al., 2017), raising the question of why alignment was observed in our data. To test the confidence similarity more formally, a bootstrapping procedure was applied that compared the correlation coefficient between mean confidence within a dyad against the distribution of correlation coefficients observed by randomly pairing participants. A corresponding analysis was performed on the spread of confidence distributions within vs. across dyads. Mean and variance were used as they have been shown to be informative statistics of someone's confidence and robust across tasks (Ais et al., 2016). For these analyses, participants were shuffled into nominal dyads and the correlation between means and the correlation between standard deviations were computed. This procedure was repeated 1000 times, each time with a new random pairing. Figure 7.1 shows the distribution of Pearson's r correlation coefficients expected by random pairing of participants and the one observed on empirical dyads. Empirical Pearson's r (solid line) was more than 3.9 σ away from the distribution centre (greater than 99.9% of values) for the mean and 2.21 σ away from the centre (greater than 98.8% of values) for the standard deviation. Members of a

same dyad aligned their confidence distributions (i.e., how they used the confidence scale), both in terms of mean and standard deviation.

To test for changes in this confidence alignment over time, we computed the average Kullback-Leibler (KL) distance between pre-social information confidence distribution (divided in 5 bins) of the two members in the first half and second half of the experiment. A one-way ANOVA on KL-distance showed no reliable effect of time ($F(1, 18) = 2.83, p = .10, \eta^2_G = .06$), although the similarity between dyad members confidence distributions reduced numerically over time as seen in Figure 7.1, right panel.

Less surprisingly, also dyad members' post-social information confidence distributions were extremely correlated. The z-scores relative to random pairing were 3.68 for the correlation of means and 2.26 for correlation of standard deviations), suggesting a high level of post-social information confidence alignment.

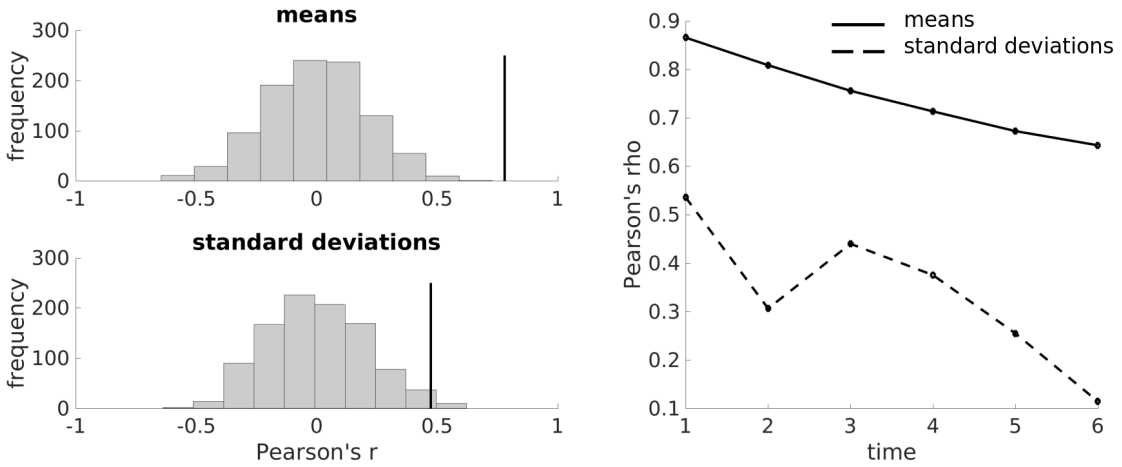


Figure 7.1: Confidence alignment in dyads in Experiment 4. Left: distribution of correlation coefficients expected by chance for both mean and standard deviation. The solid line represents the correlation observed in empirical dyads. Right: The solid line represents the correlation observed between means of confidence distributions belonging to members of the same dyad. The dashed line represents the correlation between standard deviations.

Alignment could result from participants using the confidence scale similarly across trials or from shared features of the stimuli on a trial-level. Although the box stimuli

were created independently by the two parallel staircase procedures, other variables, like room noise, might have contributed to trial-level perceptual covariation between same dyad members. To check whether this correlation arose from common trial-level variables, we computed correlation coefficients between the same dyad members' confidence judgments across trials. Empirical average correlation between dyad members confidence across trials was only 0.72 σ away from the center of the distribution obtained by random pairing (1000 repetitions), suggesting that confidence alignment was not purely due to trial-level features, but reflected a more general similarity in the use of the confidence scale across trials. The fact that the observed correlation was still above the chance distribution mean notwithstanding separate staircase procedures, can be a by-product of the similar use of the confidence scale already described between dyad members.

Finally we wanted to check whether the confidence alignment effect was due to covariation in the accuracy of each member. Because of the application of the staircase procedure to each participant, members of the same dyad experienced equal number of correct responses *on average*. However correlation between dyad members' accuracies could arise on a trial-by-trial level, that would lead to covariations also in confidence (Fleming & Lau, 2014). We computed the Goodman-Kruskal's gamma statistic, which measures the rank correlation between trial-level choice accuracy of members of the same dyad, divided by decision (pre-social information vs. post-social information) and condition (Non-Interactive vs. Interactive). A 2x2 ANOVA showed an effect of decision ($F(1, 23) = 107.22, p < .001$) but not of condition ($F < 1$) nor an interaction between these factors ($F < 1$). Gamma values collapsed across conditions showed that dyad members response accuracy was not correlated pre-social exchange ($t < 1$, against 0 mean), but it was after communication took place ($t(23) = 12.02, p < .001, d = 2.45$).

Experiment 5 and 6 replicated key Experiment 4 results on alignment (see Appendix C for details), indicating that the effect was robust across changes in experimental conditions and incentive schemes. Different explanations could be offered to explain the confidence alignment results of Experiment 4. The following section will describe the three main hypotheses considered, namely the pre-existing similarity hypothesis, the behavioural anchoring hypothesis, and the strategic alignment hypothesis. Experiment 7 was designed to discriminate among the three alternative hypotheses by manipulating the possibility to interact with others. During the first phase of the experiment, participants performed the dot-count task alone and their individual confidence distributions were recorded. In the second phase they interacted with their dyad partner in a similar manner to the Interactive conditions of Experiments 4-6. Finally in the third phase they performed the task alone once more.

Three alternative explanations

The first hypothesis for the emergence of confidence alignment - named the "similarity hypothesis" - claims that the recruitment procedure introduced a selection bias whereby people with shared confidence variance were recruited in pairs. The recruitment advertisement asked each volunteer to bring along a friend of theirs of the same gender on the day of the experiment. Several studies in social psychology have shown that a bidirectional reinforcement loop exists between liking, similarity, and interaction (E. Smith & Mackie, 2007). People who spend more time together are known to align both behaviourally and emotionally (Griffitt, 1969; Zajonc, Adelmann, Murphy, & Niedenthal, 1987). On top of that the fact that members of the same dyad were of the same gender could have had the effect of introducing further covariation in their confidence (Barber & Odean, 2001; Buchan et al., 2008). Thus members of the same dyad shared similar confidence traits due to their past interactions and demograph-

ics, which naturally led to greater confidence correlations compared to members of different dyads.

An alternative hypothesis - the "anchor hypothesis" - is that participants' specific use of the confidence scale provided a behavioural anchor that their partners' used to drive their own judgments. Anchoring is a well-known heuristic and its effects are robust and difficult to erase even when participants are informed that the anchor provides no useful information (Englich, Mussweiler, & Strack, 2006). Phenomena of behavioural contagion are well known in the social psychology literature (Dezecache et al., 2013; C. D. Frith, 2007; Moussaïd et al., 2015; Wheeler, 1966) and in the animal kingdom (Massen, Šlipogor, & Gallup, 2016). According to this view, the simple fact that participants knew how the other person was using the scale might have influenced them to use it in a similar manner (e.g. clustering to the extremes or to the middle part). Notice that in previous experiments participants did not have to be in the Interactive condition in order to know how their partner was using the scale as in all conditions confidence judgments were shared between dyad's members. According to this explanation alignment is irrelevant to the specific task used and does not reflect a reward maximizing strategy.

Finally, a final hypothesis - the "interaction hypothesis" - is that during interaction, confidence alignment represents an effective strategy to maximise sharing of task-relevant confidence variance while minimising the effects of task-irrelevant variance. Confidence is known to be affected by several task-irrelevant variables, including gender (Barber & Odean, 2001), profession (Broihanne et al., 2014), mental health (Huq et al., 1988), personality (Campbell et al., 2004) and culture (Mann, 1998). These "trait" variables affect the overall confidence of an individual. Variability in trait confidence however does not carry any information about the perceptual stimulus and should thus be disregarded. Variation in "state" confidence on the contrary

reflects the internal variation of perceptual uncertainty around an individual "trait" mean. Assuming that the most influential opinion is the most confident one, misaligned partners risk to generate a situation where one individual (the overconfident one) is consistently more influential than the other one. In a joint decision task where the dyad always chooses the option supported by the highest confidence (Bahrami et al., 2010; Pescetelli et al., 2016), this situation would mean that dyad's accuracy cannot be expected to exceed the accuracy of the most confident participant. Aligning confidence distributions has the effect of reducing the effect of trait confidence variation preserving state confidence variability. Group accuracy can thus overcome individual accuracies by making underconfident individuals more influential on group choices.

The interaction paradigm was modified to disentangle the first two hypotheses from the last one. Participants performed the dot task alone at first, then with modalities similar to the Interactive condition described in Experiments 4-6 and finally alone again. Confidence distributions were recorded before, during and after interaction with their partners and alignment between partners was computed for each phase. If confidence alignment is a product of unspecific increased similarity due to past interactions ("similarity hypothesis") we should expect to observe alignment from the very beginning of the experiment. Confidence sharing among members should not be necessary for confidence alignment to be observed and alignment should not differ across phases (Figure 7.2). On the other hand, under the anchor hypothesis, we should expect that little or no alignment should be observed in the present paradigm, because extensive initial experience with the task performed alone should anchor participants to their own independent confidence judgments instead to their partner's (Figure 7.2). Finally, the interaction hypothesis predicts that a steep increase in confidence alignment should be expected under the new paradigm as soon as participants get to know each other's confidence distributions, as a product of an

optimisation strategy that allows participants to reduce the harmful effect of variance in trait confidence and maintain useful variance in state confidence (Figure 7.2).

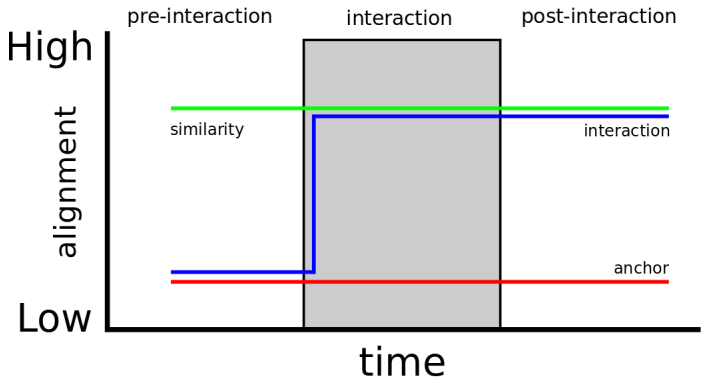


Figure 7.2: Predictions made by the three hypotheses under consideration about the alignment pattern that should be observed in the experimental data.

Experiment 7

Method

Participants Participants (mean age: 21.14±2.70) formed 24 dyads (19 female dyads). Volunteers were recruited through local advertisements websites and University recruitment platform. Interested volunteers were asked to bring a friend of the same gender along on the day of the experiment. Participants were compensated for their time and according to performance with money and/or university credits.

Task and Manipulation. The experiment comprised 432 trials divided in three phases of six blocks each. All trials started with the dot-task (Boldt & Yeung, 2015) previously described. Participants responded on a semi-continuous confidence scale ranging from 100% *Sure Left* to 100% *Sure Right*. Participants entered their answers with modalities described in Experiments 4-6. Each side of the scale comprised 50 levels. In phases I and III, after both participants had confirmed their answers, feedback was given about each participant's accuracy and a new trial began. In phase II of the experiment, after both participants had confirmed their individual answers,

the social exchange part began with modalities similar to the Interaction condition already described in Experiments 4-6. As in Experiments 4-6, the social exchange part of the trial lasted for 4 seconds, during which time the cursors' positions along the confidence scale were continuously recorded at 5 Hz (21 data points per trial). Feedback was then provided and a new trial began (see Figure 7.3). Participants were informed that an extra bonus could be achieved by accurately reporting their final confidence during individual decisions and their continuous confidence during their social decisions.

Before the beginning of phase I, one block of six trials served as practice with the perceptual task. Before the beginning of phase II, participants received a new set of instructions explaining the input modalities and incentive scheme during the social exchange part, according to the Interactive condition modalities. No practice was given.

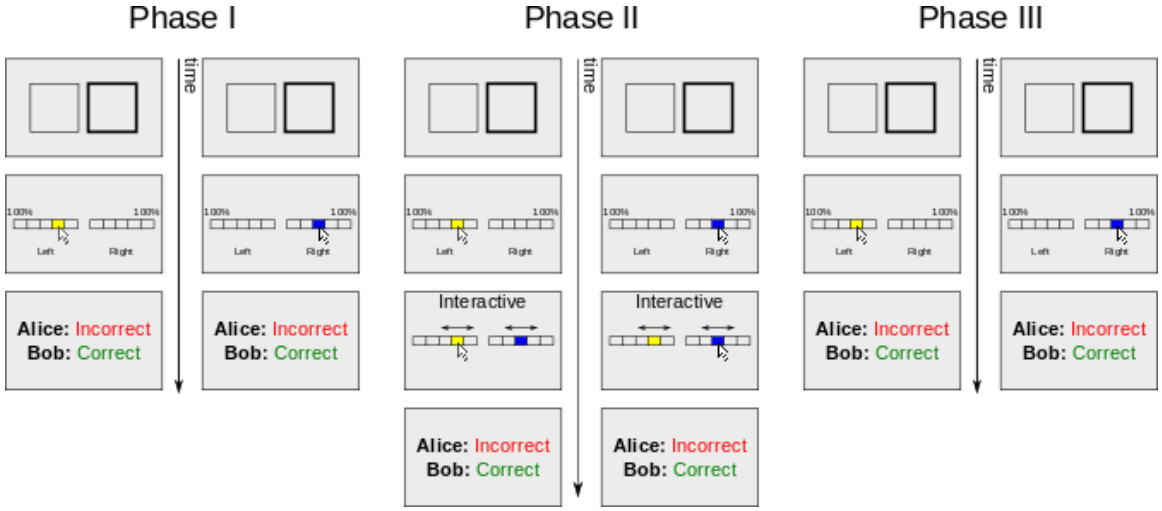


Figure 7.3: Experimental paradigm employed in Experiment 7. Participants perform the perceptual task on their own expressing on each trial a choice and a confidence judgment. In the middle phase of the experiment participants are shown their partner's choice and confidence after expressing their own private views and are asked to update their initial answers in a similar fashion as in the Interactive condition of Experiments 4-6. In the last phase, participants are again asked to perform the task alone and their choice and confidence are recorded on every trial.

Results

Preliminary analysis. All the routine analyses that were performed on previous experiments were also performed on the current study although not central to the hypotheses at test. This was done to ensure that the pattern of results did not differ from previous experiments. Results showed that the pattern of updates was similar to Experiments 4-6, with larger updates taking place around 1 seconds and reaching stability by the end of the update window. Similarly to Experiments 4-6, irrational confidence changes during the social phase were more frequent in disagreement than agreement ($F(1, 47) = 8.70, p = .004, \eta^2_G = .08$) and both choice accuracy and calibration significantly improved from pre- to post-social information phase ($t(47) > 5.10, p < .001, d \sim 1$).

Confidence distributions. Figure 7.4 shows the individual confidence distributions over all confidence levels of the scale, divided by experimental phase. Each row represents one dyad, with phases coded as different colours. Participants belonging to the same dyad are placed near each other. Some clustering can be noticed around text landmarks (60%,70%,80%, etc.), probably due to instruction's emphasis on confidence calibration.

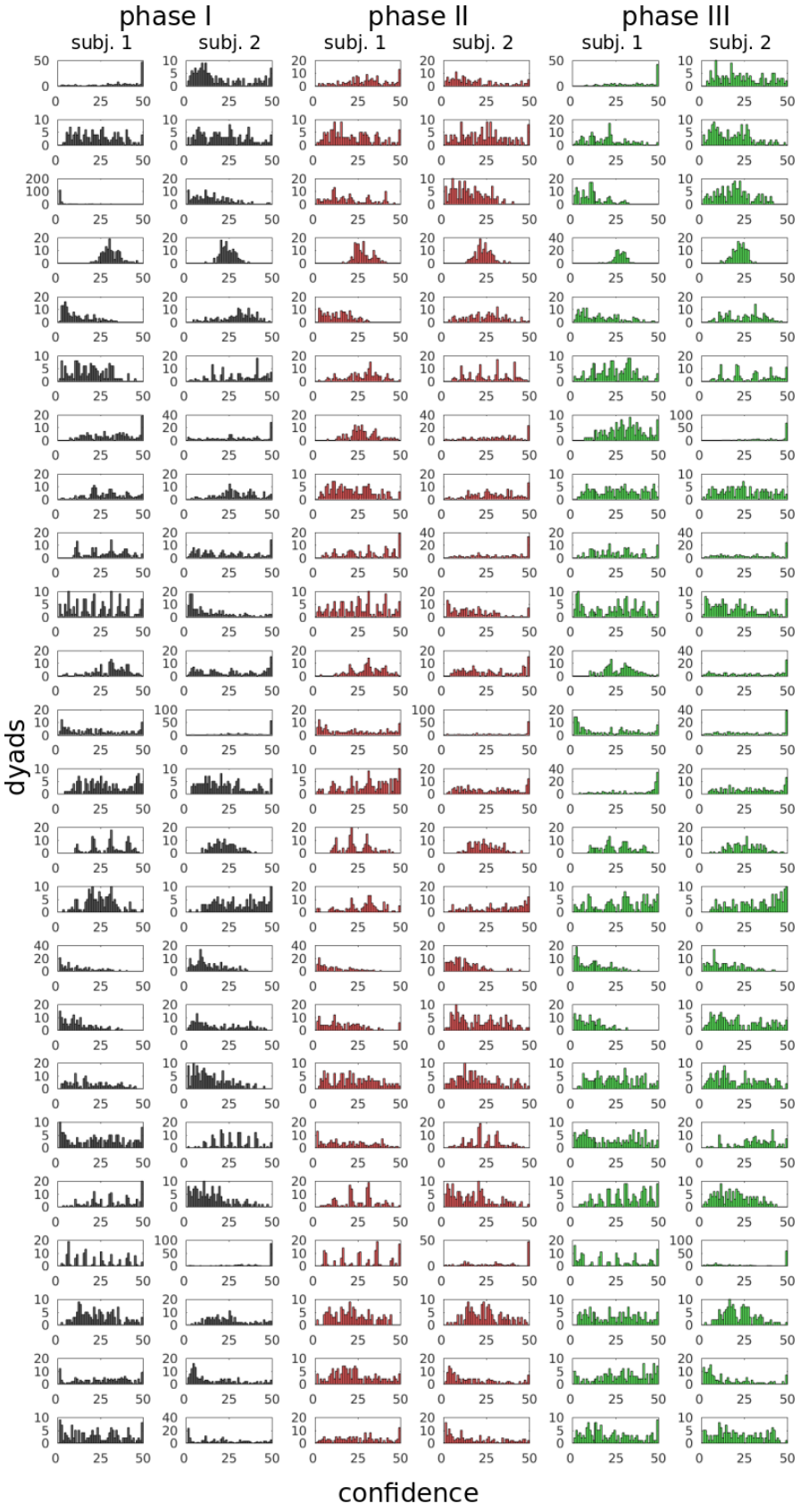


Figure 7.4: Individual confidence distributions over available confidence levels (x-axis), divided by experiment phase (color code). Each row corresponds to one dyad; y-axis indicates frequency of use.

Confidence alignment. The central test of the experiment was a comparison of confidence alignment over the three experimental phases. A bootstrapping procedure, as described in previous experiments, was applied to to compare observed correlation values in each phase with what is to be expected by randomly pairing participants. Results show that compared to a bootstrapped distribution the observed correlation between confidence means went from the 74 percentile in phase I to 97 percentile in phase II and 94 percentile in phase III (Figure 7.5, top row). Correlation of standard deviations on the contrary remained high during the whole experiment, as shown by the bottom panels (98, 99 and 99 percentile, in the three phases respectively).

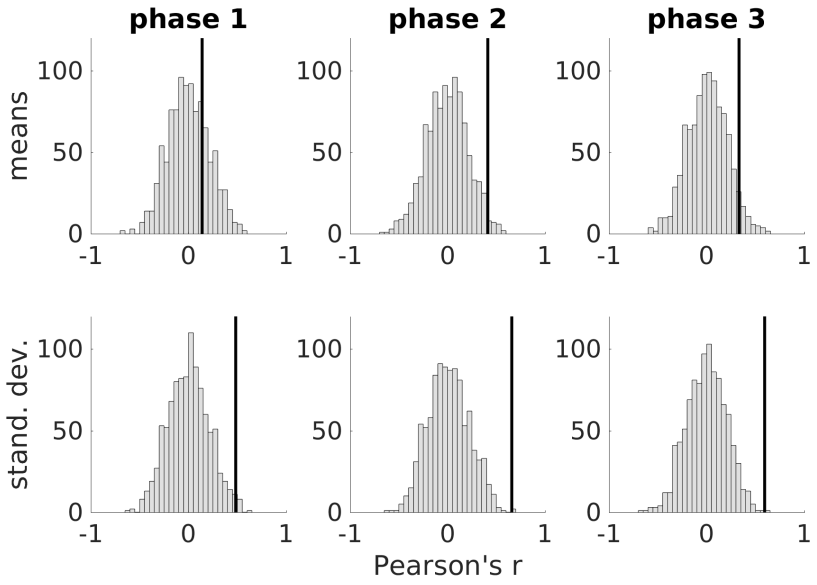


Figure 7.5: Bootstrap procedure across the three experimental phases. Correlation coefficients between means (top row) and standard deviations (bottom row) of the empirically observed confidence distributions are compared against the distribution of correlation coefficients expected by a random pairing procedure.

Empirical Pearson's correlation coefficients between pre-advice confidence means of participants belonging to same dyad is also shown in Figure 7.6 top panel. The observed correlation sharply increased during the interaction phase of the experiment, suggesting a causal relation with the experimental manipulation. Results were not strongly altered when controlling for variations in accuracy and threshold using a

partial correlation analysis (red bars in Figure 7.6 top panel). Compared to Experiment 4, correlation coefficients are smaller in magnitude, probably indicating that some anchoring to confidence scale from phase I must have taken place.

The increased correlations offers support for the hypothesis that alignment significantly increased over experimental phases. To test this formally, alignment was reformulated, for each dyad and each phase, as the absolute difference in mean confidence between the two dyad members, which produced a value for each dyad instead of across dyads. Alignment reflects the fact that participants shifted their pre-advice confidence distributions toward each other, so values closer to zero indicate greater alignment. Figure 7.6, bottom panel, plots distance as negative values to correspond with correlation coefficients plotted in the top panel, where more positive values indicate closer alignment. Values were then subjected to a one-way ANOVA which revealed that phase significantly altered alignment ($F(2, 46) = 4.48, p = .01, \eta_G^2 = .06$). Planned comparisons showed that overlap was significantly larger during the interaction phase than during the pre-interaction phase ($t(23) = 2.80, p = .01, d = 0.58$), confirming that alignment was directly caused by the experimental manipulation. A marginally significant difference was also observed between phase III and phase I ($t(23) = 1.76, p = .09, d = 0.35$), but not between phase II and phase III ($t(23) = 1.22, p = .23, d = 0.25$).

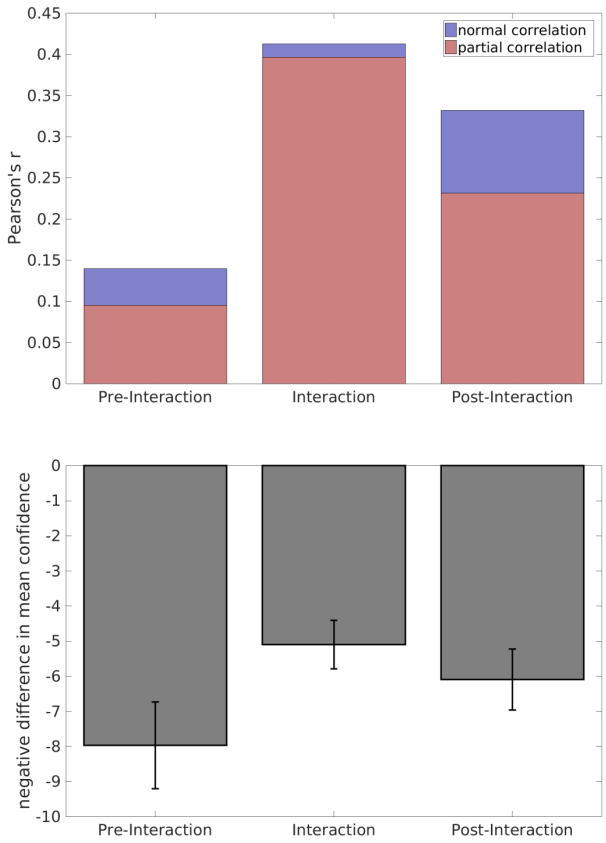


Figure 7.6: How confidence alignment between participants of similar dyads changed over the course of the experimental phases. Top panel: Pearson's correlation coefficient between means of the pre-advice confidence distributions of members belonging to the same dyads (blue bars); similar results were obtained when controlling for variations in accuracy and perceptual threshold (red bars). Bottom panel: negative distance between confidence means of pre-advice distributions. Also this method to quantify alignment suggests alignment increases from phase I to phase II.

The next set of analyses investigated what drives alignment and whether confidence shifts are symmetrical between the two participants. Noting that task difficulty was set to match the accuracy of all participants at $\sim 71\%$, two potential factors were considered, namely average confidence and confidence calibration during the pre-interaction phase. A linear regression tested whether participants tended to shift their confidence toward the most confident participant or toward the most calibrated one. The absolute difference between phase I and II in the mean pre-advice confidence acted as dependent variable. Average confidence and confidence calibration in phase

I acted as independent variables. Results show that smaller confidence shifts occurred in participants who were highly calibrated ($\beta = -20.59, SE = 5.39, p < .001$) or highly confident ($\beta = -0.44, SE = 0.11, p < .001$). Interestingly, the interaction term between the two factors was also significant and positive ($\beta = 0.82, SE = 0.20, p < .001$), suggesting that greater shifts were observed in more calibrated participants when they were also highly confident. Conversely, lower shifts occurred in highly confident but uncalibrated participants and in highly calibrated but uncertain participants.

	Estimate	SE	tStat	DF	p
intercept	12.74	2.9237	4.3573	44	$p < .001$***
calibration	-20.594	5.3974	-3.8155	44	$p < .001$***
confidence	-0.4488	0.11693	-3.8383	44	$p < .001$***
calibration:confidence	0.82702	0.20815	3.9732	44	$p < .001$***

Table 7.1: Linear coefficients and standard errors of a multi linear model run to predict participant's absolute confidence shift from phase I to phase II of the experiment.

The benefits of confidence alignment. The results presented so far are inconsistent with the similarity hypothesis, which states that alignment is the product of pre-existing similarity traits between members of a same dyad. The results are also inconsistent with the anchor hypothesis, which predicts that anchoring effects should not be observed if participants had the chance to consolidate their use of the confidence scale before social exchange. They are, however, consistent with the interaction hypothesis, stating that alignment should be observed as a consequence of confidence sharing between the two members.

A more flexible version of the anchor hypothesis can however be formulated, which states that people can still get biased towards their partners by looking at the their partner's use of the confidence scale when this information becomes available during the interaction phase. This modified version of the anchor hypothesis thus makes identical predictions than to the interaction hypothesis (namely increased alignment

in phase II) and it would be impossible to distinguish between the two only based on the analyses presented so far. The modified anchor hypothesis differs from the interaction hypothesis, however, in that it does not make specific predictions about whether and how accuracy would change from phase I to phase II. On the contrary, the interaction hypothesis predicts that alignment should be correlated with good performance, as this is a strategy to optimise information sharing about the individual likelihood of a correct response.

Accuracy improvement was defined as the difference between pre- and post-social exchange choice accuracy during phase II, averaged across dyad members, with dyad as the random effect. A marginal negative correlation was found between accuracy improvement and difference between confidence means of members belonging to a same dyad during phase one ($r(22) = -.37, p = .07$), suggesting that greater accuracy improvements are observed when participants' confidence distributions started off relatively aligned to each other. This result provided preliminary evidence for the interactive interpretation.

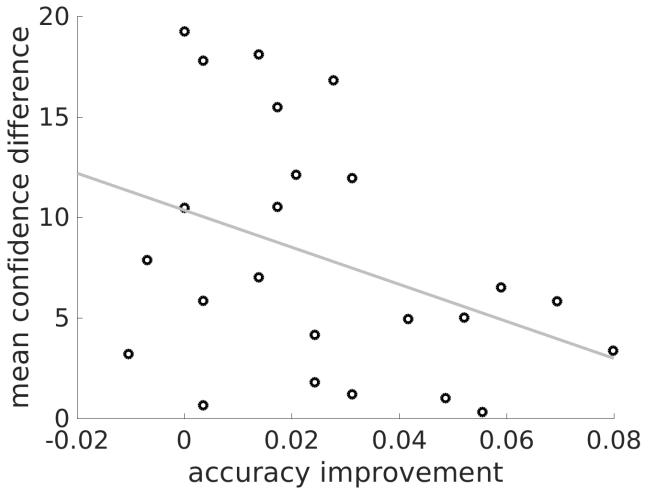


Figure 7.7: Marginally significant negative correlation existing between confidence alignment during pre-interaction phase (phase I) and accuracy improvement during the interaction phase (phase II). The correlation suggests that greater initial misalignment decreases the chances of accuracy improvement at interaction.

A final analysis estimated the accuracy level that would be expected had the participants not shifted their confidence distributions towards each other. The approach taken was to estimate what pre-advice confidence $\hat{C}$ was to be expected without alignment and then add the empirically observed confidence shift. $\hat{C}$ was estimated as:

$$\hat{C} = \mu_{pre} + \sigma_{pre} * z \tag{7.1}$$

where z is the z-score of pre-advice confidence observed during the interaction phase (phase II) and μ_{pre} and σ_{pre} are the mean and standard deviation of the pre-advice confidence distribution observed during the pre-interaction phase (phase I). To estimate the hypothetical final accuracy of participants had they not aligned confidences, observed confidence change δ_C (equation 2.3, Chapter 2) was added on top of $\hat{C}$: $\hat{C}_{post} = \hat{C} + \delta_C$. This resulted in an estimated post-advice confidence according to pre-alignment coordinates, in which post-advice confidence values lower than zero indicate changes of mind from pre- to post-advice decisions. Thus, estimated post-advice accuracy $\widehat{Cor}$ differed from observed post-advice accuracy Cor whenever $sign(\hat{C}_{post}) \neq sign(C_{post})$, indicating changes of mind that would have not occurred without alignment (or vice versa resisting a change of mind that would have occurred without alignment). Results are shown in Figure 7.8 left panel and show that observed accuracy (Cor) was greater than estimated accuracy ($\widehat{Cor}$) indicating that alignment produced grater average final accuracy than if alignment had not happened ($t(47) = 5.10, p < .001, d = 0.75$). These findings provide further support for the interaction hypothesis, which proposes that confidence alignment occurs to facilitate effective sharing of information in social exchange.

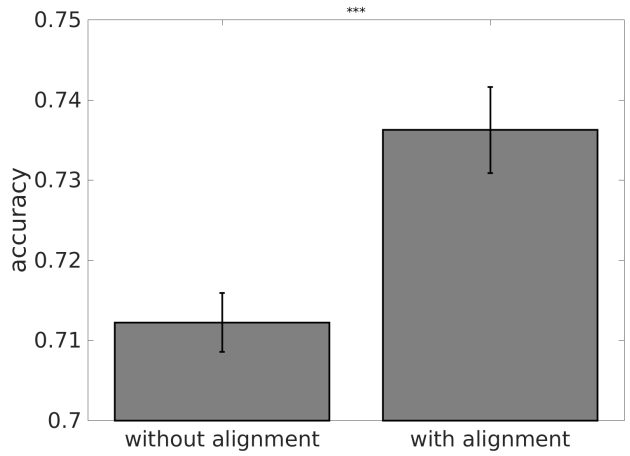


Figure 7.8: Accuracy that would be expected had the participants not shifted their confidence distribution toward one another.

Discussion

Experiment 7 aimed to disentangle contrasting explanations of confidence alignment between dyad members. Two of them, the similarity and anchor hypotheses, ascribed confidence alignment to contingencies that were irrelevant to the task. In contrast, the interaction hypothesis proposes that confidence alignment is an effective strategy to reduce variance in trait confidence (which here is not relevant to the task) and maximise variance in state confidence (task-relevant). Based on this interpretation (a) alignment should be observed only after participants learn the other person's confidence distribution (e.g., mean confidence) and (b) it should produce benefits to accuracy. Both predictions were met in the observed data, suggesting that confidence alignment is a strategy emerging for optimising cooperation.

It should be pointed out that the alignment strategy is not necessarily a conscious decision made by participants, but must be interpreted as a behavioural coordination strategy the dyad converges to when trying to solve the task. Confidence alignment might well be a general trait of human social interactions that allows people to solve group coordination problems (Bang, Summerfield, & Lau, 2017; Fusaroli

et al., 2012; Shea et al., 2014). Alignment was in the present task the optimal strategy to adopt given that participants shared equal accuracies. However in situations where partners differ in their expertise alignment can produce suboptimal results. If we assume members are calibrated, whenever accuracy levels differ across group members confidence alignment makes the accurate partner under-confident and the inaccurate partner over-confident, a situation that produces over-reliance on the inaccurate person's judgment. Confidence calibration has been shown to be beneficial to group performance (Pescetelli et al., 2016). When listening to advice, people have nevertheless been shown to be more sensitive to absolute confidence than calibration (Pescetelli et al., 2016; Price & Stone, 2004), at least when gathering objective feedback is impossible or costly (Sah et al., 2013). Finally confidence alignment in situations of great accuracy imbalance can produce an equality bias (Mahmoodi et al., 2015), whereby group choices favour too much inaccurate members and too little accurate ones.

As it can be seen from Figure 7.6 the effects of confidence alignment carried on during phase III of the experiment, although reduced in size. An interesting question is whether the confidence alignment effect can be used to favour confidence calibration in uncalibrated participants. An ongoing collaborative project (Van den Broeck, Pescetelli, & Yeung, 2017) investigates these questions by pairing participants with virtual advisers devised to have a complementary confidence distribution to the participant. If the participants align confidence with these virtual "tutors", this should have the effect of reducing confidence in over-confident individuals and increasing confidence in under-confident ones. If successful, the implications will impact clinical populations (David, Bedford, Wiffen, & Gilleen, 2012) as well as produce applications for improved group cooperation and performance.

The following Chapter will summarise the main findings of this work and identify common underlying issues and threads. Unanswered questions will be stated and future possible directions suggested.

8

General Discussion

Chapter Abstract

This Chapter concludes the present work on the use of metacognitive processes to guide the navigation of our social world. I summarise the paradigms used and the results found. I will highlight common threads that connect the different pieces together and draw integrative conclusions. Finally, the present work is put into the broader context of existing literature and emphasis is placed on how it expands this literature and what new questions it raises.

First line: a Judge-Adviser paradigm

Work motivation

The research presented in this thesis was motivated by existing literature in decision making and social psychology suggesting a direct link between judgments of confidence and the influence of people's opinions on others' believes and group choices (Bahrami et al., 2010; Bonaccio & Dalal, 2006; Price & Stone, 2004). Although confidence is an indicator of the probability that a given choice is correct, people have been shown to vary in the degree to which their confidence judgments are accurately indicative of their actual underlying probability of reporting a correct response (Baird, Smallwood, Gorgolewski, & Margulies, 2013; Fleming et al., 2010; Song et al., 2011), an ability called calibration (Fleming & Lau, 2014). Calibration is important for group performance as it favours the selection of the response associated to the least uncertainty (Pescetelli et al., 2016). Previous research shows that when people have access to objective feedback, they trust calibrated individual over confident ones (Tenney et al., 2007). However when feedback becomes costly to acquire or impossible to define, confidence is the best predictor of influence and trust (Sah et al., 2013). Thus confidence (and its calibration) can be regarded as a social signal that allows others to know when our judgment is most to be trusted.

A large tradition in decision-making and learning has focused on models of reinforcement learning (Sutton & Barto, 1998). These models of decisions are fundamentally based on an outcome signal that is typically associated with rewards and feedback. But situations of feedback absence are abundant in our daily life (Weiss & Shanteau, 2003), yet in these contexts people are still able to make decisions and learn (Guggenmos et al., 2016). In the absence of feedback, group performance in joint decision tasks has been shown to exceed performance of individuals as long as group members are allowed to share the uncertainty in their choices (Bahrami et al., 2012a). These observations raise the question of what signals participants follow

when learning in the absence of objective reliability cues (own and external sources'). Given that confidence is defined as the (subjective) estimation of the probability of an event happening (e.g. a decision made being correct), it has here been suggested that it can be used as a noisy internal proxy to feedback in situations when objective feedback is missing.

The first line of studies put together the findings reviewed above into a unitary and coherent paradigm in which the effects of reliability and calibration were studied with and without objective feedback. Three studies aimed to (a) identify the likely mechanisms driving learning about the reliability of social sources in the absence of objective feedback; (b) understand how different adviser profiles affect the formation of trust in the adviser; (c) understand whether internal metacognitive signals in the advisee play a role in the perception of the advice. The paradigm used to investigate such questions is a modified Judge-Adviser System paradigm, similar to that often used in organisational psychology experiments (Sniezek & Buckley, 1989, 1995). In this new version of the paradigm, a participant (the judge) makes a decision about the state of a perceptual quantity (e.g. the number of dots contained in a delimited field), then is shown the opinion of a virtual partner and is finally allowed to adjust his/her initial opinion based on the advice received. Manipulation of the advice profile and analysis of belief updates as well as explicit trust judgments given by participants allow the researcher to quantify what factors influence trust formation and advice perception when feedback is explicitly provided versus when it must be inferred from contextual and internal cues.

People distinguish advice even in the absence of feedback

In the absence of objective feedback, classical reinforcement learning models, like delta rule, have difficulties explaining the mechanisms of learning because they are all highly dependent on a prediction error term that quantifies the difference between an expected reward and an outcome, which is missing in no-feedback scenarios. But in

many everyday cases, outcome rewards and feedback signals are known to co-vary with several satellite variables, including agreement among different sources and internal metacognitive estimates like uncertainty and confidence judgments. In principle, any variable correlating with the outcome (choice accuracy in the present work) can be used as a (noisy) replacement for the actual outcome itself. Experiments 1-3 showed that when available agreement provides a simple heuristic cue that participants use to estimate advice reliability.

The results of Experiment 1 confirmed and expanded previous results (Tenney et al., 2007) showing that when feedback is available, greater advice influence and trust were predicted by both obvious advice features, like accuracy, and more subtle cues, like calibration. Importantly, when feedback was removed the overall pattern of results did not substantially change suggesting that participants were still able to recognise (and value) accurate advisers and calibrated advice.

Across the three experiments, results indicate that people seem to use agreement in confidence as a cue to adviser accuracy. A simple formal model of advice-taking behaviour and trust formation shows that good approximations of participants' trust and belief update patterns could be reproduced in variants that had access to metacognitive information (Confidence variant). The success of this model was benchmarked against a simpler model without metacognitive access (Consensus variant) and to a model with access to objective feedback (Accuracy variant). Appendix A describes how similar conclusions are reached by a delta-rule model adapted to the conditions of Experiment 3.

Inference in the absence of feedback: highs and lows

The agreement-in-confidence strategy works well when advice is independent from one's own judgment, like in Experiment 1. But issues arise when similar biases affect adviser's and advisee's judgments. Experiment 2 contrasted the Consensus and Confidence model variants by dissociating the accuracy and agreement rates of

different advisers. This dissociation was possible by manipulating the probability of agreement conditional on the participant's accuracy. So, for instance, an agreeing but inaccurate adviser would tend to agree with the participant on trials when the participant is wrong.

Results showed that when provided with feedback participants trusted more and relied more on accurate advisers over inaccurate ones and on agreeing advisers over disagreeing ones. The effect of agreement was surprising given that participants could form an objective representation of the advisers' accuracy. Given the task, this was a suboptimal strategy as consensus was not predictive of accuracy. It shows that inferring reliability from agreement with one's own opinions is a cognitive heuristic that participants cannot avoid using, rather than a strategy adopted only when no other objective standard is available. When feedback was removed, agreement of the advice with participant's judgments became the strongest predictor of trust and influence. Although an effect of accuracy on advice influence could be seen, the results showed that participants heavily relied on easily available cues (agreement) rather than on implicit ones (metacognitive signals). Interestingly however, the Confidence model produced patterns of results that were more similar to the Consensus model than to the Accuracy model. The reason is that this model, although it weighted consensus by internal confidence, still needed consensus to form a trust judgment. In other words, a metacognitively informed strategy must rely on agreement while an agreement strategy does not have necessarily to rely on confidence.

Results of Experiment 2 showed how agreement probability with one's own opinions is a powerful cue to the adviser's perceived reliability. Experiment 3 overcame the limitations of Experiment 2 by removing differences among advisers in agreement rates. In this new experiment, agreement probability across advisers did not differ overall but differed based on the participant's pre-advice confidence. By manipulating the probability of agreement conditional on participant's accuracy and initial

confidence, we obtained an adviser who shared the same bias as the participant (i.e. agreement rate scaled with initial participant's confidence), an adviser who showed opposite bias (i.e. agreement probability is inversely proportional to initial participant's confidence) and an adviser who was unbiased (no difference in agreement rate over different levels of initial participant's confidence).

Results showed that when provided with feedback, participants tended to trust the anti-bias adviser the most and the bias-sharing the least, consistent with the informational value of the advice. However, when feedback was removed, participants tended to trust and rely more on the bias-sharing adviser, consistent with a confidence-based strategy. The unbiased adviser surprisingly showed the lowest trust ratings and influence, while the anti-bias adviser lay in between the first two. As predicted, a model endowed with metacognitive access trusted the bias-sharing adviser more than the anti-bias adviser, while a model without access to metacognitive information did not discriminate among advisers. Far from being unrealistic laboratory constructions, manipulations of Experiment 2 and 3 reproduce realistic scenarios where people can cluster around specific biases, use different sampling strategies of their environments or simply vary in perceiving external information due to differences in personality traits or psychiatric condition.

The mechanisms described by the model and the behaviour observed in participants suggest that agreement with someone else is a simple and effective strategy to estimate the underlying accuracy of a (social) informational source when feedback is not available or costly to acquire. Results however also point at the risks to which such a strategy might be prone. The findings above highlight the fact that our participants assumed by default that the other person's opinion was independent from their own. This follows from the fact that agreement among two sources speaks in favour of the reliability of the judgment agreed upon only if the sources are independent of each other. As an example, imagine you are the editor of a peer-reviewed journal

and you are asking for the opinion of different reviewers about the quality of a paper that has been submitted to you. If the reviewers agree that the paper is excellent you can be sure that the paper is indeed excellent *only if* you know that the reviewers independently scored it instead of simply passing each other notes. If agreement among reviewers comes from a correlation among their judgments (e.g. exchanging notes) then the reliability of the agreed judgment dramatically decreases. The fact that agreement was such a strong predictor of trust and influence in my experiments even when objective feedback was easily available suggests that participants assumed independence between their own judgment and their partners' even when objective feedback could be used to learn that this assumption was wrong (e.g. Experiment 2, feedback condition). Assumptions of judgment independence among members are also thought to explain hidden-profile results, because correlated information that is shared among group members elicits an illusion of agreement when sampling from members' views (Kerr & Tindale, 2004; Stasser & Titus, 2003). Our findings show that when the independence assumption is not met, trust can be placed on the wrong advisers (as in Experiment 3) or not placed on potentially accurate advisers (as in Experiment 2).

The presence of the agreement bias is likely to be an adaptive strategy for quickly gaining information in conditions when feedback is not readily available or costly to acquire. These situations are quite common in daily life scenarios (Weiss & Shanteau, 2003), such as in medical practice and education, although unfortunately underrepresented in the decision-making literature (although see Guggenmos et al., 2016, for an exception). Although appropriate in many naturalistic scenarios, the independence assumption that our participants seem to have made might generate quick error cascades and misplacement of trust in several modern-life environments. Increasingly more people get their news and share their views on online platforms (Mitchell, Gottfried, Barthel, & Shearer, 2016), where small-world network topologies (Watts & Strogatz, 1998) can create clustered pockets of stagnant information (Jasny et al.,

2015; Sunstein, 2001). These pockets or "echo-chambers" can easily produce agreement among their members generating a false feeling of confidence, particularly in situations where expertise is difficult to verify and feedback to acquire (e.g., political or socio-economic matters). More importantly, our findings show that, even when presented with information from sources that equally agree with one's own view, a subject might still prefer and be more influenced by partners who agree on matters for which the subject already holds a strong opinion. This leads to an inconvenient paradox whereby partners who can potentially improve our judgments the most (namely the ones who do not share our biases) are also the ones who are least likely to be perceived as reliable.

Trust and Influence: correlated but potentially dissociable

The use of explicit reports of trust and advice influence as dependent variables allowed me to test for dissociations between explicit attitudes toward advisers and observed behaviour (i.e., confidence update). In all three experiments, the trust and influence measures tended to be strongly correlated and their results consistent with each other, indicating that, when asked to report their trust explicitly, participants gave answers that matched their observed opinion update behaviour.

The two measures however seemed to be dissociated in the No-Feedback condition of Experiment 2, where participants did not show an effect of accuracy when directly asked, but did rely on accurate advisers more than inaccurate ones when updating their confidence. Two explanations were offered. The first one suggests that perhaps the pre-processing steps used to generate the trust measure caused some information loss. As detailed in Chapter 4, analysis of participants' unprocessed responses showed that this explanation is unlikely. Indeed, even when directly asked about advisers' objective accuracy (Question 1 of the questionnaire), responses showed an effect of adviser agreement rate but not of accuracy, indicating that participants could not distinguish advisers based on their objective reliability.

A more likely explanation is in terms of halo-dumping effect (Clark & Lawless, 1994), according to which participants tend to use the most available dimension to discriminate among stimuli, when offered with a unidimensional scale. It is possible that agreement rate in our experiment was such an obvious cue that, when directly asked, participants ignored the much subtler variations in accuracy (an implicit cue in the No-Feedback condition) and discriminated different advisers according to agreement only (a explicit cue in the No-Feedback condition).

The importance of judgment variability

In the absence of objective feedback, participants in Experiment 3 showed an unexpected numerical non-significant difference between the unbiased adviser and the anti-bias adviser. The unbiased adviser showed lower trustworthiness and influence compared to the anti-biased counterpart. This fact was surprising given that the unbiased adviser was expected, given its profile, to lie in between the other two (whatever direction of preference participants expressed). Although not significant, the results raise the question of why such a pattern was consistently found in both measures investigated. One possibility is that participants' strategy to form a reliability estimate about their partners was more complex than described by our simple model.

Weiss and Shanteau (2003) provided a useful framework for the definition of expertise in the absence of explicit accuracy measures. Taking medical practice - a situation where outcomes are known much later than decisions are made - as example they insightfully suggest that expertise should be defined as the ratio between between-class variance and within-class variance. In other words the judgments of an expert should be discriminative across different classes (e.g. different diagnoses for different diseases), thus creating large between-class variance, but consistent within each class (e.g. same diagnosis for the same disease), thus creating small within-class variance. It is possible that a similar logic might have been applied by the participants tested in Experiment 3. Without an external reference (i.e. objective feedback)

participants might have resorted to use internal confidence to categorise trials (e.g. low-confidence, medium confidence and high confidence trials). The agreement pattern of the unbiased adviser would have in this case showed the lowest variance across these subjectively constructed categories.

Although it is not suggested here that participants actively engaged in such mental strategy, it must be pointed out that a shared logic can explain important features characterising judgments of expertise in the absence of objective references. Indeed, low variance across classes of events (objective or subjectively defined) implies that each instance carries less information than if classes are more variable. Participants might have associated that the probability of agreeing with the unbiased adviser did not help them to discriminate trials in which they should have felt confident in their answer from trials in which they should have felt uncertain.

The second research line: Moving beyond static social stimuli

The first three experiments made extensive use of the traditional Judge-Adviser System used in social and organizational psychology (Budescu & Rantilla, 2000; Sniezek & Buckley, 1989; Sniezek & Van Swol, 2001). Although useful to understand opinion change and advice taking, the studies are also limited by the strict structure of the paradigm. Compared to real-life situations the JAS paradigm constraints social exchange into static stages where information is transmitted unilaterally (i.e. from the adviser to the participant). Novel work in social neuroscience as well as psychiatry and virtual reality shows the importance of understanding social interactions in more ecological contexts, where decision makers are active agents instead of passive observer of social stimuli (Auvray et al., 2009; Dumas et al., 2014; Froese et al., 2014; Mattout, 2012; Schilbach, 2016). A new paradigm tried to move beyond Judge-Adviser systems by including more naturalistic social exchanges based on real-time recursive interactions in a highly controlled environment.

Interaction is characterised by non-linear dynamics

Experiment 4 compared directly non interactive scenarios with interactive ones, keeping however the amount of task-relevant information constant. The results showed an overall agreement in the pattern of opinion updates between the two conditions, but importantly also some differences. In interaction, more update transitions and longer times to reach equilibrium were observed during the update window, indicating greater recursivity. Interaction led participants to an increase of the agreement effect (i.e., the increase in confidence after agreement) and a reduction of the disagreement effect (i.e., the decrease of confidence observed after disagreement). The results can be explained by a modulation of partner's confidence change during interaction. When participants could see how their own opinion affected their partners, confidence changes of members of a same dyad became coupled. In agreement, greater increases in partners' confidence pushed participants to increase their confidence more, thus suggesting confidence escalation, particularly when initial opinions were uncertain. On the contrary larger decreases of confidence after disagreement in one participant corresponded to lower decreases in their partner, probably due to the fact that opinion changes were interpreted as the participant holding unreliable opinions.

Experiment 4 seemed to show that simply changing the dynamics by which two individuals share the same amount of information affects the final state of the dyad (described by each member's opinion). Experiment 5 was run with the goal of replicating the main results found in Experiment 4 and disprove competing explanations of the same. One of the alternative hypotheses was that participants escalated their confidence simply because they did not remember their initial opinion. Experiment 5 thus added an interactive condition where the original opinion of the participant was presented along with their current opinion and current partner's opinion. Results showed that the addition of an initial opinion reminder did not affect the overall effect of the interactive condition, thus rejecting the memory failure hypothesis.

Another alternative explanation for the interactive effects was that participants showed escalating updates because they tended to forget the initial opinion of their partner and used their partner's current confidence instead. Using a similar logic to Experiment 5, Experiment 6 tried to disprove this hypothesis by introducing an interactive condition where participants were reminded about the initial opinion of their partner. It compared such condition with the original interactive condition (Experiment 4-5) and the memory condition with reminder of one's own initial opinion (Experiment 5). Results showed that the overall behavioural pattern was not affected by the introduction of opinion reminders although the presence of the partner's opinion reminder seemed to make this opinion more salient and thus influential. Given that the major interactive effects (e.g., confidence changes coupling, enhanced agreement, irrational changes asymmetry etc.) replicated in all conditions and differences among conditions were rare in all analyses considered, the memory hypotheses (either related to the self or to the partner) could not be accepted.

What you tell me is not what I hear

In Experiments 4-6, participants' most common behaviour was to not update confidence at all. Even when participants did update their confidence, on a small but consistent amount of trials their updates were irrational, increasing their confidence after disagreement and decreasing it after agreement. A significant interaction between consensus and condition indicated that more irrational increases and fewer irrational decreases were observed in Interactive than Non-Interactive trials. Interestingly, irrational changes were asymmetrical for agreement and disagreement, so that increases after disagreement were more frequent, on average, than decreases after agreement, a phenomenon that cannot be easily explained by static aggregation rules like averaging, summing or Bayesian integration (Bang et al., 2014; Larrick & Soll, 2006; Pescetelli et al., 2016), but is predicted by a recursive update model.

Social information seemed to be discounted by participants, who thus showed an egocentric bias (Yaniv & Kleinberger, 2000), whereby one's own opinion is weighted more than somebody else's opinion. Average weights on partners' opinions were around 0.50-0.60, meaning that partners' opinions were weighted about half of the partner's actual stated judgments. The effects were however modulated by dominance, condition and consensus. The effect of consensus indicated that participants also showed a confirmation bias (Nickerson, 1998), whereby partners' opinions were discounted more by the participant if they were against the participant's original opinion and discounted less if they supported it.

The fact that both members of a same dyad provided confidence judgments both before and after exchanging social information, allowed me to compare the objective social evidence supporting a participant's judgment provided by their partner and its effect on the participant's opinion, thus effectively quantifying social information perception. Such comparison suggested that social information was categorised by participants into either highly supportive of their own initial opinion or neutral. Such was the case even though objective social information was uniformly spread across the whole range (from confident disagreement to confident agreement). An intriguing idea is that participants are not simply using social information as it is communicated by their partners but are actively trying to infer the state of their partners (partner is correct vs. partner is wrong). On each trial, once this categorisation is performed, participants act accordingly, either ignoring advice when they think advice is wrong or completely following the advice when they think it is correct. Consequently, this active inference process allows to transform uncertain supporting evidence (e.g., uncertain agreement) so to maximise the impact of social information on one's own judgment.

Confidence reflects relevant and irrelevant cues

In accordance with an active inference hypothesis, Experiments 4-6 showed that participants' evolving confidence in their choice during interaction was influenced by how much their partners were updating confidence themselves. These results suggest that, instead of limiting themselves to task-relevant information, participants might have used all cues that were available to them to infer their partners' likelihood of being correct.

Given the evolutionary pressure to quickly gather information in situations of uncertainty, it makes sense that the human cognitive architecture has adapted to use all available information to make an inference (e.g., is my partner likely to give a correct response?). This strategy allows to quickly reach decisions even when information is scarcely available by taking into account a host of circumstantial variables that are known to co-vary with problem-specific evidence, but that are not themselves strictly task-relevant (Gigerenzer et al., 1991; Gigerenzer & Todd, 1999; Hertwig & Gigerenzer, 1999).

Interpreting someone's changes of mind as another cue for confidence is sensible in many daily life situations. Indeed, if somebody's beliefs are fickle we have reasons to believe s/he must be uncertain. In the case of social interaction however this leads to unwanted self-reinforcing dynamics as one person is using the impact that his/her own opinion is having on the other person as evidence for the opinion itself.

Imagine as an example two lovers Narcissus and Echo[1] who disagree on a binary decision (get married or not?). Narcissus strongly wants to get married and manages to convince poor Echo. Echo's change of mind should not be taken by Narcissus as evidence for the quality of his choice because this would fail to take into account the fact that Narcissus himself convinced Echo to change her mind. The Narcissus-Echo

[1]From the myth told in Ovid's Metamorphoses: The nymph Echo manages to make her feelings known to the self-absorbed Narcissus by repeating his last words. The story ends with a love rejection and a suicide to remind us of the disadvantages of interpreting information originating from ourselves as evidence of our good skills.

effect does not take into account the correlations in the sources of error emerging when adopting the fickle-belief heuristic to cases involving the self. Although the strategy of interpreting unstable opinions as a proxy to infer probability correct seems to be adaptive in many situations (e.g. when judging others' changes of mind originating from others' opinions), it has the risk of leading to over-confident errors when applied to interactions involving the self, particularly if members are uncalibrated (Echo changes her mind but Narcissus had made a poor judgment).

Once again, we find that participants assumed independence between their partner's confidence (and change thereof) and their own judgments, an assumption that in the case of interactive conditions was wrong. Nevertheless, as evidenced by Experiments 4-6 results on accuracy, the positives of incorporating task-relevant and irrelevant cues seem to outweigh the downsides making the strategy overall successful. Arguably, negative effects of interaction on performance would be likely observed if, as in Experiments 2-3, participants' initial judgments were not independent (e.g., Pescetelli et al., 2016). The Narcissus-Echo effect can be ascribed to one of several heuristics that allow humans to navigate a complex world with minimal effort, by computing good-enough solutions to otherwise intractable and computationally demanding problems (Gigerenzer & Brighton, 2009).

Confidence alignment

A serendipitous result of Experiment 4 was the observation of confidence alignment among members of the same dyad. Two participants' pre-social information confidence distributions were more similar to each other if the two belonged to the same dyad than to different dyads. In joint decision making tasks, confidence is used to arbitrate disagreement in a near-optimal manner (Bahrami et al., 2010), but a coordination problem arises when group members vary in their confidence trait (Barber & Odean, 2001; Broihanne et al., 2014; Campbell et al., 2004; Huq et al., 1988; Mann, 1998), confidence calibration (Song et al., 2011), and task sensitivity (Bahrami et al.,

2012b). The coordination problem arises from using members' expressed confidence to select the judgment, on a given trial, with the highest probability correct (Bang, Aitchison, et al., 2017; Bang, Summerfield, & Lau, 2017). Confidence alignment and equality bias (Mahmoodi et al., 2015) are simple heuristics that solve with minimal computation this coordination problem through convergence to a social norm (Schelling, 1980). By aligning their confidence distributions, participants can reduce the negative impact of task-irrelevant variance in their confidence judgments (trait confidence (Ais et al., 2016)) while maintaining task-relevant variability (state confidence).

The observation of confidence alignment in our participants was replicated in Experiments 5-6 (Appendix C) and further explored in Experiment 7. Experiment 7 tested alternative explanations for the confidence alignment phenomenon. Confidence alignment was suggested to be due to pre-existing similarities (similarity hypothesis), anchoring effects (anchor hypothesis), or strategic communication (interaction hypothesis). The study was designed to disentangle these explanations. Results showed that confidence distribution were misaligned before interaction took place and rapidly aligned after participants got to know each other's distributions. It was concluded that social interaction caused confidence alignment and that alignment was strategic for the dyad to achieve greater accuracy.

Results of Experiment 7 further showed that alignment benefits dyads, so that if individuals had not aligned their accuracy would have been significantly lower, as implied by our analysis. Furthermore, there was a numerical trend in the direction of greater accuracy benefits for dyads whose participants were more aligned from the start of the experiment. Accuracy benefits due to confidence alignment were likely observed because our participants were matched in terms of average accuracy. Presumably, alignment when dyad members have different task sensitivities (Bahrami et al., 2012b) or different confidence calibrations (Pescetelli et al., 2016) would negatively impact accuracy.

Asking for advice: when my information is not enough

Findings from Experiment 4-6 showed that participants often did not use the advice they received from their partners. Interestingly advice use was not dependent on the initial uncertainty of the participant. A Bayesian observer would show larger updates for smaller prior confidence levels, but this is not what was observed in the data. What was instead seen is that (1) the probability of ignoring advice was roughly equal across levels of prior confidence and (2) advice influence for low prior confidence levels was smaller than predicted by a normative account. An on-going undergraduate project in the lab (Hauperich, Pescetelli, & Yeung, 2017) investigated whether different levels of task difficulty and prior confidence impacted the probability for requesting advice and the impact that the advice had on the initial opinion. Results show that difficulty did not affect the two dependent variables but subjective confidence did. Participants were more prone to ask for advice when their initial confidence was low but they also showed, in the same trials, confidence changes that were smaller than expected by a normative account. A possible interpretation of the phenomenon is that participants attribute the same bias to others, so that when they are unsure about their own decision they expect also others to be uncertain. This can make sense in a world where people share similar perceptual system's capacity and a difficult trial is perceived as difficult by whichever observer.

To some extent, this conclusion is at odds with the findings highlighted in the previous paragraphs suggesting that people tend to wrongly think others' judgments are independent from their own (independence attribution error). When it comes to listening to advice, however, people seem to show a halo effect whereby their own uncertainty is attributed to others (non-independence attribution error). Future work should address these apparently opposite results and find an overarching principle that might underlie both effects.

An alternative and perhaps more parsimonious hypothesis is that different costs are associated with errors due to one's own judgment and errors due to other people's

judgments. Low confidence trials are by definition trials that are closer to the decision boundary where confidence changes are more likely to turn into changes of mind (i.e. changing interval). The regret coming from erring after following someone else's opinion might be greater than the regret coming from erring due to one's own mistake. People have been shown to make decision that are consistent with a regret-minimising strategy (Coricelli et al., 2005). This asymmetry in costs associated with own and others' errors can potentially explain why confidence changes characterised by high uncertainty are found to be smaller than confidence changes characterised by intermediate confidence.

Future directions

Interacting with dynamic models: bridging the two lines

Experiments 4-7 showed how the manipulation of the modalities information is communicated among social partners can affect participants' behaviour. The researcher interested in social phenomena however might feel uncomfortable with the daunting task of abandoning fully controlled paradigms to embrace tasks in which uncontrollable participants interact within controlled environments. This must not necessarily be the case. I am currently trying to bridge the gap existing between the two series of experiments presented in this work using a method developed in the study of visuo-motor imitation, called Human Dynamic Clamp. This method entails a human participant interacting in real-time with a model of human behaviour characterised by several well-defined parameters. The Human Dynamic Clamp paradigm was introduced by Dumas et al. (2014) inspired by the dynamic clamp paradigm used in cellular biology. In its biological counterpart, a living cell membrane potential is measured and fed into a model of the membrane synaptic conductance. The model can compute in real-time the current that has to be inserted back to the living cell. This model-cell pairing allows to study the behaviour of the cell *as if* the cell contained the synaptic conductance modelled with the dynamic clamp. Dumas et al.

(2014) reused this idea by pairing well-established models of hand imitation with a human counterpart, to test whether the dynamics observed in the model-human dyad could reproduce the dynamics observed in human-human dyads, thus validating the model. The model controlled a visual screen output representing individual hand frames that were manipulated by the model in real-time giving the impression of natural movement. The human participant on the other hand controlled a physical handle that recorded hand position and velocity, and fed this information to the model. Manipulation of different model parameters allowed the authors to reproduce a wide range of behaviours observed with real participants, including stable dynamics, in-phase/anti-phase switching and unstable dynamics. Interestingly participants reported perceiving agency and intentionality in their virtual partner.

Inspired by this paradigm, I tried to implement a simpler version of it using the interactive dot and confidence update paradigm. Details of methods and results are reported in the Appendix E. A short summary is provided below. The study, although not bringing any ground-breaking result to the present work, aims to be a proof of concept that such human-model pairings can provide insightful suggestions to theory and new experimental questions.

Methods. 20 participants (age=26.6±4.29; 12 females) were recorded during the poster session of a local conference on a volunteer basis. Participants received chocolate bars for compensation. Participants completed 200 trials of the dot-count task in one of two conditions. In the Self condition, participants rated their confidence in their decision with similar modalities as in Experiment 4-7, and were asked to update their confidence continuously for a 2 seconds after the first response was confirmed. This condition served as baseline to monitor post-decisional processes naturally arising after a decision (Pleskac & Busemeyer, 2010). In the Other condition, participants were also asked to continuously update their initially expressed confidence during 2 second update window, but this time they were paired with a

dynamic model imitating continuous human confidence updates. Participants were told to collaborate with the model if they thought this would help their performance. Crucially, the model provided no information to the participant (thus task-relevant information was matched to the Self condition), as its confidence ratings were sampled from a Gaussian noise distribution - with variable standard deviations across participants - centred on the participant's current confidence in agreement trials and on the participants current confidence minus 50 confidence points in disagreement trials. Participants received feedback on every trial so that they could figure out that the model was uninformative. The experiment aimed at showing whether the addition of a uninformative dynamic social component to the confidence update task was enough to generate some of the non-linear dynamics observed in human-human dyads, like for example confidence escalation.

Results. Full results of the experiment are reported in Appendix E. Confidence increased marginally more in the Other condition than in the Self condition ($p = .06$). Median participants' confidence change during Other condition was plotted along the opinion surface. Participants seemed to resist confidence change when their opinion was dominant (Figure E.4, Appendix E), but this was not the case when they were dominated (Figure 8.1). Figure 8.1 shows that greater confidence updates were observed in trials characterised by weak agreement (point x on the graph) and unbalanced disagreement (point y on the graph). Notice that the same regions of interest are observed in Experiments 4-7. Confidence changes in the Other condition over the 2-second update window were also characterised by greater recursivity as defined in Experiment 4 ($t(19) = 2.15, p = .04, d = 0.42$), indicating that interacting with a dynamic model generated significantly more updates.

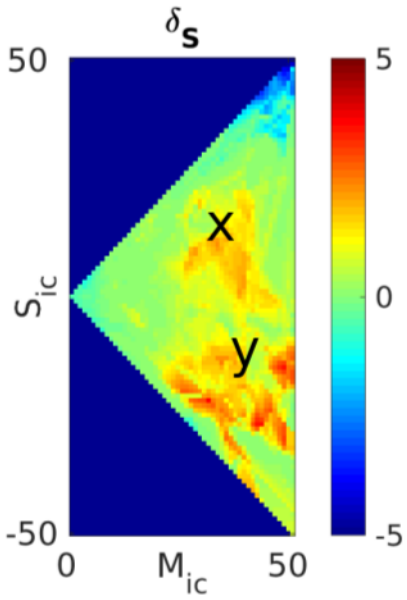


Figure 8.1: Median participants' confidence change in opinion space observed in the Other condition. The plot shows only trials when the participant held the dominated view (model on x-axis, participant on y-axis). Here, agreement led participants to increase their initial confidence and in disagreement to reduce it (or change their mind). Subscript ic indicate "initial confidence".

Analysis of performance was then carried out to test whether being part of a dynamic dyad affected performance when compared to situations characterised by non-social dynamic updates (Other vs. Self). Participants could update their initial opinion in both conditions, allowing me to distinguish performance improvements that could be expected by simple regression to the mean or post-decisional processes from performance improvements that were due to the experimental manipulation. Both first and second order performance measures were considered. First order performance change was measured by the difference between post-update and pre-update decision accuracy (positive values indicating improvements). Second order performance (i.e. confidence calibration) was evaluated by subtracting the area under the type II ROC curve at pre-update time (A^{pre}_{ROC}) from the area at post-update time (A^{post}_{ROC}). Positive values indicate improvements in metacognitive evaluation.

Figure 8.2 shows the average participants' performance change in first and second order task and divided by condition. In the Self condition, neither judgment accuracy or confidence calibration differed from zero (i.e. no change from pre-update phase) ($p > .18, B_{H[0,.05]} < .48$). When participants were paired with the noise model however, judgment accuracy decreased metacognitive accuracy increased. In the Other condition, accuracy was significantly lower ($t(19) = -2.24, p = .03, d = -0.50, B_{H[0,.05]} = 3.62$) and calibration significantly higher ($t(19) = 2.85, p = .01, d = 0.63, B_{H[0,.05]} = 13.81$) than zero, indicating that accuracy decreased but calibration increased. The dissociation between accuracy and calibration is an interesting one given that these measures usually correlate (Fleming & Lau, 2014). The effect is probably due to changes of mind happening in low confidence correct trials. Changes of mind in these trials (point y on the surface plot) had the likely effect of turning into low confidence incorrect trials, thus explaining the decrease in accuracy and increase in calibration.

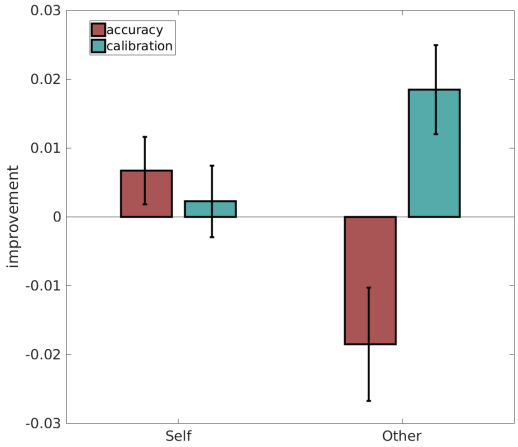


Figure 8.2: Dissociation between changes in first and second order performance, observed when participants were paired with the noise dynamic model.

Relevance. This pilot experiment showed that even a simple algorithm modelling the behaviour of an uninformative interactive agent can affect the human counterpart producing patterns of confidence changes that are characteristic of human-human

dyads (Experiments 4-7) - namely larger confidence shifts when participants' independent judgment started from middle portions of the confidence scale (e.g., confidence levels of 20-30). Contrary to what a Bayesian norm would prescribe, but similarly to what already observed in Expeirments 4-7, confidence changes were smaller when the participant's independent judgment was very unsure (e.g., confidence levels of 0-20). Interesting effects can be produced even in these simple situations, as shown by the opposite effects on accuracy and confidence calibration (Fleming & Lau, 2014). A simple manipulation, namely being coupled with a Gaussian noise model mirroring one's own decision, dissociates changes observed in the two measures.

Being aware of fundamental limitations of this simple design and that the findings were possibly due to the specific contingencies of the model used, I do not claim that the experiment introduced a radical new result. The experiment however wants to be proof of concept that it is possible to overcome the trade-off between control and generalisability when studying social interactions. It is further argued that the trade-off itself is not a real limitation for researchers but actually a space of viable investigations offering unexplored research avenues.

Rethinking the social landscape. A useful framework for future research in social interaction is to picture it as a continuum spanning several dimensions (Figure 8.3). On the one hand, dyads and groups can span the whole range from being composed of all humans (e.g. naturalistic interaction, observational studies, etc.) to all machines (e.g. simulation studies, agent-based models, etc.). On the other hand, social exchange can vary along the dimension of the directionality of the information flow. In one-directional paradigms information is transferred from a sender to a receiver who then decides how to aggregate it with personal believes (e.g. judge-adviser systems). In bi-directional studies on the contrary information is bidirectionally shared among two or more individuals (e.g. Experiment 4-8, interactive conditions). A further dimension can be added characterising the dynamics of the

social exchange which can range from being a one-step process where initial estimates are aggregated only once (stationary opinion aggregation studies (e.g., Migdal et al., 2012), advice-revise paradigms, judge-adviser systems, etc.) to a recursive process where information is actively transformed over time until dyads/groups reach an equilibrium state (Experiments 4-8).

Given the full spectrum of possible designs that the researcher interested in social questions can explore, it seems clear that conclusions can be reached only from the coordinated convergence of results from different paradigms. Furthermore, although here we investigated these questions using a belief update paradigm, it should be noted that similar principles could in principle be applied to different social research questions. Research on emotional contagion, collective dynamics, motor imitation, theory of mind and moral decision-making could all benefit from adopting this wider perspective.

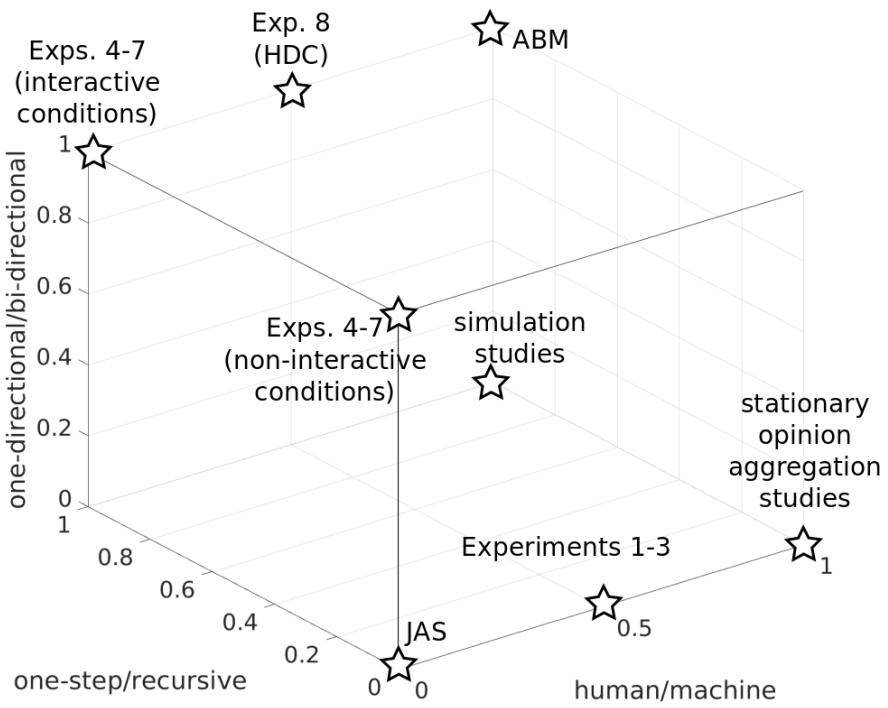


Figure 8.3: The figure shows the space of possible investigation in cognitive social sciences and a non-exhaustive list of paradigms in their relative locations.

Larger, bigger, greater

One of the most pressing questions originating from the findings of the present work is how the results might scale up when considering larger group sizes and how they expand to different problem spaces. In previous chapters, the idea that preference for bias-sharing and agreeing advisers can generate echo-chambers in social networks with no access to feedback has been put forward. Agent-based models, Human Dynamic Clamp methods and social network analysis offer three complementary levels of investigation where this hypothesis can be put to test. Existing results from Navajas et al. (2017) suggest that hierarchical deliberation and integration of opinions might not suffer from the same biases that affect interactions at the local scale. The authors asked a live crowd to respond to general knowledge questions individually and then to discuss their views in small groups to reach a consensus. Consensus estimates were less biased and more diverse compared to a uniform aggregation of the independent opinions (cf. Lorenz et al., 2011). Consequently, averaging post-deliberation estimates was a better strategy than aggregating individuals views, indicating benefits of argumentative aggregation strategies (Mercier & Sperber, 2011). Hierarchical integration in tasks where the popular opinion is often incorrect has still to be assessed (Koriat, 2012b; Krause et al., 2010; Migdal et al., 2012).

At the same time, these effects should be investigated in different problem domains. Future studies should address whether similar biases are found when participants are engaging with problems involving more than two options, multiple decision dimensions, value-based decisions and non-perceptual tasks. There are reasons to believe that similar phenomena might be found. Phenomena of social influence and imitative behaviour are observed in several non-perceptual domains and multidimensional problem spaces (Mason & Watts, 2012). Furthermore, earlier studies on group polarisation (Myers & Lamm, 1976), attitude change (Festinger, 1954; Griffitt, 1969; Simons, Berkowitz, & Moyer, 1970), hidden profile paradigm (Stasser & Titus, 2003)

and *groupthink* (Janis, 1972) seem all to suggest that preference for agreeing opinions is a widely observed phenomenon. These questions are theoretically interesting and allow us to define the conditions under which these cognitive biases are more likely to be consequential. In these cases, the definition of agreement and inter-subject confidence co-variation is harder to define compared to simpler cases of perceptual binary judgment, but recent work from Gershman, Pouncy, and Gweon (2017) provides an elegant general framework describing how humans can infer structures of similarity among observed others. The authors use a probabilistic approach and simple assumptions about groups to model the behaviour of a rational agent inferring the group membership of observed decision makers from their choice patterns. The model infers the posterior distribution over latent group assignments, which is shown to produce similar patterns to what observed in real participants. This work offers a formal solution, grounded in computational theories of structure learning (Gershman & Niv, 2010), that can be flexibly adapted to several domains.

Alternative hypotheses for the interactive effect

Results of the second line of experiments showed robust effects that real-time interaction has on opinion change. The interpretation that was offered to explain the difference with non-interactive conditions was that interaction produced different dynamics between participants due to the recursive nature of the social exchange (the dynamics hypothesis).

Results from the mixed-effects model however might be indicative of an alternative explanation. The regression analysis showed that, in the interactive condition, participants' confidence updating behaviour was modulated by their partner's updates. According to this alternative interpretation (the heuristic hypothesis) making participants aware of their partner's updates in the interactive condition might have offered them an indirect indicator of the partner's true underlying confidence (as discussed above) over and beyond the presence of recursive processes. In other words,

interactive and non-interactive conditions did not differ only in the dynamics of the information exchange (one was recursive the other was not), but also because interactive conditions simply offered participants *more* information about their partners.

Future experiments should address this issue by introducing a non-interactive condition in which participants are informed (through a static prompt) also about their partner final confidence (*Non-Interactive+*). This condition introduces the same knowledge about the partner's update magnitude that the heuristic hypothesis supposes is producing differences between interaction and non-interaction. A heuristic hypothesis would predict that such condition should be equal to the interactive one as in both participants are informed about their partner's original opinion and update magnitude. A dynamic hypothesis on the contrary would predict that the *Non-Interactive+* condition should be equal to the non-interactive condition because they both lack recursive communication.

Using tailored dynamic tutors

One of the most exciting avenues of investigation regards the applicability of adaptive dynamic models of human interaction to improve people's metacognitive skills, calibration and judgment accuracy. It has long been known that by interacting with others we can improve our judgments and reduce our biases (Surowiecki, 2004). However it is often difficult to control for the precise environmental conditions where good interaction can take place. Results have been often found in opposite directions to the point that group decision-making has been said to deserve the Dr. Jekyll and Mr. Hyde Award for psychology (Hertwig, 2012). The idea of a Human Dynamic Clamp methodology paired with the knowledge of the mechanisms behind opinion change in the presence of others allow to create personalised virtual tutors tailored to the behavioural profiles of people with the aim of improving their judgments by means of interaction. For example, the alignment effect could be used to appropriately modify uncalibrated people's confidence distributions. An ongoing undergraduate project

(Van den Broeck et al., 2017, in preparation) is exploring this possibility by pairing participants with virtual tutors who show complementary confidence distributions so that over-confident participants are matched with under-confident tutors and vice versa. The expectation is that the tendency to align with the tutor will improve participants' calibration profile. Questions remain, however, whether alignment affects people's true underlying metacognitive ability or simply their mapping function from internal states to response scale. Results of Experiment 7 showed that once departed from each other partners' confidence alignment decreased, suggesting that perhaps the effects are transitory.

Similarly, virtual tutors could be tailored to the participants in terms of opinion distance. We showed that advice is often declined or ignored. Previous studies have shown that others' opinions must be in a certain range of distance from the participant's to be taken into account (Moussaïd, Kämmer, Analytis, & Neth, 2013; Schultze, Rakotoarisoa, & Schulz-Hardt, 2015). Tailored dynamic tutors can then be used to make people's opinions more malleable by presenting humans with advice that is more likely to affect them and dynamically adjusting this over time.

Non-linear dynamics among partners in the escalation of confidence can thus be used to arbitrarily shift people opinions depending on expected probabilities computed by appropriately chosen models. This route offers the opportunity to enhance human judgments with the knowledge of normative and statistical models with access to different information sets. The increasingly widespread use of social-bot on online platforms (Guilbeault, 2016) can thus be virtuously used to improve individual judgment. Traditionally (Minsky, 1961), effort in machine learning and artificial intelligence has focused on surpassing human performance (Mnih et al., 2015; D. Silver et al., 2016). The use of tailored dynamic tutors would instead place itself along the different tradition of human-machine symbiosis (Licklider, 1960; Roy, 2004), exploiting uniquely-human and uniquely-machine capabilities. People's perception of

human and artificial advisers is however different even when controlling for information and accuracy (Boorman et al., 2013; Dietvorst et al., 2015). Any application of tailored dynamic tutors should thus take these effects into account. In this way, the work presented in this thesis presents new avenues for improving individual decisions with observations from multiple observers (virtual and/or human) and how judgments should be shared and integrated across agents.

Conclusion

This Chapter concludes my work for this doctorate degree, integrating the findings of my experiments conducted with one core question in mind: How do internal metacognitive states inform and guide our behaviour in the exploration of our social world? We knew that confidence is intertwined with our daily decisions and the way we learn from them. Given our limited cognition, metacognitive signals offer useful measures that allow us to navigate an uncertain world, telling us when to seek for more information and when uncertain outcomes are more likely to happen. We discovered here that confidence is also entangled with our social world, in which other cognitive agents are also embarking into an inference process about unknown external states. Our confidence - whether stated or perceived - gives to others some knowledge about ourselves and gives to ourselves knowledge about our mind. This work demonstrates how, thanks to confidence, we can collectively succeed and warned about how, thanks to confidence, we can collectively fail.

Many studies have been presented but many more could and should be done if we want to exhaustively answer the questions posed here. It is my hope that this work gave the reader a brief but insightful peek into a complex and fascinating landscape that we are only starting to understand.

References

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transaction on the Web*, *6*(2), 9. Retrieved from `https://hal.archives-ouvertes.fr/hal-00718085`

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016, jan). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377–386. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0010027715300846` doi: 10.1016/j.cognition.2015.10.006

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Computational Biology*, *11*(10), 1.

Ariely, D. (2008). *Predictably irrational: the hidden forces that shape our decisions* (HarperColl ed.).

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic.

Asch, S. E. (1956). Studies of Independence and Conformity: a Minority of One Against a Unanimous Majority. *Psychological Monographs: General and Applied*, *70*(9, whole no. 416), 1–70.

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological review1*, *67*, 1–15.

Auvray, M., Lenay, C., & Stewart, J. (2009, apr). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology*, *27*(1), 32–47. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0732118X07000748` doi: 10.1016/j.newideapsych.2007.12.002

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006, may). Neural correlations, population coding and computation. *Nature reviews. Neuroscience*, *7*(5), 358–66. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/16760916` doi: 10.1038/nrn1888

Bach, D. R., & Dolan, R. J. (2012, aug). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature reviews. Neuroscience*, *13*(8), 572–86. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/22781958` doi: 10.1038/nrn3289

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2013, apr). Collective enumeration. *Journal of experimental psychology. Human perception and perfor-*

mance, *39*(2), 338–47. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3607463{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1037/a0029717

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012a, feb). Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of experimental psychology. Human perception and performance*, *38*(1), 3–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3268462{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1037/a0025708

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012b, apr). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1350–1365. Retrieved from http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0420 doi: 10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010, aug). Optimally interacting minds. *Science (New York, N.Y.)*, *329*(5995), 1081–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20798320 doi: 10.1126/science.1185718

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013, oct). Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *33*(42), 16657–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24133268 doi: 10.1523/JNEUROSCI.0786-13.2013

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Bang, D., Aitchison, L., Moran, R., Castanon, S. H., Rafiee, B., Mahmoodi, A., ... Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(0117), 1–7. doi: 10.1038/s41562-017-0117

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., ... Bahrami, B. (2014, may). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, *26*, 13–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24650632http://linkinghub.elsevier.com/retrieve/pii/S1053810014000324 doi: 10.1016/j.concog.2014.02.002

Bang, D., Summerfield, C., & Lau, J. Y. (2017). *On confidence in individual and group decision-making* (DPhil thesis, University of Oxford). Retrieved from https://ora.ox.ac.uk/objects/uuid:e86852b9-d167-44bb-9e0f-add2183bf1f1

Barber, B. M., & Odean, T. (2001). Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment. *The Quarterly Journal of Economics*, *116*(1), 261–292.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008, nov). Associative learning of social value. *Nature*, *456*(7219), 245–9.

Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2605577{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1038/nature07538

Bessi, A. (2016, dec). Personality traits and echo chambers on facebook. *Computers in Human Behavior*, *65*, 319–324. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0747563216305817 doi: 10.1016/j.chb.2016.08.016

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., ... Quattrociocchi, W. (2016, aug). Users Polarization on Facebook and Youtube. *PLOS ONE*, *11*(8), e0159641. Retrieved from http://dx.plos.org/10.1371/journal.pone.0159641 doi: 10.1371/journal.pone.0159641

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*, 700–765. doi: 10.1037/0033-295X.113.4.700

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The Impact of Evidence Reliability on Sensitivity and Bias in Decision Confidence. *Journal of Experimental Psychology: Human Perception and Performance*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28383959{%}0Ahttp://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000404 doi: 10.1037/xhp0000404

Boldt, A., & Yeung, N. (2015, feb). Shared Neural Markers of Decision Confidence and Error Detection. *Journal of Neuroscience*, *35*(8), 3478–3484. Retrieved from http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0797-14.2015 doi: 10.1523/JNEUROSCI.0797-14.2015

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press. doi: 0195131584

Bonaccio, S., & Dalal, R. S. (2006, nov). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597806000719 doi: 10.1016/j.obhdp.2006.07.001

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013, dec). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–71. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3878380{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1016/j.neuron.2013.10.024

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.108.3.624 doi: 10.1037/0033-295X.108.3.624

Brehmer, B., & Hagafors, R. (1986). Use of experts in complex decision making: A paradigm for the study of staff work. *Organizational Behavior and Human Decision Processes*, *38*(2), 181–195. doi: 10.1016/0749-5978(86)90015-4

Brennan, A. A., & Enns, J. T. (2015). When two heads are better than one: Interactive versus independent benefits of collaborative cognition. *Psychonomic*

Bulletin & Review, *22*(4), 1076–1082.

Broihanne, M., Merli, M., & Roger, P. (2014, jun). Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Finance Research Letters*, *11*(2), 64–73. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1544612313000627 doi: 10.1016/j.frl.2013.11.002

Buchan, N. R., Croson, R. T., & Solnick, S. (2008, dec). Trust and gender: An examination of behavior and beliefs in the Investment Game. *Journal of Economic Behavior & Organization*, *68*(3-4), 466–476. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S016726810800139X doi: 10.1016/j.jebo.2007.10.006

Budescu, D. V., & Rantilla, A. K. (2000, jun). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0001691800000378 doi: 10.1016/S0001-6918(00)00037-8

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004, oct). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, *17*(4), 297–311. Retrieved from http://doi.wiley.com/10.1002/bdm.475 doi: 10.1002/bdm.475

Carandini, M., & Heeger, D. J. (2011, nov). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. Retrieved from http://www.nature.com/doifinder/10.1038/nrn3136 doi: 10.1038/nrn3136

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013, jun). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23380166 doi: 10.1016/j.neuroimage.2013.01.054

Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Herce Castañón, S., & Summerfield, C. (2014, mar). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24656259 doi: 10.1016/j.neuron.2014.01.020

Clark, C. C., & Lawless, H. T. (1994). Limiting response alternatives in time-intensity scaling: an examination of the halo-dumping effect. *Chemical Senses*, *19*(6), 583–594. Retrieved from https://ezproxy-prd.bodleian.ox.ac.uk:5876/chemse/article/19/6/583/article doi: 10.1093/chemse/19.6.583

Clifford, C. W. G., Arabzadeh, E., & Harris, J. a. (2008, mar). Getting technical about awareness. *Trends in cognitive sciences*, *12*(2), 54–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18178511 doi: 10.1016/j.tics.2007.11.009

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature neuroscience*, *8*, 1255 – 1262. doi: 10.1038/nn1514

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine learning*, *20*, 273–297.

Couzin, I. D. (2009, jan). Collective cognition in animal groups. *Trends in cognitive sciences*, *13*(1), 36–43. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19058992 doi: 10.1016/j.tics.2008.10.002

D'Ausilio, A., Novembre, G., Fadiga, L., & Keller, P. E. (2015, mar). What can

music tell us about social interaction? *Trends in Cognitive Sciences*, *19*(3), 111–114. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661315000066 doi: 10.1016/j.tics.2015.01.005

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012, may). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1379–90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22492754 doi: 10.1098/rstb.2012.0002

Daw, N. D., Niv, Y., & Dayan, P. (2005, dec). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16286932 doi: 10.1038/nn1560

De Martino, B., Camerer, C. F., & Adolphs, R. (2010, feb). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3788–92. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2840433{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1073/pnas.0910230107

De Martino, B., O'Doherty, J. P., Ray, D., Bossaerts, P., & Camerer, C. (2013, sep). In the Mind of the Market: Theory of Mind Biases Value Computation during Financial Bubbles. *Neuron*, *79*(6), 1222–1231. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0896627313005680 doi: 10.1016/j.neuron.2013.07.003

de Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des vois*. Paris: de l'Imprimerie Royale.

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84. Retrieved from http://dx.doi.org/10.1016/j.conb.2013.12.005 doi: 10.1016/j.conb.2013.12.005

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016, dec). Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, *6*(1), 37825. Retrieved from http://www.nature.com/articles/srep37825 doi: 10.1038/srep37825

Denrell, J. (2005). Why Most People Disapprove of Me: Experience Sampling in Impression Formation. *Psychological Review*, *112*(4), 951–978. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.112.4.951 doi: 10.1037/0033-295X.112.4.951

Dezecache, G., Conty, L., Chadwick, M., Philip, L., Soussignan, R., Sperber, D., & Grèzes, J. (2013, jan). Evidence for unintentional emotional contagion beyond dyads. *PloS one*, *8*(6), e67371. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3696100{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1371/journal.pone.0067371

Dienes, Z., & Mclatchie, N. (2017, mar). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 1–12. Re-

trieved from http://link.springer.com/10.3758/s13423-017-1266-z doi: 10.3758/s13423-017-1266-z

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033 doi: 10.1037/xge0000033

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... Johannesson, M. (2015, nov). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. Retrieved from http://www.pnas.org/content/112/50/15343.abstract.html?etoc doi: 10.1073/pnas.1516179112

Dumas, G., de Guzman, G. C., Tognoli, E., & Kelso, J. a. S. (2014, aug). The human dynamic clamp as a paradigm for social interaction. *Proceedings of the National Academy of Sciences*, 3726–3734. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.1407486111 doi: 10.1073/pnas.1407486111

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010, jan). Inter-brain synchronization during social interaction. *PloS one*, *5*(8), e12166. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923151{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1371/journal.pone.0012166

Dunbar, R. (2003, oct). The Social Brain : Mind, Language, and Society in Evolutionary Perspective. *Annual Review of Anthropology*, *32*(1), 163–181. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.anthro.32.061002.093158 doi: 10.1146/annurev.anthro.32.061002.093158

Dunbar, R., & Shultz, S. (2007, sep). Evolution in the social brain. *Science (New York, N.Y.)*, *317*(5843), 1344–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17823343 doi: 10.1126/science.1145463

Edelson, M., Sharot, T., Dolan, R. J., & Dudai, Y. (2011, jul). Following the Crowd: Brain Substrates of Long-Term Memory Conformity. *Science*, *333*(6038), 108–111. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1203557 doi: 10.1126/science.1203557

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200. doi: 10.1177/0146167205282152

Epstein, J. M. (2013). *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*. Princeton and Oxford: Princeton University Press.

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

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.105.3.482 doi: 10.1037/0033-295X.105.3.482

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 1954.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. Retrieved from http://content.apa.org/journals/amp/34/10/906 doi: 10.1037/0003-066X.34.10.906

Fleming, S. M. (2016). Changing our minds about changes of mind. *eLife*, *5*. doi: 10.7554/eLife.14790

Fleming, S. M., & Dolan, R. J. (2012, apr). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. Retrieved from http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0417 doi: 10.1098/rstb.2011.0417

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012, apr). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286. Retrieved from http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2012.0021 doi: 10.1098/rstb.2012.0021

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012, may). Prefrontal Contributions to Metacognition in Perceptual Decision Making. *Journal of Neuroscience*, *32*(18), 6117–6125. Retrieved from http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.6489-11.2012 doi: 10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., & Lau, H. C. (2014, jul). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. Retrieved from http://www.frontiersin.org/Human{_}Neuroscience/10.3389/fnhum.2014.00443/abstract doi: 10.3389/fnhum.2014.00443

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2014). Action-Specific Disruption of Perceptual Confidence. *Psychological science*. doi: 10.1177/0956797614557697

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010, sep). Relating introspective accuracy to individual differences in brain structure. *Science (New York, N.Y.)*, *329*(5998), 1541–3. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3173849{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1126/science.1191883

Friston, K., & Frith, C. (2015, nov). A Duet for one. *Consciousness and Cognition*, *36*, 390–405. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S105381001400230X doi: 10.1016/j.concog.2014.12.003

Frith, C. D. (1999, nov). Interacting Minds–A Biological Basis. *Science*, *286*(5445), 1692–1695. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.286.5445.1692 doi: 10.1126/science.286.5445.1692

Frith, C. D. (2007, apr). The social brain? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1480), 671–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1919402{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1098/rstb.2006.2003

Frith, C. D. (2012, aug). The role of metacognition in human social interactions.

Philosophical transactions of the Royal Society of London. Series B, Biological sciences, *367*(1599), 2213–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22734064 doi: 10.1098/rstb.2012.0123

Frith, U., & Frith, C. D. (2003, mar). Development and neurophysiology of mentalizing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *358*(1431), 459–73. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1693139{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1098/rstb.2002.1218

Froese, T., Iizuka, H., & Ikegami, T. (2014, jan). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports*, *4*. Retrieved from http://www.nature.com/articles/srep03672 doi: 10.1038/srep03672

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012, jul). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological science*, *online Jul*, 1–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22810169 doi: 10.1177/0956797612436816

Gallagher, H. L., & Frith, C. D. (2003, feb). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, *7*(2), 77–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12584026

Galton, F. (1907, mar). Vox Populi. *Nature*, *75*(1949), 450–451. Retrieved from http://www.nature.com/doifinder/10.1038/075450a0 doi: 10.1038/075450a0

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003, dec). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. Retrieved from http://www.springerlink.com/index/10.3758/BF03196546 doi: 10.3758/BF03196546

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current opinion in neurobiology*, *20*(2), 251–256. doi: 10.1016/j.conb.2010.02.008

Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017, mar). Learning the Structure of Social Influence. *Cognitive Science*. Retrieved from http://doi.wiley.com/10.1111/cogs.12480 doi: 10.1111/cogs.12480

Gigerenzer, G. (2008). Why Heuristics Work? *Perspectives on Psychological Science*, *3*(1), 20–29.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in cognitive science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.98.4.506 doi: 10.1037/0033-295X.98.4.506

Gigerenzer, G., & Todd, P. A. R. G. (1999). *Simple Heuristics that Make us Smart*. Oxford: Oxford University Press.

Graefe, A., & Armstrong, J. S. (2011, jan). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting*, *27*(1), 183–195. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169207010000889 doi: 10.1016/j.ijforecast.2010.05.004

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley and Sons.

Griffitt, W. B. (1969). Personality similarity and self-concept as determinants of interpersonal attraction. *Journal of Social Psychology*, *78*(1), 137. Retrieved from https://ezproxy-prd.bodleian.ox.ac.uk:7316/docview/1290677422?accountid=13042

Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016, mar). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, *5*. Retrieved from http://elifesciences.org/lookup/doi/10.7554/eLife.13388 doi: 10.7554/eLife.13388

Guilbeault, D. (2016). Growing Bot Security : An Ecological View of Bot Agency. *International Journal of Communication*, *10*(June), 5003–5021.

Gürçay, B., Mellers, B. A., & Baron, J. (2015, jul). The Power of Social Influence on Estimation Accuracy. *Journal of Behavioral Decision Making*, *28*(3), 250–261. Retrieved from http://doi.wiley.com/10.1002/bdm.1843 doi: 10.1002/bdm.1843

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–97.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016, aug). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *40*(6), 1496–1533. Retrieved from http://doi.wiley.com/10.1111/cogs.12276 doi: 10.1111/cogs.12276

Harvey, N., & Fischer, I. (1997, may). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597897926972 doi: 10.1006/obhd.1997.2697

Harvey, N., & Harries, C. (2004, jul). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting*, *20*(3), 391–409. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169207003001109 doi: 10.1016/j.ijforecast.2003.09.012

Hastie, R., & Kameda, T. (2005, apr). The robust beauty of majority rules in group decisions. *Psychological review*, *112*(2), 494–508. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15783295 doi: 10.1037/0033-295X.112.2.494

Hauperich, A.-K., Pescetelli, N., & Yeung, N. (2017). *Confidence Drives Post-decisional Search of Information from Social Sources.*

Hayek, F. A. (1945). The Use of Knowledge in Society. *The American Economic Review*, *35*(4), 519–530.

Haynes, J.-D., & Rees, G. (2005, may). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature neuroscience*, *8*(5), 686–

91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15852013 doi: 10.1038/nn1445

Heath, C., & Gonzalez, R. (1995). Interaction with Others Increases Decision Confidence but Not Decision Quality: Evidence against Information Collection Views of Interactive Decision Making. *Organizational Behavior and Human Decision Processes*, *61*(3), 305–326. doi: 10.1006/obhd.1995.1024

Heider, F., & Simmel, M. (1944, apr). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243. Retrieved from http://www.jstor.org/stable/1416950?origin=crossref doi: 10.2307/1416950

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/h0074579 doi: 10.1037/h0074579

Hertwig, R. (2012, apr). Tapping into the Wisdom of the Crowd–with Confidence. *Science*, *336*(6079), 303–304. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1221403 doi: 10.1126/science.1221403

Hertwig, R., & Gigerenzer, G. (1999). The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors. *Journal of Behavioral Decision Making*, *12*, 275–205.

Hertz, U., Romand-Monnier, M., Kyriakopoulou, K., & Bahrami, B. (2016). Social influence protects collective decision making from equality bias. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 164–172. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000145 doi: 10.1037/xhp0000145

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The Emerging Conceptualization of Groups as Information Processors. *Psychological Bulletin*, *121*(1), 43–64.

Hong, L., & Page, S. E. (2004, nov). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–9. Retrieved from http://www.pnas.org/content/101/46/16385.abstract doi: 10.1073/pnas.0403723101

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988, nov). Probabilistic judgements in deluded and non-deluded subjects. *The Quarterly Journal of Experimental Psychology Section A*, *40*(4), 801–812. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/14640748808402300 doi: 10.1080/14640748808402300

Jamieson, K. H., & Cappella, J. N. (2008). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford: Oxford University Press. Retrieved from https://books.google.co.uk/books?id=139Oa4MOsAgC{&}redir{_}esc=y

Janis, I. L. (1972). *Victims of groupthink*. Boston: Houghton Mifflin.

Jasny, L., Waggle, J., & Fisher, D. R. (2015, may). An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change*, *5*(8), 782–786. Retrieved from http://www.nature.com/doifinder/10.1038/nclimate2666 doi: 10.1038/nclimate2666

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2 March), 263–291.

Kamitani, Y., & Tong, F. (2005, may). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, *8*(5), 679–85. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808230{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1038/nn1444

Kao, A. B., & Couzin, I. D. (2014, apr). Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133305–20133305. Retrieved from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2013.3305 doi: 10.1098/rspb.2013.3305

Kao, A. B., Miller, N., Torney, C., Hartnett, A., & Couzin, I. D. (2014, aug). Collective Learning and Optimal Consensus Decisions in Social Animal Groups. *PLoS Computational Biology*, *10*(8), e1003762. Retrieved from http://dx.plos.org/10.1371/journal.pcbi.1003762 doi: 10.1371/journal.pcbi.1003762

Kepecs, A., & Mainen, Z. F. (2012, apr). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322–1337. Retrieved from http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2012.0037 doi: 10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. a., & Mainen, Z. F. (2008, sep). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18690210 doi: 10.1038/nature07200

Kerr, N. L., & Tindale, R. S. (2004, feb). Group Performance and Decision Making. *Annual Review of Psychology*, *55*(1), 623–655. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.55.090902.142009http://www.annualreviews.org/doi/10.1146/annurev.psych.55.090902.142009 doi: 10.1146/annurev.psych.55.090902.142009

Khalighinejad, N., & Haggard, P. (2016). Extending experiences of voluntary action by association. *PNAS*, *113*(31), 8867–8872.

Kiani, R., Corthell, L., & Shadlen, M. (2014, dec). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0896627314010964 doi: 10.1016/j.neuron.2014.12.015

Kiani, R., & Shadlen, M. N. (2009, may). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science (New York, N.Y.)*, *324*(5928), 759–64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19423820http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2738936 doi: 10.1126/science.1169405

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007, sep). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, *8*(3), 159–66. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2649419{&}tool=pmcentrez{&}rendertype=abstract doi: 10

.1007/s10339-007-0170-2

King, J.-R., & Dehaene, S. (2014, may). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1641), 20130204. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24639577 doi: 10.1098/rstb.2013.0204

King, J.-R., Pescetelli, N., & Dehaene, S. (2016, dec). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0896627316308017 doi: 10.1016/j.neuron.2016.10.051

Knill, D. C., & Pouget, A. (2004). The Bayesian brain : the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. doi: 10.1016/j.tins.2004.10.007

Ko, Y., & Lau, H. (2012, may). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1401–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22492756 doi: 10.1098/rstb.2011.0380

Körding, K. P., & Wolpert, D. M. (2004, jan). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247. Retrieved from http://www.nature.com/doifinder/10.1038/nature02169 doi: 10.1038/nature02169

Koriat, A. (2012a, jan). The self-consistency model of subjective confidence. *Psychological review*, *119*(1), 80–113. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22022833 doi: 10.1037/a0025648

Koriat, A. (2012b, apr). When are two heads better than one and why? *Science (New York, N.Y.)*, *336*(6079), 360–2. Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.1216549http://www.ncbi.nlm.nih.gov/pubmed/22517862 doi: 10.1126/science.1216549

Krause, J., Ruxton, G. D., & Krause, S. (2010, jan). Swarm intelligence in animals and humans. *Trends in Ecology & Evolution*, *25*(1), 28–34. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169534709002298 doi: 10.1016/j.tree.2009.06.016

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), 1121–1134.

Kvam, P. D., & Pleskac, T. J. (2016). Strength and weight: The determinants of choice and confidence. *Cognition*, *152*, 170–180. Retrieved from http://dx.doi.org/10.1016/j.cognition.2016.04.008 doi: 10.1016/j.cognition.2016.04.008

Larrick, R. P., & Soll, J. B. (2006, jan). Intuitions About Combining Opinions: Misappreciation of the Averaging Principle. *Management Science*, *52*(1), 111–127. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1050.0459 doi: 10.1287/mnsc.1050.0459

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and So-*

cial Psychology, *37*(6), 822–832. Retrieved from http://content.apa.org/journals/psp/37/6/822 doi: 10.1037/0022-3514.37.6.822

Lau, H. C. (2008, jan). A higher order Bayesian decision theory of consciousness. *Progress in brain research*, *168*, 35–48. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18166384

Le Bon, G. (1895). *La psychologie des foules* (Félix Alca ed.). Paris. Retrieved from http://socserv.mcmaster.ca/econ/ugcm/3ll3/lebon/Crowds.pdf

Le Pelley, M. E. (2012). Metacognitive monkeys or associative animals? Simple reinforcement learning explains uncertainty in nonhuman animals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 686–708. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0026478 doi: 10.1037/a0026478

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*, 4–11.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011, may). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9020–5. Retrieved from http://www.pnas.org/content/108/22/9020.abstract doi: 10.1073/pnas.1008636108

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. doi: 10.1038/nn1790

Mackay, C. (1841). *Extraordinary Popular Delusions and the Madness of Crowds* (Wordsworth ed.). Ware, UK: Wordsworth Edition Limited.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: a User's Guide* (second Edi ed.). London: Lawrence Erlbaum Associates.

Mahmoodi, A., Bang, D., Ahmadabadi, M. N., & Bahrami, B. (2013, dec). Learning to Make Collective Decisions: The Impact of Confidence Escalation. *PLoS ONE*, *8*(12), e81195. Retrieved from http://dx.plos.org/10.1371/journal.pone.0081195 doi: 10.1371/journal.pone.0081195

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., ... Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, 201421692. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1421692112 doi: 10.1073/pnas.1421692112

Maniscalco, B., & Lau, H. (2012, mar). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22071269 doi: 10.1016/j.concog.2011.09.021

Mann, L. (1998, oct). Cross-cultural Differences in Self-reported Decision-making Style and Confidence. *International Journal of Psychology*, *33*(5), 325–335. Retrieved from http://doi.wiley.com/10.1080/002075998400213 doi: 10.1080/002075998400213

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014, aug). The wisdom of select crowds. *Journal of personality and social psychology*, *107*(2), 276–99. doi:

10.1037/a0036677

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *PNAS*, *109*(3), 764–9. doi: 10.1073/pnas.1110069108

Massen, J. J. M., Šlipogor, V., & Gallup, A. C. (2016). An Observational Investigation of Behavioral Contagion in Common Marmosets (Callithrix jacchus): Indications for Contagious Scent-Marking. *Frontiers in Psychology*, *7*, 1190. doi: 10.3389/fpsyg.2016.01190

Mattout, J. (2012). Brain-Computer Interfaces: A Neuroscience Paradigm of Social Interaction? A Matter of Perspective. *Frontiers in Human Neuroscience*, *6*(114). Retrieved from http://journal.frontiersin.org/article/10.3389/fnhum.2012.00114/abstract doi: 10.3389/fnhum.2012.00114

McClelland, J. L., & Rogers, T. T. (2003, apr). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. Retrieved from http://www.nature.com/doifinder/10.1038/nrn1076 doi: 10.1038/nrn1076

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111. doi: 10.1017/S0140525X10000968

Metcalfe, J., & Greene, M. J. (2007). Metacognition of Agency. *Journal of Experimental Psychology: General2*, *136*(2), 184–199. doi: 10.1037/0096-3445.136.2.184

Meyniel, F., Sigman, M., & Mainen, Z. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0896627315008284 doi: 10.1016/j.neuron.2015.09.039

Migdal, P., Raczaszek-Leonardi, J., Denkiewicz, M., & Plewczynski, D. (2012, dec). Information-sharing and aggregation models for interacting minds. *Journal of Mathematical Psychology*, *56*(6), 417–426. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0022249613000096 doi: 10.1016/j.jmp.2013.01.002

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE*, *January*, 8–30.

Mitchell, A., Gottfried, J., Barthel, M., & Shearer, E. (2016). *The Modern News Consumer* (Tech. Rep.). Pew Research Center. Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/13/2016/07/08140120/PJ{_}2016.07.07{_}Modern-News-Consumer{_}FINAL.pdf

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. doi: 10.1038/nature14236

Moore, J. W., Dickinson, A., & Fletcher, P. C. (2011, sep). Sense of agency, associative learning, and schizotypy. *Consciousness and Cognition*, *20*(3), 792–800. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1053810011000031 doi: 10.1016/j.concog.2011.01.002

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence : Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. Retrieved from http://dx.doi.org/

10.1016/j.cogpsych.2015.01.002 doi: 10.1016/j.cogpsych.2015.01.002

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory Integration in Macaque Visual Cortex Depends on Cue Reliability. *Neuron*, *59*, 662–673. doi: 10.1016/j.neuron.2008.06.024

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135.

Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015, may). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences*, *112*(18), 5631–5636. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1421883112 doi: 10.1073/pnas.1421883112

Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013, nov). Social Influence and the Collective Dynamics of Opinion Formation. *PLoS ONE*, *8*(11), e78433. Retrieved from http://dx.plos.org/10.1371/journal.pone.0078433 doi: 10.1371/journal.pone.0078433

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015, dec). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, *4:e11946*, 3478–3484. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26687008http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4749550http://elifesciences.org/lookup/doi/10.7554/eLife.11946 doi: 10.7554/eLife.11946

Murty, V. P., DuBrow, S., & Davachi, L. (2015, apr). The Simple Act of Choosing Influences Declarative Memory. *Journal of Neuroscience*, *35*(16), 6255–6264. Retrieved from http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.4181-14.2015 doi: 10.1523/JNEUROSCI.4181-14.2015

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627. doi: 10.1037/0033-2909.83.4.602

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2017, feb). Deliberation increases the wisdom of crowds. *BioArXiv*. Retrieved from arxiv:1703.00045http://arxiv.org/abs/1703.00045

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *The Psychology of Learning and Motivation*, *26*, 125–173.

Nickerson, R. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of general psychology*, *2*, 175–220.

Oaksford, M., & Chater, N. (1991, mar). Against Logicist Cognitive Science. *Mind & Language*, *6*(1), 1–38. Retrieved from http://doi.wiley.com/10.1111/j.1468-0017.1991.tb00173.x doi: 10.1111/j.1468-0017.1991.tb00173.x

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004, apr). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, N.Y.)*, *304*(5669), 452–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15087550 doi: 10.1126/science.1094285

Patel, D., Fleming, S. M., & Kilner, J. M. (2012, dec). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4853–4860. Retrieved from http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2012.1847 doi: 10

.1098/rspb.2012.1847

Peirce, C. S., & Jastrow, J. (1884). On Small Differences in Sensation. *Memoirs of the of Sciences*, *3*, 73–83.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, & Law*, *1*, 817–845.

Persaud, N., McLeod, P., & Cowey, A. (2007, feb). Post-decision wagering objectively measures awareness. *Nature neuroscience*, *10*(2), 257–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17237774 doi: 10.1038/nn1840

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965. Retrieved from http://dx.doi.org/10.1037/xge0000180.supphttp://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000180 doi: 10.1037/xge0000180

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. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0019737 doi: 10.1037/a0019737

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016, feb). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. Retrieved from http://www.nature.com/doifinder/10.1038/nn.4240 doi: 10.1038/nn.4240

Price, P. C., & Stone, E. R. (2004, jan). Intuitive evaluation of likelihood judgment producers: evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, *17*(1), 39–57. Retrieved from http://doi.wiley.com/10.1002/bdm.460 doi: 10.1002/bdm.460

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from https://www.r-project.org/

Rabbitt, P. M. (1966, feb). Errors and error correction in choice-response tasks. *Journal of experimental psychology*, *71*(2), 264–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5948188

Ramsoy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. Retrieved from http://link.springer.com/10.1023/B:PHEN.0000041900.30172.e8 doi: 10.1023/B:PHEN.0000041900.30172.e8

Rauhut, H., & Lorenz, J. (2011, apr). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, *55*(2), 191–197. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0022249610001185 doi: 10.1016/j.jmp.2010.10.002

Rausch, M., & Zehetleitner, M. (2016). Visibility Is Not Equivalent to Confidence in a Low Contrast Orientation Discrimination Task. *Frontiers in psychology*, *7*, 591. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27242566http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4874366 doi: 10.3389/fpsyg.2016.00591

Rizzolatti, G., & Sinigaglia, C. (2010, apr). The functional role of the parieto-frontal

mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience*, *11*(4), 264–274. Retrieved from http://www.nature.com/doifinder/10.1038/nrn2805 doi: 10.1038/nrn2805

Roediger III, H. L., Wixted, J. H., & Desoto, K. A. (2012, jul). The Curious Complexity between Confidence and Accuracy in Reports from Memory. In *Memory and law* (pp. 84–117). Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199920754.001.0001/acprof-9780199920754-chapter-4 doi: 10.1093/acprof:oso/9780199920754.003.0004

Rosenberg, L. B. (2015, jul). Human Swarms, a real-time method for collective intelligence. In *Proceedings of the european conference on artificial life 2015* (pp. 658–659). The MIT Press. Retrieved from https://mitpress.mit.edu/sites/default/files/titles/content/ecal2015/978-0-262-33027-5-ch117.pdf doi: 10.7551/978-0-262-33027-5-ch117

Roy, D. (2004). 10x - human-machine symbiosis. *BT Technology Journal*, *22*(4), 121–124.

Sah, S., Moore, D. A., & Maccoun, R. J. (2013). Cheap talk and credibility: The consequences of confidence and accuracy on advisor credibility and persuasiveness. *Organizational Behavior and Human Decision Processes*, *121*, 246–255. doi: 10.1016/j.obhdp.2013.02.001

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010, dec). Measuring consciousness: is one measure better than the other? *Consciousness and cognition*, *19*(4), 1069–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20133167 doi: 10.1016/j.concog.2009.12.013

Schelling, T. C. (1980). *The Strategy of Conflict*. Harvard: Harvard University Press.

Schiffer, A.-M., Siletti, K., Waszak, F., & Yeung, N. (2016). Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *NeuroImage*. doi: 10.1016/j.neuroimage.2016.08.057

Schilbach, L. (2014, may). On the relationship of online and offline social cognition. *Frontiers in Human Neuroscience*, *8*. Retrieved from http://journal.frontiersin.org/article/10.3389/fnhum.2014.00278/abstract doi: 10.3389/fnhum.2014.00278

Schilbach, L. (2016, jan). Towards a second-person neuropsychiatry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150081. Retrieved from http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2015.0081 doi: 10.1098/rstb.2015.0081

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013, aug). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(04), 393–414. Retrieved from http://www.journals.cambridge.org/abstract{_}S0140525X12000660 doi: 10.1017/S0140525X12000660

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006, jan). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–30. Retrieved from http://

www.ncbi.nlm.nih.gov/pubmed/16171833 doi: 10.1016/j.neuropsychologia.2005.07.017

Schotter, A. (2003). Decision Making with Naive Advice. *American Economic Review*, *93*, 196–201.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*(12), 4595–4610.

Schultze, T., Rakotoarisoa, A.-F., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making2*, *10*(2), 144–171. Retrieved from http://journal.sjdm.org/14/141112a/jdm141112a.pdf

See, K. E., Morrison, E. W., Rothman, N. B., & Soll, J. B. (2011, nov). The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes*, *116*(2), 272–285. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597811000975 doi: 10.1016/j.obhdp.2011.07.006

Seth, A. K. (2008, sep). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and cognition*, *17*(3), 981–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17588775 doi: 10.1016/j.concog.2007.05.008

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014, feb). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661314000230 doi: 10.1016/j.tics.2014.01.006

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016, jan). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. Retrieved from http://www.nature.com/doifinder/10.1038/nature16961 doi: 10.1038/nature16961

Silver, N. (2012). *The Signal and the Noise: The Art and Science of Prediction.* Penguin.

Simion, F., Macchi-Cassia, V., Turati, C., & Valenza, E. (2001, mar). The origins of face perception: specific versus non-specific mechanisms. *Infant and Child Development*, *10*(1-2), 59–65. Retrieved from http://doi.wiley.com/10.1002/icd.247 doi: 10.1002/icd.247

Simons, H. W., Berkowitz, N. N., & Moyer, R. J. (1970). Similarity, credibility, and attitude change: A review and a theory. *Psychological Bulletin*, *73*(1), 1–16. Retrieved from http://content.apa.org/journals/bul/73/1/1 doi: 10.1037/h0028429

Slegers, D. W., Brake, G. L., & Doherty, M. E. (2000). Probabilistic Mental Models with Continuous Predictors. *Organizational Behavior and Human Decision Processes*, *81*(1), 98–114. doi: 10.1006/obhd.1999.2869

Smith, E., & Mackie, D. (2007). Liking and Loving. In *Social psychology* (3rd editio ed., p. 367). Psychology Press.

Smith, J. D., Couchman, J. J., & Beran, M. J. (2012, apr). The highs and lows of theoretical interpretation in animal-metacognition research. *Philosophical Trans-*

actions of the Royal Society B: Biological Sciences, *367*(1594), 1297–1309. Retrieved from http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2011.0366 doi: 10.1098/rstb.2011.0366

Smith, K. (2013). Reading minds. *Nature*, *502*, 428–430.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*, 135–168.

Sniezek, J. A., & Buckley, T. (1989). Social influence in the advisor-judge relationship. In *Annual meeting of the judgment and decision making societ.* Atlanta, Georgia.

Sniezek, J. A., & Buckley, T. (1995, may). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597885710400 doi: 10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. M. (2001, mar). Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational behavior and human decision processes*, *84*(2), 288–307. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11277673 doi: 10.1006/obhd.2000.2926

Soll, J. B., & Larrick, R. P. (2009, may). Strategies for revising judgment: how (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 780–805. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19379049 doi: 10.1037/a0015145

Soll, J. B., & Mannes, A. E. (2011, jan). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, *27*(1), 81–102. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169207010000877 doi: 10.1016/j.ijforecast.2010.05.003

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011, dec). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and cognition*, *20*(4), 1787–92. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3203218{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1016/j.concog.2010.12.011

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.108.1.183 doi: 10.1037/0033-295X.108.1.183

Stasser, G., & Titus, W. (2003). Hidden Profiles : A Brief History. *Psychological Inquiry*, *14*(3), 304–313.

Stokes, M. G. (2015, jul). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1364661315001023 doi: 10.1016/j.tics.2015.05.004

Sunstein, C. R. (2001). *Republic.com.* Princeton University Press.

Surowiecki, J. (2004). *The Wisdom of Crowds. Why the Many are Smarter than the Few.* (Abacus ed.). London: Little, Brown Book Group.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: an introduction.*

Cambridge, MA: MIT Press.

Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British journal of social psychology*, *44*(3), 443–461.

Tenney, E. R., MacCoun, R. J., Spellman, B. a., & Hastie, R. (2007, jan). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17362377 doi: 10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011, jul). Accuracy, confidence, and calibration: how young children and adults assess credibility. *Developmental psychology*, *47*(4), 1065–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21443337 doi: 10.1037/a0023273

Tenney, E. R., Spellman, B. A., & Maccoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368–1375. doi: 10.1016/j.jesp.2008.04.006

The MathWorks Inc. (2016). *MATLAB version 9.0.0 (R2016a)*. Natick, Massachusetts, United States: The MathWorks Inc.

Todd, J. (1955). The Syndrome of Alice in Wonderland. *Canadian Medical Association Journal*, *73*(9), 701–704.

Tost, L. P., Gino, F., & Larrick, R. P. (2012, jan). Power, competitiveness, and advice taking: Why the powerful don't listen. *Organizational Behavior and Human Decision Processes*, *117*(1), 53–65. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597811001233 doi: 10.1016/j.obhdp.2011.10.001

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... Nelson, C. (2009, aug). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, *168*(3), 242–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3474329{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1016/j.psychres.2008.05.006

Treutwein, B. (1995). Adaptive Psychophysical Procedures. *Vision Research*, *35*(17), 2503–2522.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, More than Confidence, Explain the Good Performance of Reasoning Groups. *Journal of Experimental Psychology: General*, *143*(5), 1958.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016, mar). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1519157113 doi: 10.1073/pnas.1519157113

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. Retrieved from http://mind.oxfordjournals.org/cgi/doi/10.1093/mind/LIX.236.433 doi: 10.1093/mind/LIX.236.433

Turner, M. E., & Pratkanis, A. R. (1998, feb). Twenty-Five Years of Groupthink Theory and Research: Lessons from the Evaluation of a Theory. *Organizational Behavior and Human Decision Processes*, *73*(2-3), 105–115. Retrieved from

http://linkinghub.elsevier.com/retrieve/pii/S074959789892756X doi: 10.1006/obhd.1998.2756

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. Retrieved from http://www.jstor.org/stable/1738360Copy

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. Retrieved from http://content.apa.org/journals/rev/90/4/293 doi: 10.1037/0033-295X.90.4.293

Van den Broeck, Y., Pescetelli, N., & Yeung, N. (2017). *Exploiting alignment to improve calibration.*

Van Swol, L. M. (2011, jan). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0169207010000415 doi: 10.1016/j.ijforecast.2010.03.002

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5:e12192*. doi: 10.7554/eLife.12192

Vandormael, H., Herce Castañón, S., Balaguer, J., Li, V., & Summerfield, C. (2017, mar). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences*, *114*(10), 2771–2776. Retrieved from http://www.pnas.org/lookup/doi/10.1073/pnas.1613950114 doi: 10.1073/pnas.1613950114

Vickers, D. (1979). *Decision Processes in Visual Perception.* London: Academic Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(June), 440–442. Retrieved from http://www.nature.com/nature/journal/v393/n6684/full/393440a0.html doi: 10.1038/30918

Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (1967). *Pragmatics of Human Communication:a study of interactional patterns, pathologies and paradoxes.* New York: W. W. Norton & Company.

Weiss, D. J., & Shanteau, J. (2003, jan). Empirical assessment of expertise. *Human factors*, *45*(1), 104–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12916584

Wheeler, L. (1966). Toward a theory of behavioral contagion. *Psychological Review*, *73*(2), 179–192. Retrieved from http://content.apa.org/journals/rev/73/2/179 doi: 10.1037/h0023023

Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, *18*, 107–126.

Yaniv, I. (2004a, apr). The Benefit of Additional Opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. Retrieved from http://cdp.sagepub.com/lookup/doi/10.1111/j.0963-7214.2004.00278.x doi: 10.1111/j.0963-7214.2004.00278.x

Yaniv, I. (2004b, jan). Receiving other people's advice: Influence and benefit. *Or-*

ganizational Behavior and Human Decision Processes, *93*(1), 1–13. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749597803001018 doi: 10.1016/j.obhdp.2003.08.002

Yaniv, I., & Kleinberger, E. (2000, nov). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational behavior and human decision processes*, *83*(2), 260–281. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11056071 doi: 10.1006/obhd.2000.2909

Yates, J., Price, P. C., Lee, J.-W., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting*, *12*(1), 41–56. doi: 10.1016/0169-2070(95)00636-2

Yeung, N., & Summerfield, C. (2012, may). Metacognition in human decision-making: confidence and error monitoring. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *367*(1594), 1310–21. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3318764{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1098/rstb.2011.0416

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of Postdecisional Processing of Confidence. *Journal of experimental psychology. General*, *144*(2), 489–510. doi: 10.1037/xge0000062

Zajonc, R. B., Adelmann, P. K., Murphy, S. T., & Niedenthal, P. M. (1987, dec). Convergence in the physical appearance of spouses. *Motivation and Emotion*, *11*(4), 335–346. Retrieved from http://link.springer.com/10.1007/BF00992848 doi: 10.1007/BF00992848

Zarnoth, P., & Sniezek, J. A. (1997, jul). The Social Influence of Confidence in Group Decision Making. *Journal of Experimental Social Psychology*, *33*(4), 345–366. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0022103197913263 doi: 10.1006/jesp.1997.1326

Appendix A

A reinforcement learning model in the absence of feedback

One of the questions that motivated the present work was how can people learn about the reliability of their advisers, when objective feedback is not available. The questions is challenging because most current models of learning, like the delta-rule, are based on a outcome signal (e.g., feedback or reward) that is missing in feedback-free scenarios. Although the model presented in Chapter 3 provides a useful working example of how this could be achieved through subjective trial-level confidence, I felt the need to explore how a reinforcement learning rule could also integrate a confidence signal and what performance in our task could be expected by such model.

The classic delta rule is a gradient descent rule that updates current expected values (for example expected reward) proportionally to its error, namely the difference between the observed outcome r (in our case the objective accuracy feedback $\in \{0,1\}$) and the expected outcome w (in our case the expected accuracy of the advice θ):

$$w_t = w_{t-1} + \alpha(r - w_{t-1}) \tag{A.1}$$

where α is the learning rate. Similarly to what done in the main text, three models can be designed depending on feedback condition and metacognitive access. In the feedback condition, r is the feedback participants receive on adviser's performance (AccuracyRL). In no-feedback condition, a model without metacognitive access can be

designed by replacing the feedback signal r with agreement (ConsensusRL). Finally, in feedback-free scenarios with metacognitive access, the equation can be re-adapted so that instead of using the objective outcome r, the model uses the estimated probability of the outcome $\hat{r}$ - obtained from participants' expressed confidence judgments - to update expected accuracy separately for each advisor v:

$$w_t^v = w_{t-1}^v + \alpha(\hat{r} - w_{t-1}^v) \tag{A.2}$$

$$\hat{r} = p^A \neg p^{\neg A} \tag{A.3}$$

where p is the participant's confidence expressed as a probability judgment over the outcome RIGHT, according to equation 3.10 in main text (with $p + \neg p = 1$). A and $\neg A$ assume values of 0 and 1 in disagreement and values 1 and 0 in agreement respectively.

Experiment 3

I applied the model only to Experiment 3, given that this experiment was the one that best discriminated between a consensus-based strategy and agreement-in-confidence strategy. Also, given that the model has a free parameter α, the model was fit only to participants in the relevant feedback group. The α parameter was fit to each individual participant so to minimise the root mean squared residual between empirical confidence updates and predicted confidence update. As shown in Figure A.1, the AccuracyRL model did not show any effect of adviser on average expected adviser reliability w ($F(2, 48) = 1.16, p = .32, \eta_G^2 = .005$). The ConsensusRL model did not show any effect of adviser either ($F < 1$). A ConfidenceRL model, on the contrary, showed a marginal effect of adviser ($F(2, 44) = 3.15, p = .05, \eta_G^2 = .03$). The bias-sharing adviser was significantly trusted more by the model compared to the anti-bias adviser ($t(22) = 2.18, p = .04, d = 0.43$), but not compared to the

unbiased adviser ($t(22) = 1.02, p = .31, d = 0.19$). A marginal difference also existed between the unbiased and anti-bias adviser ($t(22) = 1.76, p = .09, d = 0.31$). Figure A.1 shows w values (corresponding to θ values in the main model), representing the model average expected adviser reliability. Mean fitted α for the ConfidenceRL model was 0.05 ± 0.15.

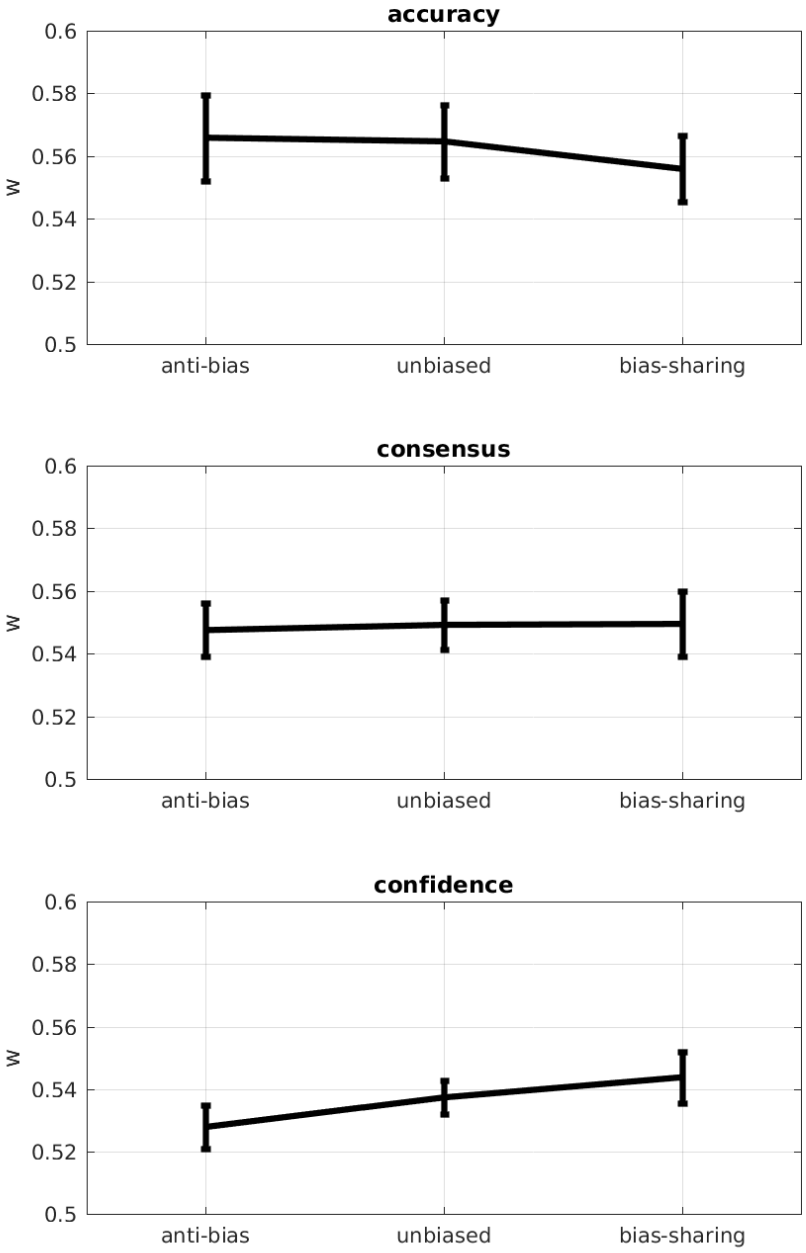


Figure A.1: Reinforcement learning model adapted to each feedback condition.

Appendix B

Analysis of reaction times during social exchange

Experiments 4-7 showed that interactive conditions significantly alter the independence of members' judgments, making confidence change magnitudes positively correlated in agreement and negatively correlated in disagreement. Results suggested that participants were interpreting the magnitude of their partners' updates as another indicator of their partner's confidence. Another possibility, however, is that in the Interactive condition participants interpreted their partner's speed to react as a cue for uncertainty. In other words, the more quickly participants were willing to opinion (i.e., move their own cursor), the more uncertain they would have been perceived by their partners. The following analysis was done to control for this "speed to respond" explanation.

Experiment 4

Due to the fixed time allocated to the social part, a direct measure of reaction times was not available. To test for this hypothesis we fitted a sigmoid curve to each trial's confidence time series, namely each member's cursor's positions over the five seconds each social exchange lasted. Slope ϕ and offset λ along the time axis were free parameters to be estimated on each trial. The offset parameter λ was used as a proxy for reaction times and entered into an ANOVA with factors dominance (more confident vs. less confident), condition and consensus (agreement vs. disagreement).

Results showed a significant effect for all main effects ($F(1, 47) > 7.25, p < .009$) and a significant interaction between dominance and consensus ($F(1, 47) = 4.98, p = .03$). Pairwise comparisons showed that members who started off less confident than their partners were slower in updating their cursor ($t(47) = 7.91, p < .001$), making the hypothesis that cursor's stickiness was interpreted as a cue for confidence implausible. Furthermore, estimated reaction times λ were slower in the Interactive than Non-Interactive condition ($t(47) = 2.69, p = .009$) and faster in agreement than in disagreement ($t(47) = -7.71, p < .001$), indicating that, if anything, longer reaction times were associated with more uncertainty (Patel, Fleming, & Kilner, 2012).

Experiment 5

Fitted λ values were entered into a three-way repeated measures ANOVA with factors opinion dominance, condition and consensus. Results show a significant effect for dominance ($F(1, 47) = 6.59, p = .01, \eta^2_G = .006$), pointing out that the dominated individual showed significantly greater reaction times than the dominant individual ($t(47) = 2.56, p = .01$). Disagreement trials showed significantly greater update times than agreement trials ($F(1, 47) = 10.18, p = .002, \eta^2_G = .03$). A significant effect of condition was found ($F(2, 94) = 4.36, p = .01, \eta^2_G = .002$) suggesting different conditions produced different confidence update times. Pairwise tests among conditions showed that the Interactive condition had significantly longer reaction times than the Non-Interactive condition ($t(47) = 2.48, p = .01$), and compared to the Interactive$_{self}$ condition ($t(47) = 2.66, p = .01$). No difference was found between the Interactive$_{self}$ and Non-Interactive conditions. Thus the effect of adding a reminder to the Interactive condition was to reduce the confidence update times found in this condition. Finally a significant interaction was found between condition and consensus ($F(2, 94) = 6.16, p = .003, \eta^2_G = .002$). Pairwise comparisons showed that the difference in reaction times between disagreement and agreement trials was significantly greater in the Non-Interactive condition than in either the Interactive condition

($t(47) = 2.03, p = .04$) and Interactive-plus-anchor condition ($t(47) = 3.80, p < .001$). No difference was found between interactive conditions ($p > .1$).

Experiment 6

A three-way ANOVA on fitted λ values showed a marginal effect of dominance ($F(1, 47) = 3.18, p = .08, \eta^2_G = .003$), a significant effect of consensus ($F(1, 47) = 28.23, p < .001, \eta^2_G = .04$) but no effect of condition ($F < 1$). Furthermore no significant two-way interaction was found (all $F < 1$) and no significant three-way interaction ($F < 1$). Further comparisons showed that also for this study members holding the dominant opinion were faster in updating their confidence in the social part (although here not significantly), suggesting that the speed of the update was unlikely to be used as an indicator of uncertainty. Agreement trials showed significantly smaller reaction times ($t(47) = 5.31, p < .001$), again showing that fast reaction times were more indicative of low uncertainty than high uncertainty. In conclusion interpreting the sluggishness of the partner's update as a sign of their confidence does not seem to be a likely strategy adopted by participants, given that more confident opinions also generated faster updates.

Appendix C

Confidence alignment

Experiment 5

A bootstrap analysis was performed on the data by randomly pairing participants into nominal dyads and computing the Pearson's correlation coefficient of the distribution means between two dyad members. The random pairing procedure was repeated 1000 times and then compared against the Pearson's correlation coefficient between distribution means observed in empirical dyads. The same procedure was repeated for the correlation between distribution standard deviations. The empirical correlation coefficients for both means and standard deviations were in the 99.9 percentile of the distribution observed in nominal dyads, suggesting that the observed correlation coefficients were significantly higher than what should be expected by chance.

To control that the high correlation coefficients observed were not driven by trial-level features observed by members of a same dyad the bootstrap procedure was repeated with trial-level confidence ratings to test whether trial-level confidence ratings were correlated. The results show that the correlation coefficient at the trial level was less than one standard deviation away (z=0.90) from the mean correlation coefficient that should be expected by chance. The result thus shows that alignment could not be attributed to common characteristics of the stimuli observed by the same dyad's members.

Finally a two-way ANOVA with factors decision type (pre-social information vs post-social information) and condition on the Goodman-Kruskal gamma correlation coefficient of trial-level accuracy was performed to test whether the alignment observed in confidence distributions was due to the co-variation in trial-level accuracy between same dyad members. Results show a significant effect of decision ($F(1, 23) = 90.93, p < .001, \eta^2_G = .51$) suggesting that members' accuracy was more correlated post-advice than pre-advice (M $\pm$ SDT: -0.06 ± 0.25 vs. 0.56 ± 0.31), but no effect for condition ($F < 1$). No significant interaction was found between the two main effects ($F < 1$).

Experiment 6

Empirical correlation coefficients between means of confidence distributions of dyad members were in the 99.99 percentile indicating that observed correlation was way larger than what should be expected by chance. A similar procedure was applied to the standard deviations, indicating the 91 percentile compared to the distribution of correlation coefficients that should be expected by chance, suggesting that participants aligned standard deviations to a less degree than mean confidences.

To test whether the alignment effect is driven by similar characteristics in the stimuli observed by the two participants, correlation coefficients were computed between trial-level confidence ratings of the two members of a same dyad. Results for correlation in trial-level confidence showed that coefficients were in the 84 percentile compared to a chance distribution, well below what observed for average confidence. The results thus corroborate the conclusions suggesting that confidence distributions were more similar when belonging to participants of a same dyad than of different dyads over and above what could be attributed to similarity in the visual setting.

Finally, a two-way ANOVA on gamma correlation coefficients of trial-level accuracy with factors decision type (pre-social information vs post-social information) and condition showed significant effects for condition ($F(2, 46) = 4.64, p = .01, \eta^2_G = .03$),

decision type ($F(1, 23) = 205.51, p < .001, \eta^2_G = 0.65$), but no significant interaction term ($F(2, 46) = 1.35, p = .26, \eta^2_G = .003$). Pairwise comparisons showed that none of the conditions showed gamma coefficients significantly different from zero before social information was exchanged ($t(23) < 1.38, p > .17$), suggesting that pre-advice accuracy of the two members was not significantly correlated.

Appendix D

Social information perception analysis

Social information perception analysis was performed in opinion space to see where differences between objective and perceived supporting evidence differed the most.

Experiment 4

Figure D.1 shows the difference between objective and perceived social evidence along the opinion surface. Positive values (warmer colours) indicate that social information is perceived and/or used as more in agreement with one's own initial decision that what objectively stated by one's partner. Negative values (colder colours) indicate that social information is perceived as more in opposition to one's views than actually stated by one's partner. As expected on average there is a bias towards interpreting social information in favour of one's initial opinion, as shown by the extent of positive areas. This effect is particularly true in disagreement and more so in dominated trials than dominant trials. Notice the larger positive areas in weak dominant agreement produced by interaction (top-right plot) compared to non-interaction (top-left plot).

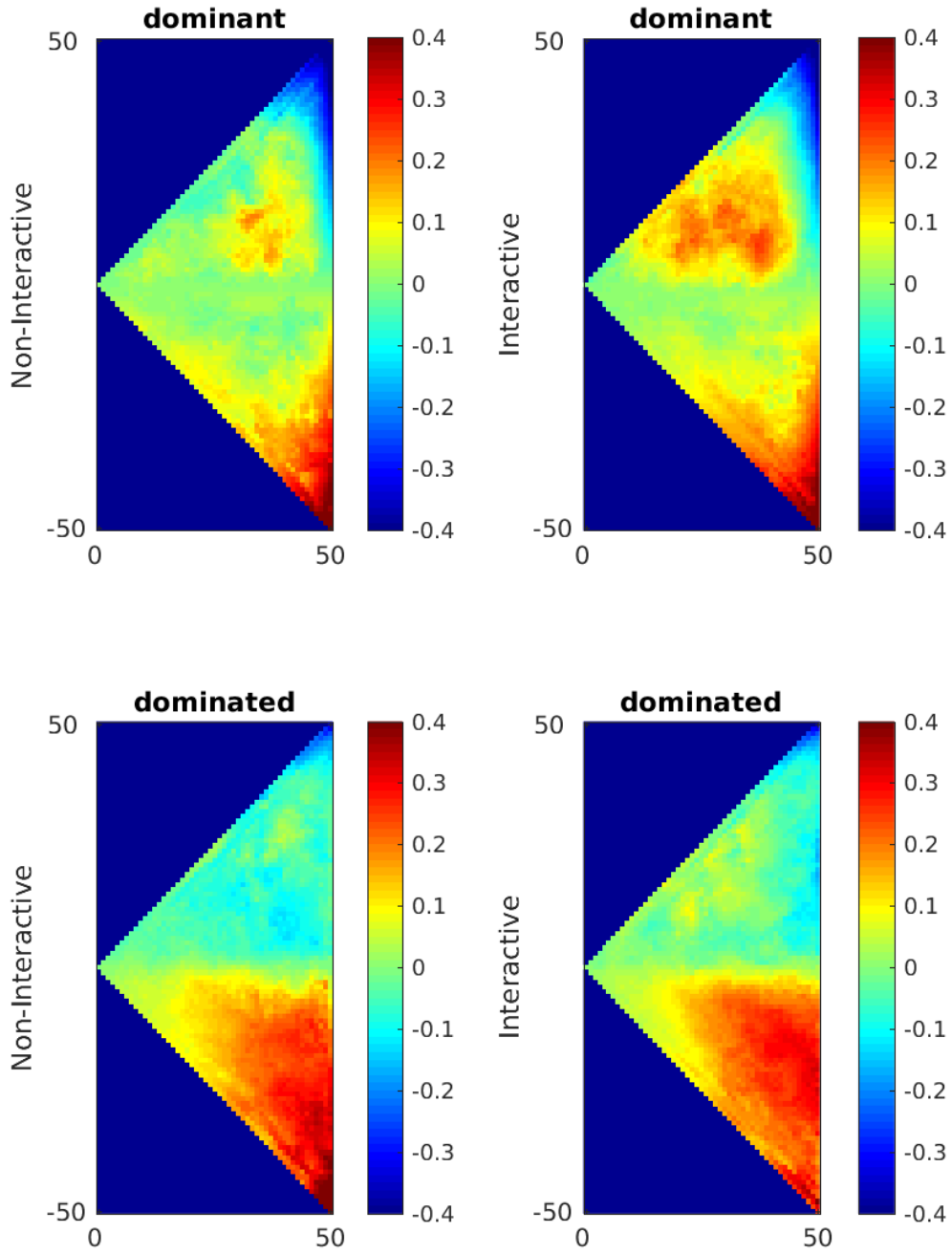


Figure D.1: Median difference between stated and perceived evidence in opinion space.

Experiment 5

Similarly, Figure D.2 shows the difference between objective and perceived social evidence. No evident differences emerge among conditions.

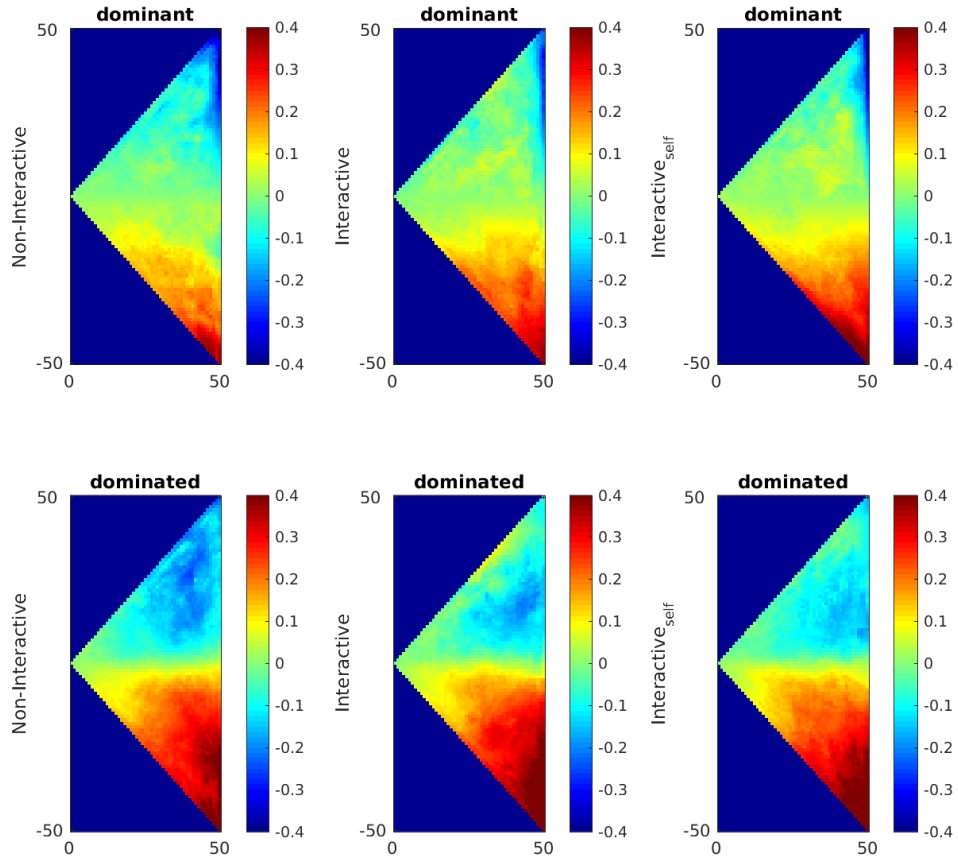


Figure D.2: Median difference between stated and perceived evidence in opinion space.

Experiment 6

Finally, Figure D.3 shows the difference between objective and perceived social evidence in Experiment 6. No evident difference is present among conditions.

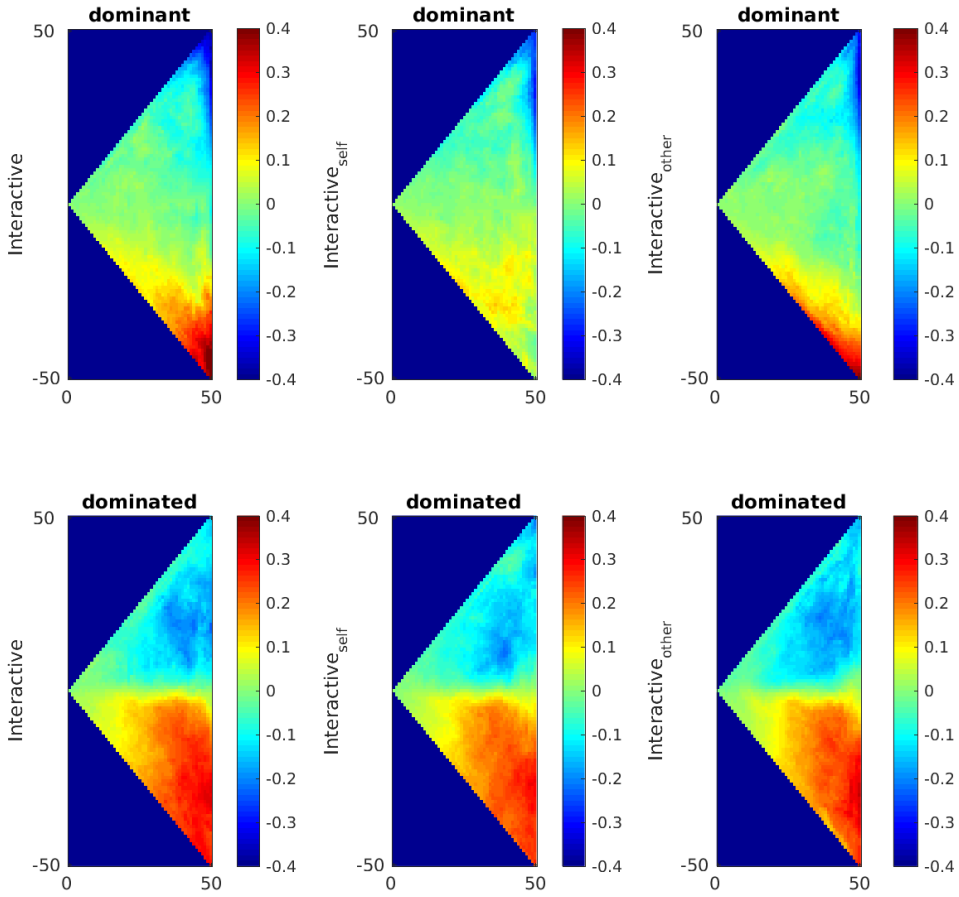


Figure D.3: Median difference between stated and perceived evidence in opinion space.

Appendix E

Experiment 8

Method

Participants. 20 participants (age=26.6±4.29; 12 females) were recorded during the poster session of a local conference on a volunteer basis. Participants received chocolate bars for compensation.

Paradigm. The experiment was composed of 200 trials. Each trial started with the dot task already described previously, with identical presentation parameters. Difficulty was titrated for each participant to reach a 70.7% accuracy in pre-interaction decisions. After participants confirmed their response, the second part of the trial started. The second part was a confidence update window which lasted 2 seconds, during which participants were asked to track their confidence changes over time while their cursor position along the x-axis was being recorded at a frequency of 5 Hz (11 data points per trial). Two conditions were randomly shuffled across trials. The Self condition allowed participants to update their original confidence if thought to be appropriate. No extra information nor advice was provided. This provided a reference to control baseline improvements due to naturally occurring post-decisional processes (Pleskac & Busemeyer, 2010). In the Other condition, the participant was paired with a dynamic model that reproduced realistic cursor updates. The model's expressed confidence originated from random Gaussian noise added to the participant's own

judgments, thus did not provide any useful information. As neither condition provided extra task-relevant information conditions are not expected to differ according to confidence, accuracy and calibration. The only variable that is manipulated is the presence of another (uninformative) agent.

All participants received feedback at the end of every trial indicating the accuracy of their final decision. Although participants could in principle used feedback to realise that the model is uninformative and thus discarding its opinion, interaction might create ripples of confidence escalation making participants more confident, particularly in weak agreement trials. The Self and Other conditions were made distinguishable to the participant by the presentation at the center of the screen of a small picture of a famous robot[1] indicating that the current trial was Other condition.

Model design. The dynamic model's decision (LEFT vs RIGHT) on a given trial was *a priori* determined to agree with the participant's decision on 60% of the trials, mimicking the nominal agreement rate of another equally accurate independent participant. The model's confidence level at the beginning of the social part was determined by taking a target value randomly drawn from a Gaussian distribution centred on a variable mean and fixed standard deviation. The mean of the distribution was centred on the human current confidence in agreement trials and on half scale away from it (i.e., 50 confidence points) in disagreement trials. For example, if at a given time point the participant is expressing a confidence level of 30, the model is most likely to have a confidence of 30 on agreement trials and -20 (20 on the opposite interval) on disagreement trials. The standard deviation was varied across participants by randomly selecting a value at the beginning of the experiment sampled from a uniform distribution between 6 and 15. On every iteration of the internal loop controlling the model's behaviour, the model stayed on the target value with 90% probability and selected a different target value with 10% probability, using the

[1] Wall-E, Disney Pixar

participant's current position as its new reference mean. This ensured that if participants shifted their confidence, the model would tend to update its confidence too, but without giving the sense of exactly mirroring participant's behaviour. When a new target value was selected, the model reached this target by continuously reducing the distance between its current position and the target over successive loop iterations. This algorithm allows a dynamic coupling whereby the model uses the current human confidence to probabilistically update its expressed confidence and the human observes the model's confidence update, which can in turn inform his/her judgment.

Unfortunately, due to a bug in the design introducing a leftwards bias in the model's confidence, disagreement trials where the participant started on the left hand side of the scale (LEFT decision) had to be removed.

Results

Manipulation check. Figure E.1 shows the distribution of the model's confidence judgments relative to the participant's momentary opinion and divided by consensus. On agreement trials the model's confidence distribution is centred on the participant's current expressed confidence (zero along the x-axis). In disagreement trials, on the contrary, the distribution is centred on -50, corresponding to a decision on the other half of the scale. The model thus provided no stimulus-relevant information to the participant and simply mirrored the participant's behaviour.

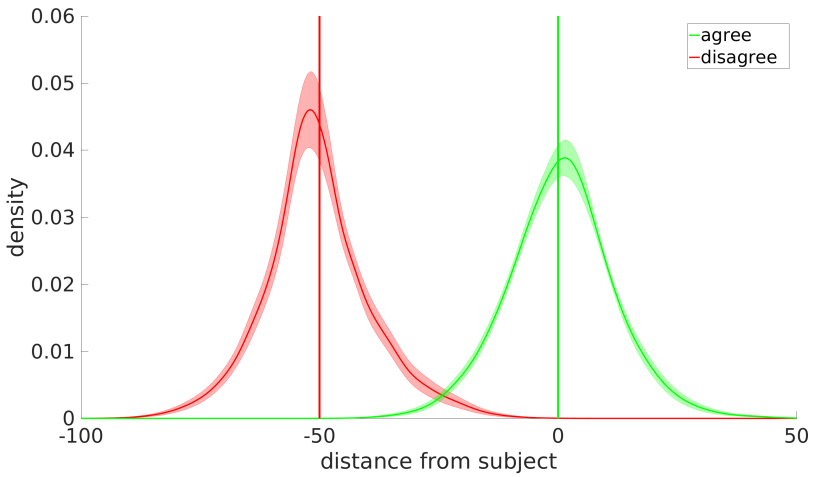


Figure E.1: The distribution of model's confidence in agreement and disagreement trials, relative to the participant's current confidence.

Confidence distributions. Figure E.2 shows the pre-advice confidence distributions characterising participants before the post-decisional phase begun. A two-way ANOVA on average confidence with factors decision type (pre- versus post-update) and condition (Self vs. Other) showed no significant effect of either factor ($F < 1$), suggesting that participants' average confidence was similar across conditions and did not significantly change from pre- to post-decision time. A marginally significant interaction ($F(1, 19) = 3.89, p = .06, \eta^2_G = .0005$) provided preliminary evidence that average confidence increased more from pre- to post-decisional time in the Other condition rather than in the Self condition.

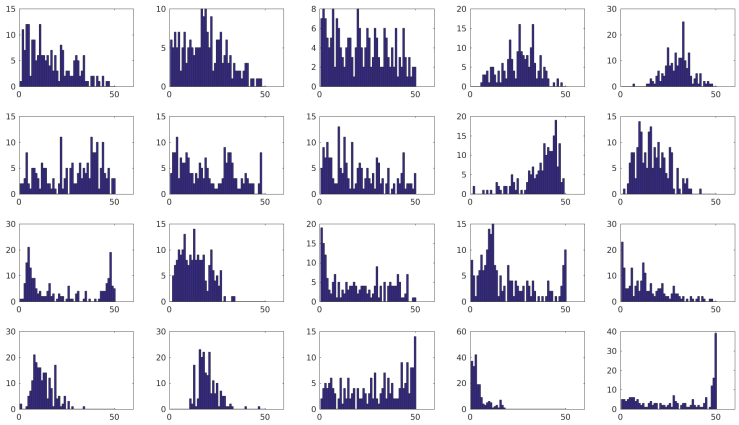


Figure E.2: Pre-advice confidence distributions recorded for each participant.

Confidence change. To better understand the reasons behind such unexpected dissociation, the distribution of human-model confidence distance in agreement and disagreement were plotted over the two seconds update interval. An animated version can be found at `https://niccolopescetellidotcom.files.wordpress.com/2017/04/hdc.gif`. Figure E.3 shows only the last time point recorded, just before the trial ended. It can be observed that part of the original mass of the disagreement distribution (Figure E.1) has now moved towards the participant's opinion (zero). These are trials in which either the model or the participant changed their mind by changing interval. Given the model's design, once either the model or the participant changed their mind, the model's confidence distribution quickly gravitated towards the participant's current level of confidence.

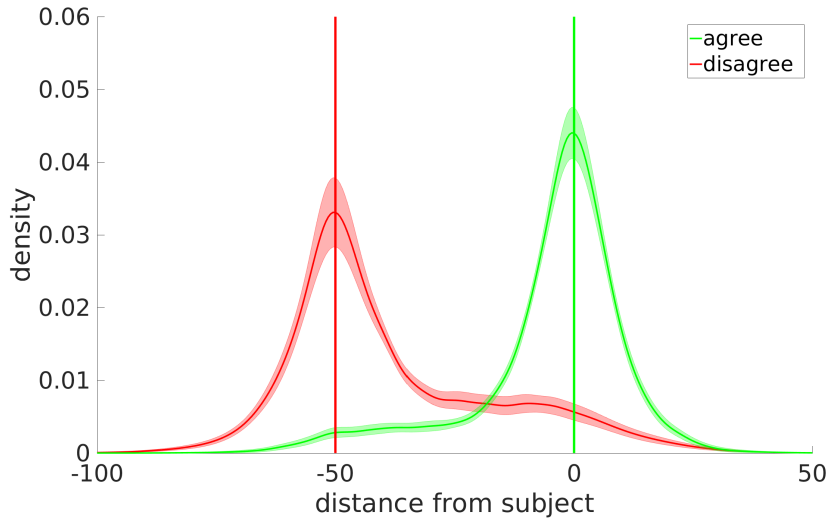


Figure E.3: The distribution of model's confidence in agreement and disagreement trials, relative to the participant's current confidence after update has taken place.

A more informative analysis however would be to plot confidence change for model and participants separately along the opinion surface. Figure E.4 shows the confidence changes observed in the Other condition divided by trials in which the participant was dominant (S_{ic} on x-axis) and dominated (S_{ic} on y-axis). Subscript ic indicates initial confidence. An real-time version of confidence change in the two can be found at `https://niccolopescetellidotcom.wordpress.com/?page_id=498&preview=true`. The top-left and bottom-right graphs show the participant's confidence changes when the participant started from the dominant opinion and the dominated opinion, respectively. Notice that color scales of participants' updates (δ_S) and model's updates (δ_M) have different ranges, indicating that participants' updates were on average smaller than the model's. Although participants' confidence was robust to change when the model's opinion was dominated (top-left graph), this was not the case when the model was dominant (bottom-right graph). In particular, participants who started in a low-confidence disagreeing position (point y) were more likely to change their mind towards the model's opinion. These trials are the ones to drive the dissociation observed between calibration improvements and

accuracy decreases. Given that the participant's accuracy was titrated to 70%, the participant was more likely to be correct than wrong. Changes of mind in low confidence correct trials had thus the effect of turning into low confidence incorrect trials, thus explaining the decrease in accuracy and increase in calibration. Similarly, when starting from low-confidence dominated agreement (point x) participants seemed to increase their confidence. Similarly to Experiments 4-6, points x and y on the figure, corresponding to uncertain agreement or disagreement respectively, are characterised by the greatest confidence shifts in interactive social exchanges. These trials are also the ones characterised by the greatest discrepancy between objective and perceived social information as described in Experiments 4-6 using inverse Bayes. Thus, when subjects are paired with an interactive agent with no task-relevant information, participants' behaviour seems to reproduce the sensitivity of these trials to biases and distortions already observed in human-human dyads.

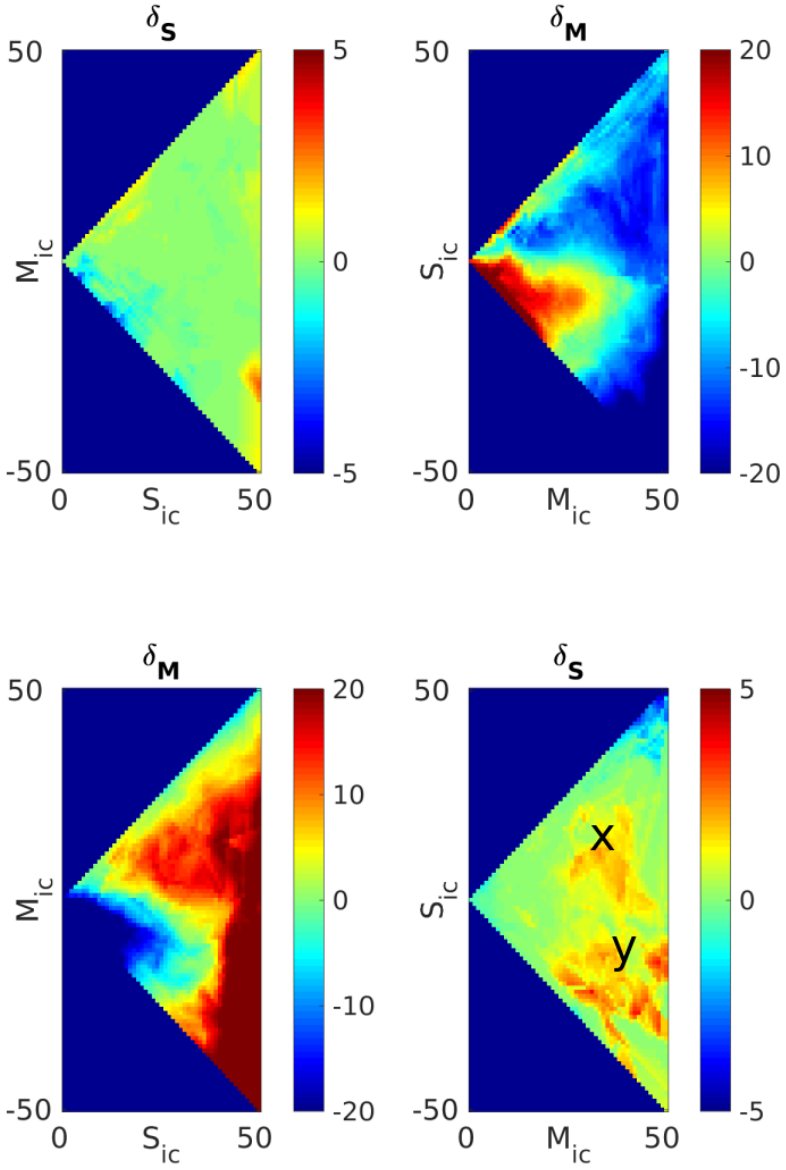


Figure E.4: The figure above shows opinion space plots observed in the Other condition. Although the participant's confidence was quite resistant to change when the participant started from a dominant opinion (top-left panel), this was not the case when it was dominated by the model (bottom-right panel). In these trials agreement led participants to increase their initial confidence and in disagreement to reduce it (or change their mind). This behaviour can turn low confident correct trial into low confidence error trials, explaining some of the findings reported above.