Where the truth lies: how sampling implications drive deception without lying

Keith Ransom[a,*], Wouter Voorspoels[b], Danielle J. Navarro[c], Amy Perfors[a]

[a]*School of Psychological Sciences, University of Melbourne, Australia*
[b]*Brain & Cognition, University of Leuven, Belgium*
[c]*School of Psychology, University of New South Wales, Australia*

Abstract

Efficient communication leaves gaps between message and meaning. Interlocutors, by reasoning about how each other reasons, can help to fill these gaps. To the extent that such *meta-inference* is not calibrated, communication is impaired, raising the possibility of manipulation for deceptive ends. We examined how people reason when acting as the perpetrator or target of deception across two related experiments. Importantly, the nature of the task precluded outright lying. Thus, deception required withholding information or providing data that was factually correct but nonetheless misleading. We find evidence for two distinct patterns of behaviour. One group of people appear to make assumptions about communicative intent based on context and message content. Senders in this group were more likely to mislead, and receivers were more effectively misled. A second group of people appeared to adopt a more defensive stance, displaying the same cautious approach in all situations. We explain this behaviour using a computational account of the kinds of inferences required by both receiver and sender. These distinct patterns arise from different assumptions about the generative process behind communication.

Keywords: Bayesian modelling, deception, meta-inference, sampling assumptions

1. Introduction

Inference on the basis of real-world communication is a complex and under-constrained problem. Messages (not unlike flat-packed furniture) rarely come complete with everything necessary to assemble what was intended. Over and above decoding the message on syntactic and semantic grounds, the receiver must also fill in gaps based on her existing knowledge and her inductive biases. In so doing, she may make assumptions about the way that the sender chooses what to say on the basis of what he means to convey (Grice, 1989). Likewise, a sender who seeks to convey

*Corresponding author
Email address: keith.ransom@unimelb.edu.au (Keith Ransom)

a given meaning may make his own assumptions about the receiver and how she will decode his meaning from the contents of his message. Critically, both sender and receiver may recognise that for each assumption they themselves make, their interlocutor may assume that they make it. This pattern of reciprocal and potentially recursive "meta-inference" may be leveraged by both parties. This can enhance their ability to communicate accurately yet efficiently and result in stronger conclusions and more decisive action.

However, meta-inferential reasoning presents challenges of its own. Meta-inference, like inference in general, is under-determined: as a result, successful inference relies on making assumptions that are appropriate to the problem space. Without adequate mechanisms to ensure that the assumptions of senders and receivers are reciprocally calibrated, communication can be significantly impaired. Communicative conventions which rest on the presumption that communication is generally truthful, cooperative and goal-directed offer a basis for calibrated meta-inference.

At the same time, such conventions raise the possibility of manipulation for deceptive ends. Once we accept that cooperative principles may not always hold, communicative meta-inference becomes a vital prerequisite. Without it, people who try to lie will fail to do so successfully and people who are lied to will be consistently taken in. The spectre of deception thus both necessitates and complicates meta-inference. The inferential problem is especially difficult because skilful deceivers may strive to avoid detection by never lying outright. This kind of deception, known as *paltering* (Schauer & Zeckhauser, 2009), occurs when the communicator takes advantage of what they know about the listener's mental state to provide information that is not strictly false but will cause the listener to draw the wrong conclusion.

How *do* people reason in contexts where the goals of sender and receiver are not necessarily aligned? In the present study, we investigate this issue in a setting where deception is warranted but outright lying does not occur. Using behavioural data from two related experiments, one sender focused and one receiver focused, we examine how meta-inference (and ultimately inference) is affected when cooperative norms may no longer apply. We then rely on a model-based analysis to address a number of further questions. First, intuitively, we expect receivers to reason from evidence (messages) differently based on the perceived intent of the sender: why should this be the case when the veracity of the evidence is beyond question? Second, do receivers use cues from the message content itself in order to gauge the sender's intent? And finally, how do these factors affect the sender when deciding whether to mislead or conceal information?

Our approach here complements descriptive accounts of pragmatic reasoning (Grice, 1989) and verbal deception (e.g., Dynel, 2011), which yield valuable insight into the measures and countermeasures that senders and receivers employ. Our contribution lies in providing a computational account of how different strategies may be weighed in the balance and how precisely such strategies give rise to different behaviours and inferences. By casting people's beliefs about the conventions that govern communication as sampling assumptions, and formalising message production as

the computational inverse of comprehension, we can examine the trade-offs involved with greater precision. We demonstrate that a form of meta-inferential signalling affects the interplay between meta-inference and inference, and our results highlight the added complexity of meta-inference when the possibility of deception arises.

Before presenting our experimental work, we characterise the meta-inferential challenge that deception without lying represents and provide an overview of the theoretical basis for our analysis.

1.1. The liar's toolbox: a meta-inferential challenge

Imagine the following scenario:

> You are a graduate student, attending an academic conference for the first time. Nervous about your presentation the next morning, you have some wine at the conference dinner to help you relax. One thing leads to another and after a night of heavy drinking, you oversleep and miss your talk. Travelling home from the conference, you meet a colleague at the airport. She asks you how your talk went. The colleague is a potential future employer, so you are keen not to look foolish.

Assuming that you'd prefer not to reveal what really happened, how can you conceal the truth from her? There are three main strategies you might consider, each corresponding to different violations of the Gricean maxims:

Outright lying: One possibility is to proffer a blatant falsehood: *"My talk went really well! I was touched by the standing ovation."* By communicating facts which the sender knows false, outright lying represents a violation of the fundamental norm of communication. But as long this violation goes undetected, the receiver may leverage the assumption of cooperation implicit in the context and draw the desired incorrect conclusion. That said, lying is often fraught with difficulty. The liar may be uncertain about what the receiver already knows and it may be easier for the receiver to detect if new facts come to light. Outright lying is thus not necessarily the safest option, even for a completely amoral and self-interested communicator.

Being uninformative: To avoid outright lying, it may be preferable to say something irrelevant or otherwise uninformative: *"The conference dinner was fun."* Where no new information is disclosed the receiver's inference is seemingly restricted to her prior beliefs. But overtly flouting maxims of relevance and quantity in this way is likely to raise suspicion. Indeed, the blatant violation of Gricean norms is often deliberately used as a communicative strategy of its own.

Misleading: A third option is paltering: providing truthful but information with a misleading implication in mind: *"I was nervous beforehand, but the session was over before I knew it and there weren't any questions I couldn't handle."* There are considerable advantages to this strategy. Outright lies often bring harsh consequences when detected. Misleading implication which, by defintion, is not part of what is explicitly conveyed, offers a sense of plausible deniability (Pinker, Nowak & Lee, 2008; Lee & Pinker, 2010) and diminished repercussions (Schauer & Zeckhauser, 2009). Perhaps because of this, people may be less likely to view this form of deception as equivalent to lying (e.g., Coleman & Kay, 1981; Hardin, 2010), although this perception may change when one is on the receiving end (Rogers, Zeckhauser, Gino, Norton & Schweitzer, 2017). Importantly, because miselading involves being genuinely informative, norms of relevance and quantity are not overtly violated. This acts to limit suspicion and reduce the risk of detection while at the same time leveraging the receiver's presumption of cooperation. Thus, by selectively sampling facts in the right way, the sender may lead the receiver to a false conclusion as a result. Of course, this strategy carries risks of its own. For one thing, it requires the sender to accurately judge the conclusions that the receiver will draw. These inferences are likely to be determined in part by the receiver's assumptions and level of prior suspicion. Allowing for individual variation on the part of the receiver makes matters more complex. Misleading thus becomes a delicate balancing act: enough information must be disclosed to avoid or reduce suspicion, but not enough that the chance of inferring the truth increases.

What kind of option do people tend to choose in this kind of situation? In a preliminary study, we asked 96 first year psychology students (87 women) at the University of Leuven to imagine seven different scenarios like the one above. Participants selected a response from seven options consisting of two lies, two uninformative statements, two misleading statements, and the truth. Fig. 1 presents their preferences, collapsed across scenarios and equivalent response options.

Two important conclusions emerge from Figure 1. Firstly, people were uncomfortable with deception: 37% of responses involved telling the full truth and only 10% were outright lies: a surprising number perhaps given that each scenario provided a clear motivation to deceive. Secondly, among those who chose not to tell the truth, people showed a clear preference for misleading over lying or being uninformative (37%, 10% and 15% respectively). This finding is consistent with other work on the topic (Montague, Navarro, Perfors, Warner & Shafto, 2011; Rogers et al., 2017).

Why do people seem to prefer to actively mislead rather than be entirely uninformative? At first glance, it seems rational to be as uninformative as possible, because you are providing no information that the receiver can use to revise her beliefs at all. Effective misleading, on the other hand, involves salting your statements with a grain of truth. It thus runs a greater risk of the receiver inferring the real truth.

An important motivation for choosing a misleading utterance over a strictly uninformative one

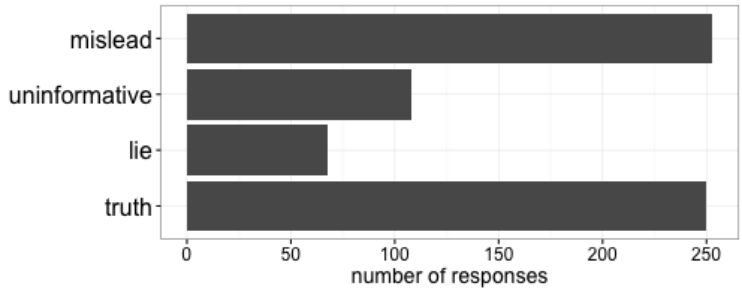


Figure 1: When choosing how to communicate in a variety of different scenarios with a clear motivation to deceive, people showed a strong preference to mislead rather than be uninformative. Telling an outright lie was the least preferred option.

is because the latter is suspicious. Consider the likely response of choosing to be uninformative in our earlier scenario:

Colleague: How did your talk go?
You: The conference dinner was fun.
Colleague: Talk didn't go so well?
You: The main conference room comfortably seats 400 people.
Colleague: That bad, huh? What happened?

Sperber, Clément, Heintz, Mascaro, Mercier, Origgi & Wilson (2010) propose that people have a toolbox of cognitive mechanisms for *epistemic vigilance* that reduces the risk of being deceived. The ability to track cooperation in others forms an integral part of such a defence. Whether through dedicated cognitive mechanisms or domain general capacities, obvious departures from communicative norms can be reliably detected by children as young as 3–6 years old (e.g., Eskritt, Whalen & Lee, 2008; Skarakis-Doyle, Izaryk, Campbell & Terry, 2014; Okanda, Asada, Moriguchi & Itakura, 2015). Responding in an uninformative way violates the principle of cooperation so blatantly that the deception is revealed.

A deceiver, sensitive to the epistemic vigilance of his counterpart may prefer instead to provide truthful but misleading utterances, a technique which may reduce or bypass such scrutiny altogether (Reboul, 2017). However, in so doing, he faces a delicate trade-off. Chosen well, such utterances may not only allay the receiver's suspicion, but by virtue of the inferential boost accorded to cooperative speakers, the receiver may be led to a false conclusion, terminating the search for further information (Bonawitz, Shafto, Gweon, Goodman, Spelke & Schulz, 2011; Montague et al., 2011). Yet if suspicion is already raised, the receiver is unlikely to fall for the false implicature and may use the information to get closer to the truth (Dynel, 2011).

This analysis points to two opposite forces, balanced in the selection of one strategy over an-

other. On one hand, the knowledge that the receiver may engage in inference about the helpfulness of the statement may lead the sender to opt for a misleading yet informative statement. On the other hand, if the sender considers that the receiver will be suspicious a priori, he may resort to being uninformative. In the following section we present a computational account of meta-inference which has the potential to capture this sort of reasoning.

1.2. Sampling assumptions as meta-inference

Consider a communication scenario where one person (the *receiver*) seeks to update her beliefs on the basis of information disclosed by another (the *sender*). The sender, for his part, selects information designed (according to his intention) to help or hinder the receiver in her efforts. We may characterise the reasoning of two such communicating parties as a form of Bayesian inference (following, for example, Shafto, Goodman & Griffiths, 2014; Goodman & Frank, 2016).

Turning first to the problem faced by the receiver: how should she update her beliefs based on the evidence provided by the sender? Let h denote one possible hypothesis that the receiver is currently considering, and $P(h)$ denote her belief in the hypothesis prior to receiving information from the sender. Then, having observed new information x (revealed by the sender) the receiver updates her belief according to:

$$P_{\text{RECEIVER}}(h \mid x) \propto P_{\text{SENDER}}(x \mid h) P(h), \tag{1}$$

where $P_{\text{SENDER}}(x \mid h)$ represents the assumption the receiver makes about the sender's sampling strategy (the way he chooses information to convey). The sender, in turn, is assumed to select information according to a sampling strategy targeted to the receiver:

$$P_{\text{SENDER}}(x \mid h) \propto (P_{\text{RECEIVER}}(h \mid x))^{\alpha} \tag{2}$$

where $P_{\text{RECEIVER}}(h \mid x)$ represents an assumption the sender makes about the belief update rule adopted by the receiver.

The goals of the sender are captured by the parameter α. A positive value for α corresponds to a sender who wishes to reveal the truth (that is, to increase the receiver's posterior belief in the correct hypothesis h); a negative value for α implies that the sender wishes to conceal the truth (by reducing the receiver's posterior belief). The magnitude of α indicates the degree to which the sender selects optimally: the larger the magnitude, the closer to optimal his selection becomes, the smaller the magnitude the more his choice resembles random selection. There are other ways to capture conflicting goals, like assigning separate utility functions for the sender and receiver with regard to truth-predicated action, but we chose this for its relative simplicity.

Using the model to describe a particular communication scenario requires Eqs. 1 and 2 to be considered simultaneously. Describing how the receiver updates her beliefs amounts to specifying the sampling strategy for the hypothetical sender that she thinks she is facing. Likewise, describing

the sender's sampling strategy requires stating an update rule for the hypothetical receiver that he considers. This may be a deeply recursive process, depending on the level of *"he thinks that she thinks that he thinks..."* reasoning that occurs. However deep the reasoning, we end up with a series of sender and receiver models nested under one another, starting at the top level with the model of the sender and receiver whose behaviour we wish to capture, progressing to the model the sender has of the receiver, and that the receiver has of the sender, and so on. Past empirical work (e.g., Colman, 2003; Vogel, Potts & Jurafsky, 2013; Stiller, Goodman & Frank, 2015; Franke & Degen, 2016) has suggested that recursive reasoning in this fashion may be limited in depth. We can avoid infinite nesting by specifying a sender with $\alpha = 0$. In this case there is no need for further nesting, since such a sender selects information at random without regard for the receiver. Likewise, any alternative update rule for the receiver that is effectively independent of the sender will also suffice as a ground term.

In any communication scenario there is a potential mismatch in the meta-inferential assumptions of the two parties. In pedagogical situations, where both parties have incentive to improve the effectiveness and efficiency of communication, such asymmetries may be of little consequence; similar qualitative patterns of inference emerge whether all assumptions are reciprocated or not. But when the goals of sender and receiver are at odds, qualitatively different patterns of reasoning may emerge depending on who is aware of the mismatch, who is aware of who is aware, and so on. By structuring a model as a series of nested sub-models, we can capture differing degrees of reciprocal awareness between sender and receiver regarding the sender's intent.

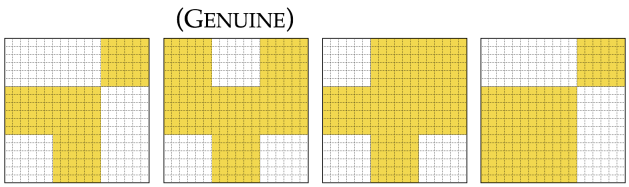
In this paper we use this computational framework to explain people's behaviour a pair of related experiments involving a simple "deception game" (see Fig. 2). In the first (Experiment 1), participants took the role of the receiver and were asked to infer the truth on the basis of potentially deceptive evidence. In the second (Experiment 2), people acted as the sender and were asked to provide evidence relevant to the hypotheses in question while at the same time preventing the truth from being discovered. We find that people's inferences and choices in this task are sensitive to the level of suspicion of the receiver. Moreover, qualitative individual differences in how people reason in the deception game correspond to different assumptions about intent and the data sampling process, as described by our computational framework.

2. Experiment 1: Reasoning from deceptive communications (receiver)

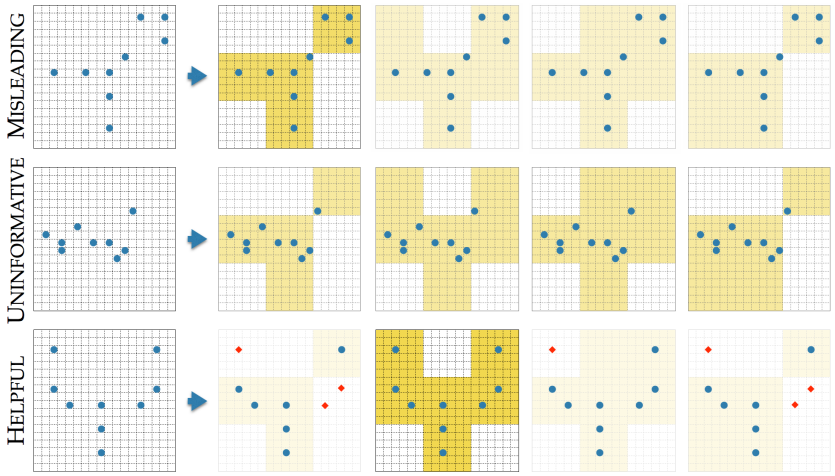
2.1. Method

2.1.1. Participants

We recruited 99 adults via Amazon Mechanical Turk, who were each paid $2.00USD for 5-10 minutes participation. One participant was excluded for browser incompatibility. The remaining 98 participants were 59% male and aged 19-64 (median age 29).



(a) Four alternative "maps" representing the common hypothesis space.



(b) Different patterns of evidence and the inferences they may licence.

Figure 2: The deception game. (a) People taking the role of the receiver (Experiment 1) and the sender (Experiment 2) see the same four "maps" in corresponding trials; the shaded area marks the region where treasure is buried. Only one of the four maps is genuine. (b) As sender, people seek to conceal the identity of the genuine map, but are nonetheless required to reveal some locations where treasure is actually buried (blue dots). They are given three options to choose from: representing *Misleading*, *Uninformative*, or *Helpful* evidence. As receiver, people attempt to infer the identity of the genuine map on the basis of the evidence provided, which varies in its potential to drive inference. The brightness of the shaded areas has been varied here to illustrate how plausible a trusting receiver might consider each map after viewing the evidence (brighter maps represent more plausible hypotheses and red dots indicate disconfirmatory evidence).

2.1.2. Procedure

A cover story informed people that they were taking part in an experiment simulating an online game based on data provided by past players of varying skill levels. People were told that they would take the role of an "explorer" (the receiver in our terminology) who must decide on a turn by turn basis which of four treasure maps is genuine based on evidence provided by a past player taking the turn of a "pirate" (the sender). The evidence consisted of points corresponding to a subset of locations drawn from the genuine map. Each point corresponded to a location where treasure was actually buried, but a sender could provide misleading or uninformative evidence through a strategic selection of points.

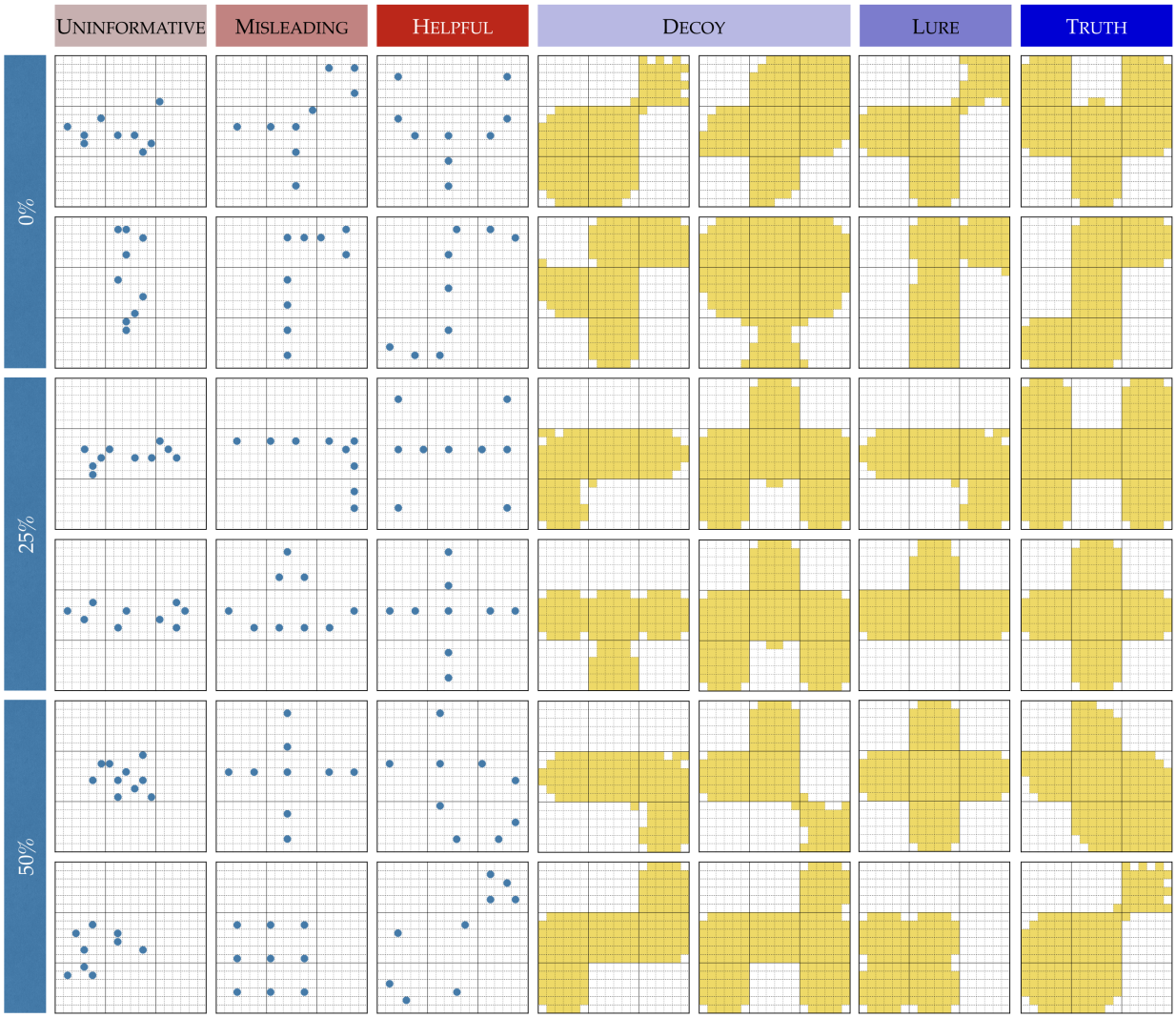


Figure 3: The experimental stimuli. Each trial in both experiments involved one of six sets of stimuli (rows), comprised of four maps (yellow regions) and three sets of evidence (blue dots). The task for the sender was to select one of the three sets of evidence to give to the receiver, while the task for the receiver was to select one of the four maps on the basis of the evidence they were given. The *Uninformative* evidence is consistent with all four corresponding maps. The *Helpful* evidence is consistent with only one map (the *Truth*). The *Misleading* evidence is designed to encourage a false conclusion (that the *Lure* map is genuine), but is also consistent with the *Truth*. The informativeness of the *Misleading* evidence was manipulated by controlling the number of *Decoy* maps with which it was consistent (row labels indicate the percentage of hypotheses (maps) ruled out by the *Misleading* evidence).

People's beliefs about the **sender's intent** in supplying the evidence was the basis of a within subjects manipulation. In the TEAMMATE condition, participants (as receivers) were told that the sender's goal had been to help a teammate identify the genuine map. In the OPPONENT condition the receivers were told that the goal of the speaker had been to keep its identity concealed.[1] Regardless of condition, participants knew that the sender could not provide false information. Thus, evidence could be relied upon to rule out a given map if any of the locations indicated did not overlap the shaded region shown on the map.

After the training session, people were shown a block of trials for the TEAMMATE condition and a block of trials for the OPPONENT condition. Within each block, participants saw each of the six map sets on three separate occasions: once in conjunction with the *Uninformative* evidence, once with *Misleading*, and once *Helpful* (see Figure 3). Thus each block consisted of 18 trials in all. The on-screen order of maps displayed in each trial, the trial order within each block, and the block order itself were all randomised. On each trial, people were required to consider the four maps and the evidence provided, and, taking into account whether the sender was a TEAMMATE or an OPPONENT, indicate which of the four maps they believed was most likely to be genuine.

2.1.3. Materials

The full set of experimental stimuli is shown in Fig. 3. Each set consisted of four maps and three pieces of associated evidence. The **quality** of the evidence was systematically varied from trial to trial. *Helpful* evidence constituted a pattern of locations that bore a close resemblance to the genuine map and ruled out three of the four alternatives. *Uninformative* evidence, in contrast, bore little similarity to any of the four maps, and could not be used to rule any out. *Misleading* evidence was designed to bear a strong resemblance to one of the three false maps. In addition, the **informativeness** of the *Misleading* evidence varied across the six sets of stimuli, ruling out either none, one or two of the *Decoy* maps, but never the genuine map (the *Truth*) or the map that it was designed to resemble (the *Lure*).

2.2. Basic results

Our first question of interest is whether people take the intention of the sender into account when interpreting the evidence offered. Fig. 4, which plots the responses of participants based on what they were told about the sender, suggests that they do indeed. To examine the strength of evidence for this finding we conducted a Bayesian multinomial logistic regression, comparing two mixed-effects models. In the EVIDENCE ONLY model, responses were predicted on the basis of the type of evidence presented in each trial. In the EVIDENCE + SUSPICION model, predictions

[1] To rule out consideration for player reputation, and to allow us to present each set of maps more than once, the instructions made it clear that people faced a different player on each trial. To reinforce this, the name and colour of the icon representing the pirate player was different for each trial.

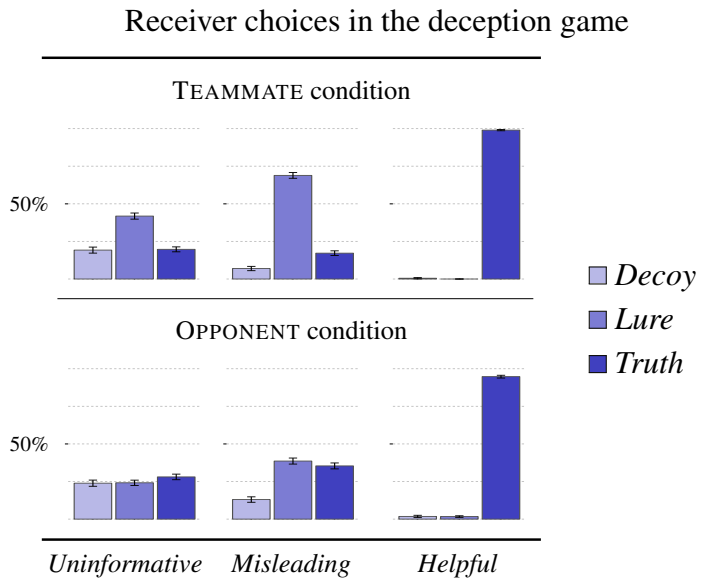


Figure 4: Proportion of participants selecting each response in Experiment 1. People playing the role of receiver were asked to identify which of four maps they believed to be genuine on the basis of the evidence provided. On each trial, people chose between two incorrect *Decoy* items, one *Lure* (a subset of the genuine map) and the *Truth*. In the TEAMMATE condition, people were told that the evidence had been provided by a helpful teammate; in the OPPONENT condition they were told that it had come from an opponent trying to conceal the truth. People correctly recognised that the *Helpful* evidence was consistent only with the *Truth*, and responded accordingly. When the evidence was *Misleading* (consistent with both the *Lure* and the *Truth*, but closest in size to the *Lure*) people were far more likely to choose the *Lure* in the TEAMMATE condition where there was reason to trust the sender. Likewise, when faced with *Uninformative* evidence (consistent with all four choices, but closest in size to the *Lure* in three out of six cases) people also displayed a preference for the *Lure* in the TEAMMATE condition. Error bars show standard error, and the proportion of responses favouring the *Decoy* items, is averaged over the two options.

also included an indicator of suspicion (based on condition). The analyses revealed strong evidence ($BF_{10} > 10^6$) in favour of the EVIDENCE + SUSPICION model over the EVIDENCE ONLY model, consistent with the notion that people reason differently depending on the context of communication.[2] When people thought the sender was trying to help them they reasoned beyond the immediate evidence, drawing strong (but mistaken) conclusions as a consequence. But when they thought the sender was trying to conceal the truth, people adopted a more conservative approach, appearing to select an option at random from amongst those not directly ruled out by the evidence.

As one might expect of meta-inferential reasoners, people interpreted the *Misleading* evidence differently depending on what they had been told about the sender. In the TEAMMATE condition,

[2]Both models included a random intercept for each individual, and were fit using the **brms** package (version 2.5.0) in R (version 3.4.3). Trials involving *Helpful* evidence were excluded because responses in favour of the *Truth* were at ceiling in both conditions.

people interpreted it as strong evidence in favour of the deceptive *Lure*, while in the OPPONENT condition they were much more cautious. The effect of suspicion in the face of *Misleading* evidence represents a five-fold reduction in relative rates of choosing the *Lure* over the *Truth* (95% CI: 3.6 to 6.7).[3] Yet even in the TEAMMATE condition more than a quarter of responses to *Misleading* evidence were in favour of items other than the *Lure*, raising the possibility that some people took different views of the evidence than others. We return to this issue in our model-based analyses.

A further curiosity is that people also interpreted *Uninformative* evidence differently, depending on what they believed about the intention of the sender — this, despite the fact that the data had no explicit evidentiary value. When *Uninformative* evidence was provided, the *Lure* map was chosen with greater frequency in the TEAMMATE condition, where cooperation was expected. While the impact of suspicion was reduced in this instance (when compared with the case for *Misleading* evidence), the effect nonetheless represents a three-fold reduction in relative rates of choosing the *Lure* over the *Truth* (95% CI: 2.1 to 4.1). Given that the deceptive *Lure* was the smallest hypothesis in size compatible with the *Uninformative* evidence in three out of six cases, and the smallest in size compatible with the *Misleading* evidence in all cases (see Fig. 3), a possible role for hypothesis size in the decision process suggests itself. We consider the nature of the meta-inferential assumptions that might account for this finding in our subsequent model-based analyses.

3. Experiment 2: Sending deceptive information

Experiment 1 demonstrated that people do take the likely intent of the sender into account when seeking to leverage the evidentiary value of information provided. Given that this is the case, do people account for this tendency in others in their own meta-inferential reasoning when they are acting as a sender? That is, do they seek to exploit the receiver's trust when they have it, and alter their strategy accordingly? In order to investigate these issues we invited people to play the deception game as a sender who was motivated to conceal the truth from their counterpart.

3.1. Method

3.1.1. Participants

We recruited 100 adults via Amazon Mechanical Turk, and paid them $1.25USD for 10-15 minutes minutes participation. Two of these participants were excluded for browser incompatibility. Data from a further 22 participants who failed to demonstrate a sufficient understanding of the experiment were excluded from subsequent analyses.[4] The remaining 76 participants were 46%

[3]This effect was quantified using our regression model extended to included an interaction between the type of evidence presented and level of suspicion.

[4]Participants were excluded if they failed to select the *Helpful* evidence on at least 40% of the CONTROL trials (where the goal was to help the other player), or if they chose the *Helpful* evidence in 40% or more LOW SUSPICION trials (where the goal was to hinder, and double bluffing was unreasonable).

female and aged 20-63 (median age 28.5).

3.1.2. Procedure

As in the first experiment, the cover story for the sender version informed people that they were taking part in an experiment based on an online game and that they would take the role of a pirate (the sender). On each trial, people were shown the genuine treasure map and three false maps, and were asked to select evidence to reveal to the explorer (the receiver).[5]

People's **sampling strategy** (deciding what evidence to disclose) was manipulated within subjects. In the CONTROL condition, the goal was to provide evidence that would help the receiver to correctly identify the genuine map. In both the TEAMMATE and OPPONENT conditions, the goal was to prevent the receiver from guessing correctly. Participants were told that the receiver was expecting evidence from a teammate (in the CONTROL and TEAMMATE conditions) or an opponent (in the OPPONENT condition). Participants were restricted in their choice of evidence to one of three options, namely: *Helpful*, *Misleading* or *Uninformative* evidence.

Experimental trials employed the same stimuli used in Experiment 1 (see Fig. 3). However, an additional four filler trials involved new stimuli, with evidence designed to reduce tactical responding; that is, whilst a seemingly random pattern of dots was characteristic of *Uninformative* evidence in the experimental trials, similar random patterns in the filler trials could be used to rule out one or more maps. Additionally, the least informative evidence in each of the filler trials was not a random pattern, but a pattern bearing a resemblance to one of the four maps. These filler trials were not analysed.

Participants undertook a training exercise similar to that used in the first experiment. In the test phase, people saw each of the ten map sets (six experimental and four fillers) three times (once per condition), making 30 trials in all. The on-screen order of maps, as well as the order of evidence items was randomised on a trial by trial basis. Trial order was also randomised, with trials from each of the three conditions randomly interleaved. The participant's goal in each trial (corresponding to the three conditions) was clearly stated via on-screen instructions. On each trial people were required to choose evidence from amongst the available options that best achieved the stated goal of the trial, taking into account the four maps shown and the identity of the genuine map.

3.2. Basic results

Experiment 1 established support for an intuitively reasonable notion: people take the sender's likely intent into account when determining the evidentiary value of information provided. The

[5]Once again, to avoid reputational concerns, people were told to assume that they were facing a different explorer on each trial, one who had not played the game before, and was unaware of the pirate's identity.

aim of the present experiment was to investigate whether people embody this intuition when motivated to deceive. Fig. 5 plots the proportion of people choosing to provide the different types of evidence to a hypothetical receiver, as a function of experimental condition. The figure suggests that, as expected, people select content to convey according to their goal and the context in which the content will be interpreted. A Bayesian multinomial logistic regression, comparing a model with condition as a predictor to an intercept only model, revealed strong evidence in support of this finding ($BF_{10} > 10^6$).[6] In the LOW SUSPICION condition, where the aim was to conceal the truth from a player expecting help from a teammate, participants made liberal use of the *Misleading* option. In contrast, when people faced an opponent in the HIGH SUSPICION condition, they overwhelmingly preferred to reveal as little as possible, selecting the *Uninformative* option in almost every case. Overall, the effect of suspicion on sender participants represents a three-fold reduction in relative rates of actively misleading rather than simply limiting disclosure (95% CI: 2.4 to 4.6).Unsurprisingly, in the CONTROL condition where the goal was to help, people were able to identify the evidence that the receiver would find most helpful, and almost always selected it.

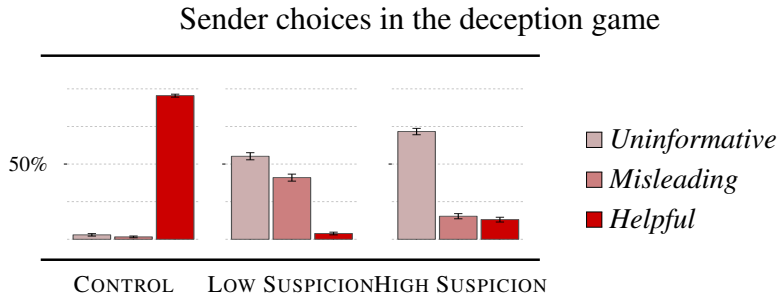


Figure 5: Proportion of participants selecting each response in Experiment 2. When acting as senders, people are only willing to provide the *Misleading* evidence when they believe the receiver does not suspect deception (left panel); when suspicion is high (middle panel) people overwhelmingly prefer the *Uninformative* alternative. In the control condition, people were asked to choose the option that would most benefit the receiver, and they did so consistently (right panel).

The pattern of behaviour across conditions — specifically, the change in the willingness to mislead — is largely what one would expect if people were reasoning about the inference of the receiver in a context sensitive manner. Despite this, the majority of senders in the LOW SUSPICION condition preferred to be uninformative. This seems contrary to the intuition that a meta-inferential reasoner expecting help should be reliably misled by misleading evidence, and the truth better

[6]Both models included a random intercept for each individual. CONTROL trials were excluded because responses in favour of the *Helpful* evidence were at ceiling.

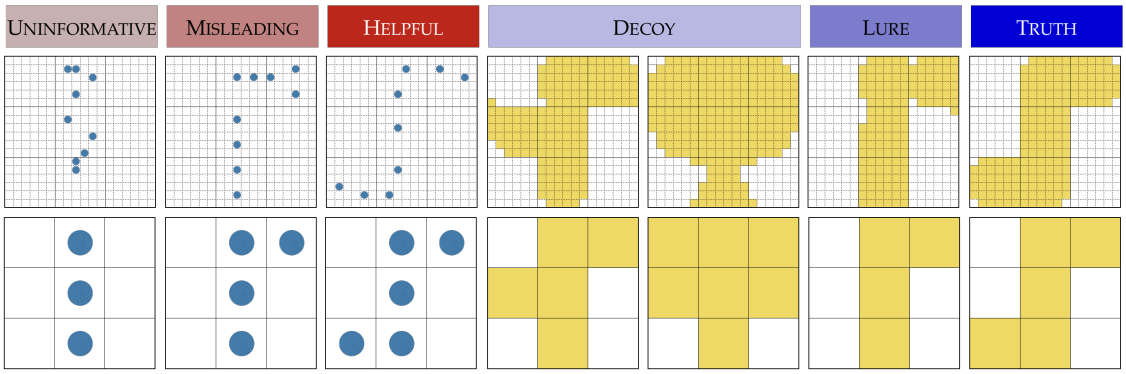


Figure 6: **Model Implementation.** We use a 3×3 grid approximation of the original stimuli to represent maps (hypothesis) and patterns of marked locations (evidence). A cell in the coarse grid representation (bottom row) is "on" if cells in the corresponding area of the original stimuli (top row) are "on".

concealed as a result. To better understand the assumptions that might drive a quantitative shift in sender preference, but not a qualitative reversal, we turn now to our computational analysis of the deception game.

4. Modelling meta-inference in the deception game

Our two experiments were designed to investigate how communicating reasoners might take account of the inferences of their interlocutor in situations where a variety of assumptions might reasonably hold. Taken together, the pattern of responses across both experiments appear largely consistent with an intuitively reasonable approach to meta-inferential reasoning. When receivers think that the sender can be trusted, they leverage that assumption to reason beyond the data; if the sender cannot be trusted no such leverage occurs. Senders, seemingly aware of this, are thus more willing to mislead trusting receivers than they are suspicious receivers. But if the veracity of data is not in question (because lying is not an option), then precisely what is it that mediates its strength as evidence? What are receivers (and consequently senders) sensitive to? Using the computational framework outlined at the start of the paper we can model various assumptions that might underpin this sensitivity, and ask which of these best captures the patterns of behaviour observed.

4.1. Model implementation

To model the deception game in a tractable way we use a simplified 3×3 grid to represent the experimental stimuli, as illustrated in Fig. 6. Thus both the hypothesis space $\mathcal{H}$ and the space $\mathcal{X}$ from which evidence is drawn, consist of the $2^9 = 512$ possible patterns of on/off grid cells. For any given trial, the hypothesis space is further restricted to one of the four maps in question

		Model			
Condition	Schema	WEAK	STRONG	ONE STEP	RECIPROCAL
(Receiver)					
TEAMMATE	$\langle Rec_T \rangle$	*Weak*	*Strong*	*Help* (*Weak*)	*Help* (...(*Weak*))
OPPONENT	$\langle Rec_O \rangle$	*Weak*	*Strong*	*Hinder* (*Weak*)	*Hinder* (...(*Weak*))
(Sender)					
CONTROL	*Help* ($\langle Rec_T \rangle$)	*Help* (*Weak*)	*Help* (*Strong*)	*Help* (*Help* (*Weak*))	*Help* (...(*Weak*))
LOW SUSP.	*Hinder* ($\langle Rec_T \rangle$)	*Hinder* (*Weak*)	*Hinder* (*Strong*)	*Hinder* (*Help* (*Weak*))	*Hinder* (*Help* (...(*Weak*)))
HIGH SUSP.	*Hinder* ($\langle Rec_O \rangle$)	*Hinder* (*Weak*)	*Hinder* (*Strong*)	*Hinder* (*Hinder* (*Weak*))	*Hinder* (...(*Weak*))

Table 1: **Four alternative models of sender and receiver behaviour in the deception game**. Receivers are assumed to reason according to Eq. (1) on the basis of the sampling assumption defined, and to respond in proportion to their strength of belief in each hypothesis. Senders are assumed to select evidence from amongst the options provided with probabilities defined according to Eq. (2). "*Help* ()" and "*Hinder* ()" denote opposite forms of intentional sampling where the selection of data is biased according to the sender's goal. "(...)" denotes a recursive and reciprocal assumption. A "*Weak*" sampling assumption means that evidence is used solely to disconfirm hypotheses, while a "*Strong*" assumption implies that data constitutes stronger evidence for smaller hypotheses, in accordance with the size principle. The schema column illustrates the common relationship amongst the sender and receiver assumptions within each model. See main text for further details.

by means of a trial-specific prior that rules out the remaining possibilities. No such restriction is placed on the evidence X, since people playing the role of the receiver were not aware of any restrictions regarding the selection of evidence, save for the fact that it was constrained to be truthful.

In the analyses that follows we consider the four models of deceptive communication summarised in Table 1. Our goal is use the models to investigate which set of assumptions best captures the behaviour observed across both experiments. Each model is actually a family of nested sub-models corresponding to the five experimental conditions (two receiver: TEAMMATE and OPPONENT, and three sender: CONTROL, LOW SUSPICION and HIGH SUSPICION). The assumptions of the trusting and suspicious receivers lie at the core of each model, so we turn to these first.

The first model we consider, the WEAK model, captures the notion that the receiver (whether trusting or suspicious) makes no assumption about the relative likelihood of an observation under each of the hypotheses in question. In terms of the computational framework, we capture this with a *weak sampling* assumption which defines the probability of an observation x in the event that hypothesis h holds, as

$$P(x|h) = \begin{cases} P(x) & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In other words, observations are used to rule out hypotheses that do not fit the evidence (i.e. where $x \notin h$), but the evidence is otherwise *uninformative* about the remaining hypotheses. The fact that a receiver adopts such an assumption, however, does not necessarily imply an absence of meta-inferential reasoning. A weak sampling assumption may capture the responses of a cautious receiver who is simply unwilling to impute any *particular* assumption on the part of the sender, perhaps in response to perceived variability in the reasoning style of others. Instead she may choose to rely only on the fact that the data itself was not false (consistent with the instructions given).

Yet in the context of the deception game, the receiver might reasonably justify a stronger assumption that leverages a perceived dependency between the evidence observed and the truth of the matter in question. The fact that only positive (and reliable) evidence may be provided constrains the sender in his choice. And importantly, the less that a given hypothesis entails (i.e. the fewer the observations compatible with it), the more the sender is constrained. According to the STRONG model, the receiver takes account of this by making a *strong sampling* assumption, where

$$P(x \mid h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Once again, such a sampling assumption need not indicate a lack of meta-inference on the part of the receiver. Rather, as long as the receiver is unwilling to assume that evidence selection is biased one way or another, then this size principle (Tenenbaum & Griffiths, 2001), which gives greater weight to smaller hypotheses, seems justified.

The next logical step in the progression of meta-inferential assumptions is for receivers to assume that senders also engage in meta-inference. A receiver who reasons in this way will expect the sender to bias selection in favour of evidence that is more informative (in the TEAMMATE condition) or less informative (in the OPPONENT condition). A single level of *"the receiver thinks that the sender thinks..."* reasoning may be modelled straightforwardly in the computational framework by a sequential instantiation of Eqs. 1 and 2. The receiver's assumption about how the sender intentionally biases the selection of evidence is captured by the α parameter: $\alpha = 1$ implies a bias in favour of more informative evidence, while $\alpha = -1$ implies the converse. For the receiver, such an intentional sampling assumption either increases or decreases the evidentiary weight that data would otherwise have under a more basic assumption, depending on the perceived intent of the sender.[7] This essentially asymmetric reasoning style, where the receiver attempts to reason one step further than the sender, forms the basis of the ONE STEP model.

[7] In the computational framework, the evidentiary weight of data ultimately stems from either its power to disconfirm hypothesis (weak sampling) or from the size principle (strong sampling). We use a weak sampling assumption as the ground term in the ONE STEP model.

The RECIPROCAL model, in contrast, describes the case where each reasoner credits the other with taking meta-inferential reasoning to its logical extreme. That is, for any assumption that the receiver makes about the sender, the sender reciprocates that assumption by assuming that she makes it, and vice versa – here there is no imbalance with regard to depth of reasoning. Notwithstanding the way that recursive and reciprocal reasoning proceeds, computationally speaking, it can only be satisfied by finding a meta-inferential equilibrium - a fixed point beyond which further recursive reasoning does not change the outcome (Shafto et al., 2014).[8]

Turning to the models of sender behaviour (see Table 1), each model represents an instance of Eq. (2) that defines the probability with which the sender selects evidence from amongst the options provided. The value of the parameter α is matched to the stated aim of each condition: in the CONTROL condition, where the goal was to help the receiver uncover the truth, we set $\alpha = 1$, while in the LOW SUSPICION and HIGH SUSPICION conditions, where the goal was to hinder, we set $\alpha = -1$. In addition to capturing the sender's intent, a sender model must define the sender's beliefs about the receiver's sampling assumption. For each of the models considered, the sender makes the same assumption about the receiver as the model itself does.[9] Each sender model thus proceeds straightforwardly from the corresponding receiver model. In the CONTROL and LOW SUSPICION conditions, the sender is assumed to model the would-be receiver in line with TEAMMATE model, while in the HIGH SUSPICION condition the sender is assumed to target the OPPONENT receiver.

4.2. Model-based analyses

We can now use the models we have defined to examine people's reasoning within the deception game. We are interested in whether people reasoned probabilistically about the generative process underlying communication within the game, and how this changed based on whether cooperation or competition was expected. Each model we have defined represents a different trade-off between the generality of the underlying assumptions and the degree to which inference is driven by those assumptions. So a comparison between model predictions and behavioural data allows us to assess how sensitive people were to the relative likelihood of evidence under one hypothesis over another.

For the receiver models, we compared predictions with responses (aggregated across all participants) to each of the 18 combinations of stimuli (six map sets and three types of evidence). Model fit, as measured by Root Mean Squared Error (RMSE), was assessed separately for each type of

[8] A single-state fixed point is not guaranteed — bi-stable equilibria may exist under reasonable assumptions, for example. However, for each of the four models a single-state fixed point exists.

[9] This need not be the case, we might wish to model a disconnect where the sender's assumption about the receiver does not match the model's direct assumption about the receiver. However, we constrain the models to be coherent in this way because it is both a theoretically reasonable and parsimonious starting point.

evidence as well as on an overall basis. Model predictions and fits to the choices of our receiver participants are shown in Fig. 7 for the TEAMMATE condition and in Fig. 8 for the OPPONENT condition. Because the *Helpful* evidence is incompatible with all but one hypotheses in every case, the receiver models predict that the receiver will identify the truth with complete certainty, fitting our behavioural data almost perfectly. For this reason, we omit those predictions from our plots, but include them in the calculation of overall fit.

In each of the models considered, the receiver makes a progressively stronger assumption about how the sender chooses what to reveal. When the receiver trusts the sender to cooperate, each additional assumption leads to progressively tighter conclusions. This cumulative ratcheting effect can be seen in Fig. 7. The figure shows that a receiver who adopts a weak sampling assumption is not easily misled. But one who believes that the sender is trying to help and that he reciprocates her assumptions, will leap to the wrong conclusion. Less intuitively, perhaps, this ratcheting effect applies even when the evidence is seemingly uninformative: information that would otherwise be ambiguous can still tighten conclusions, by virtue of the size principle. Indeed, comparing the predictions of the STRONG model to people's choices in response to *Uninformative* evidence suggests that the size principle was in effect.

This becomes important when considered from the perspective of the sender who wishes to conceal the truth. The sender's goal in this case (following directly from Eq. (2)) is to do what he can to reduce the receiver's belief in the true hypothesis (at least in relative terms). Certainly a misleading message has the potential to achieve this. But as we have seen, if the receiver's inference is consistent with the size principle then even a message that reveals no new information may act to reduce her belief in the true hypotheses. Thus, when considering these alternatives, the sender may conclude that the additional information disclosed by misleading yet informative evidence is not sufficiently offset. Fig. 9 (TEAMMATE condition) reveals that this is the case in the deception game. The figure plots the accuracy with which receivers identify the genuine map given the different types of evidence. It shows that, according to model predictions, the *Uninformative* evidence is always most effective at keeping the truth from the receiver (compromising accurate identification as a result). In the case of the WEAK model, this follows directly from what it means to be uninformative. For the remaining models, it follows from the size principle. As a direct consequence, the sender models for the LOW SUSPICION condition predict a preference for choosing the *Uninformative* option, as Fig. 10 (LOW SUSPICION condition) shows.

If being uninformative is an effective way of concealing the truth from a trusting receiver, consider then what inference a receiver in the OPPONENT condition should draw. As we have seen, a reasonable starting point for this receiver is to assume that the sender prefers to be uninformative. In our model, information becomes more informative with respect to a small hypothesis than to a large one, and hence less likely to be produced by an uncooperative sender. Yet the size principle dictates the reverse — namely smaller hypotheses consistent with the evidence are more likely.

Model fits to Receiver choices: TEAMMATE condition

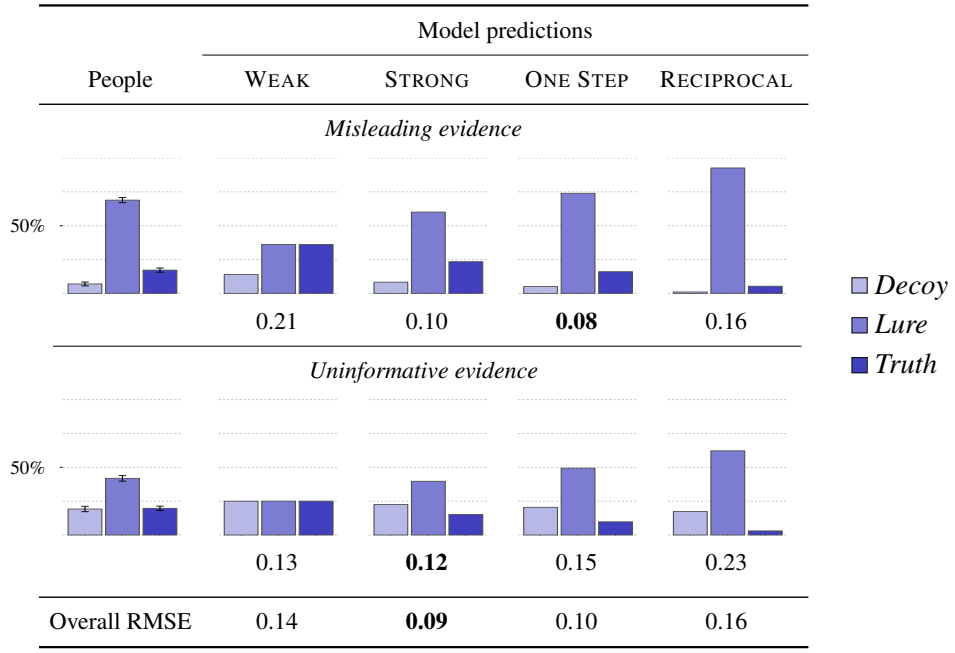


Figure 7: Predictions of four models compared with the choices of people playing the role of receiver in the TEAMMATE condition. The models are arranged in order, based on the strength and complexity of the assumptions involved. The WEAK model captures no constraints on the data, and represents a stance where the generative process is effectively ignored by the receiver. The STRONG model assumes only that the data represents positive evidence of the concept in question, and is otherwise unbiased by the sampling process. The ONE STEP model builds upon the STRONG model by assuming that the sender biases selection towards more informative content. The RECIPROCAL model assumes not only that the sender is trying to help in this way, but that both sender and receiver share a mutual awareness of each other's assumptions. The ratcheting effect of progressively layered assumptions can be seen in the top row: the more complex models increasingly favour the *Lure* item reflecting the fact that stronger assumptions licence stronger conclusions. The numbers below each graph show the model fits, as measured by Root Mean Squared Error (RMSE), with lower numbers indicating a better fit. The row at the bottom of each table shows the overall fit for each model in the given condition. While the STRONG model best captures the behaviour of participants in the TEAMMATE condition when evidence is *Uninformative*, when the evidence is *Misleading* it appears as though participants adopted a stronger assumption (although differences between the two are minor).

Model fits to Receiver choices: OPPONENT condition

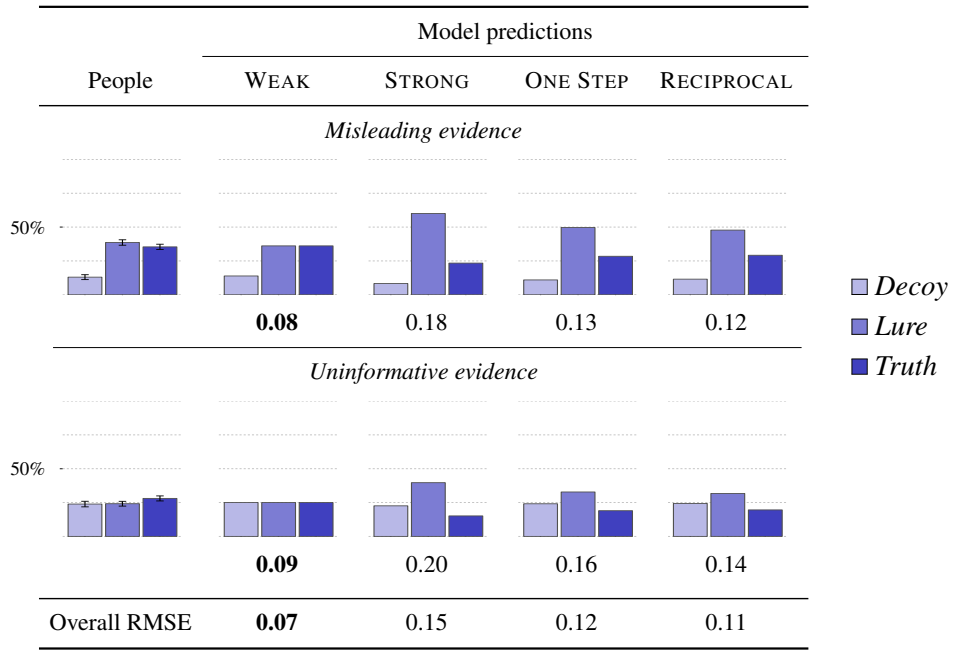
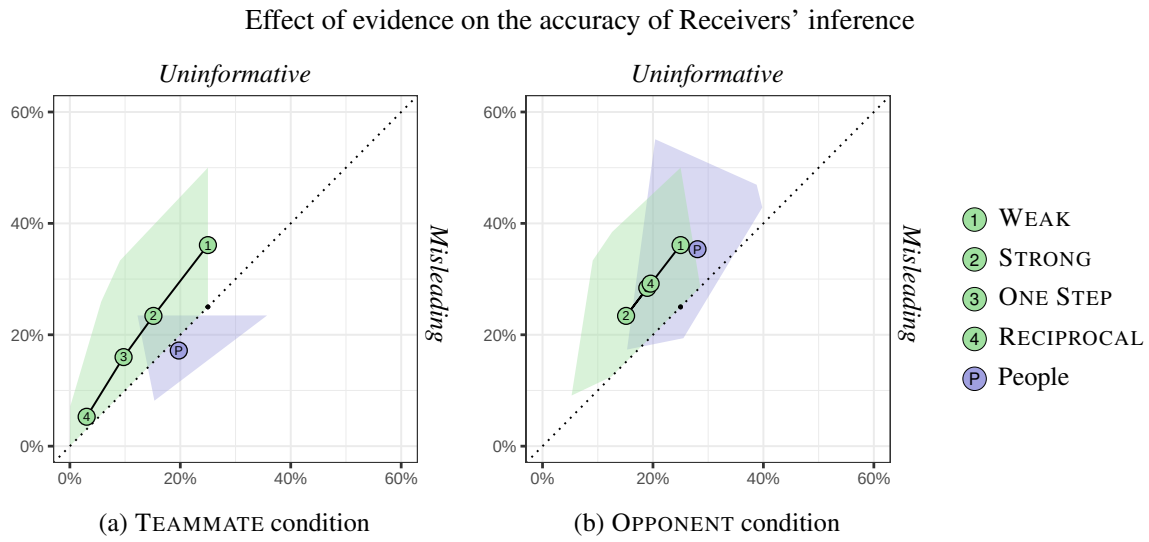


Figure 8: Predictions of four models compared with the choices of people playing the role of receiver in the OPPONENT condition. The WEAK and STRONG models, which are not context sensitive, make exactly the same predictions as described for the TEAMMATE condition. The ONE STEP and RECIPROCAL models are context sensitive however. In the OPPONENT condition, these models assume that the sender is trying to conceal the truth rather than reveal it ($\alpha = -1$). Their respective predictions reflect a trade-off between uninformativeness and the size principle, falling "between" the predictions of the WEAK and STRONG models. As described in the previous figure, lower RMSE values represent better model fits. In the OPPONENT condition it is the WEAK model, where the receiver assumes that the sender will be maximally uninformative (effectively disregarding the process by which the data is generated), that best captures people's behaviour in the OPPONENT condition.

Under the assumptions of the model, this leads the receiver to find a balance between two opposing forces. As a consequence, the inferences predicted by the ONE STEP model are less certain than those of the STRONG model but sharper than those of the WEAK model (see Fig. 8).

Somewhat paradoxically, as the receiver becomes less prepared to reason beyond the data, the sender pays a lower penalty for disclosing information. Thus, in contrast to the TEAMMATE models which predict that stronger assumptions lead to sharper conclusions, the OPPONENT models show no such pattern. Instead, progressively stronger assumptions produce predictions that follow the pattern of dampening oscillation shown in Fig. 8, and converge on the RECIPROCAL model. The predictions of the RECIPROCAL model, however, which represent an equilibrium



(a) TEAMMATE condition (b) OPPONENT condition

Figure 9: Receiver accuracy based on the type of evidence provided. The plotted points represent model predictions (green circles) and people's performance (blue circles) aggregated across the six sets of stimuli, while the polygons illustrate the spread of predictions – each vertex corresponds to a single set. (a) Model predictions in the TEAMMATE condition illustrate the effect of deceptive evidence on a trusting receiver. As people adopt stronger assumptions, their attempts to uncover the truth become increasingly inaccurate, leading to almost complete inaccuracy in the RECIPROCAL case. (b) In contrast, the OPPONENT models predict that stronger assumptions lead to little change in receiver accuracy. The plots illustrate the connection between the sender and receiver models. A sender who wishes to keep the truth from the receiver should choose the type of evidence that leads to the lowest accuracy. Regardless of the strength of the receiver's assumption, and whether or not they trust the sender, the model predictions indicate that the *Uninformative* evidence consistently leads to lower accuracy. This is in contrast to the observed accuracy of our receiver participants, who were least accurate in the TEAMMATE condition when presented with *Misleading* evidence.

where neither sender nor receiver "out thinks" the other, seem unintuitive. A more intuitive way for the receiver to take the sender's reasoning to its "logical" extreme, is to consider that he will display an optimal bias towards being uninformative (Hespanha, Ateskan & Kizilocak, 2000). As a consequence, the receiver should not attempt to reason beyond what the data falsifies - i.e. she should adopt a weak sampling assumption.[10] As Fig. 8 shows, the WEAK model best captures the behaviour of people in the OPPONENT condition. Whether we choose to model progressively stronger assumptions by an increasing bias towards the uninformative (larger negative α), or through increased depth of meta-inference, the predictions of the STRONG model, which assumes

[10] In terms of the computational model, taking the limit as $\alpha \to -\infty$ of Eq. (2), yields a likelihood function compatible with Eq. (3) – i.e. weak sampling.

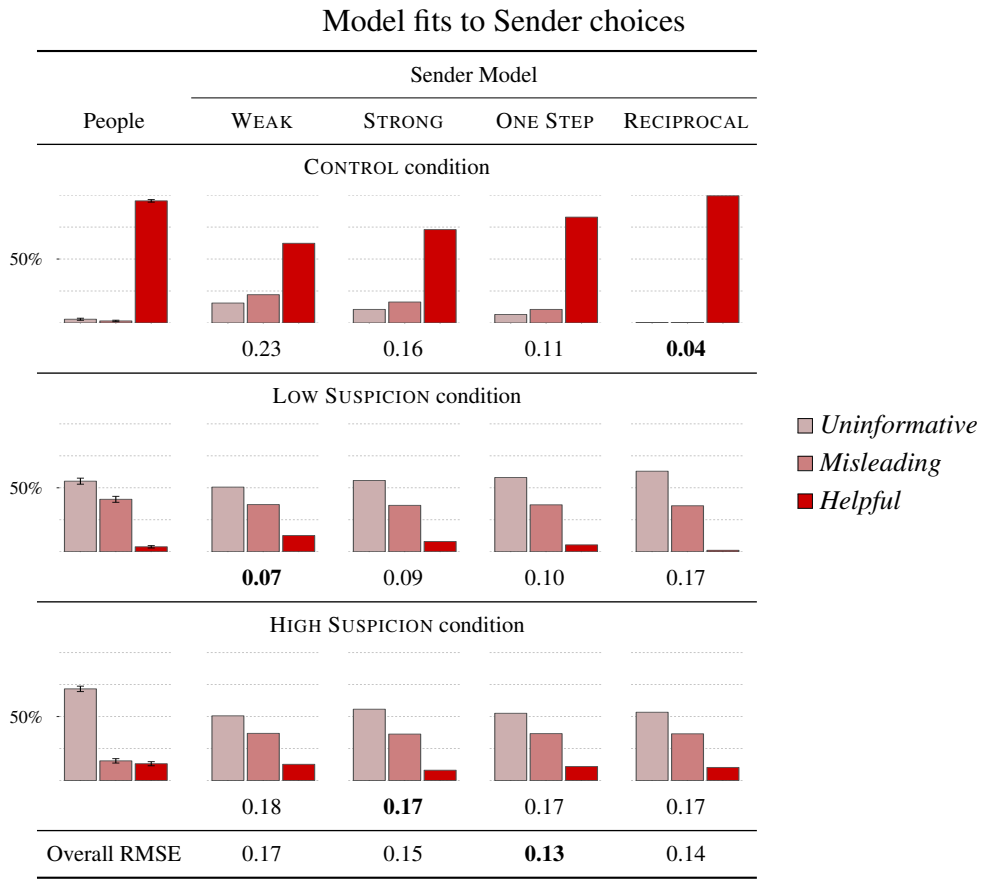


Figure 10: Predictions of four models compared with the choices of people playing the role of the Sender. In the CONTROL condition, where the sender's goal is to reveal the truth, all models predict a strong preference for the most informative (*Helpful*) message, as exhibited by participants. In the LOW SUSPICION and HIGH SUSPICION conditions however, the sender has the opposite goal – to hide the truth. The relative homogeneity of the predictions under these conditions reflects the unanimous prediction of the underlying receiver models that *Uninformative* evidence is most effective in this regard. Model fits (RMSE) are shown beneath each plot, and averaged across conditions in the bottom row of the table. Lower numbers represent better fits. While the models capture participants' overall preference for the *Uninformative* option in the LOW SUSPICION and HIGH SUSPICION conditions, the qualitative reduction in the use of the *Misleading* option is not predicted. (as the relatively poor fits in the HIGH SUSPICION condition indicate).

that data selection is unbiased, represent an upper bound on the strength of inference expected. Thus, once again, *Uninformative* evidence is most effective at concealing the truth from the receiver (see Fig. 9 – OPPONENT condition), and the sender models predict a preference for using it (see Fig. 10 – HIGH SUSPICION condition).

4.3. Discussion

Our analysis thus far has demonstrated that our framework captures two important properties of the receiver's inference. Firstly, it predicts the (obvious) effect of information content: trusting receivers draw stronger conclusions from more informative evidence. Secondly, and more importantly, it predicts an effect of assumption strength: stronger assumptions lead to stronger conclusions. While our analysis was not intended as a parameter fitting exercise, the good qualitative fits with theoretically motivated sampling assumptions suggests that the behaviour of participants in our receiver experiment is consistent with a sampling assumptions explanation. The strength of the assumption adopted depends on the perceived intent of the sender, as dictated by the setting in which communication takes place. In the TEAMMATE condition, people reasoned beyond the data, giving greater weight to those hypotheses under which the data might make sense ($\alpha > 0$). In the OPPONENT condition, where any attempt to reason beyond the data might be exploited, people behaved in line with a weak sampling assumption, using data only to falsify hypotheses ($\alpha \ll 0$).

When it comes to capturing the behaviour of our sender participants, however, the qualitative fits in Fig. 10 are less compelling. While the models matched response patterns in the LOW SUSPICION condition reasonably well, and captured people's overall preference for uninformative evidence in the HIGH SUSPICION condition, they failed to predict context sensitive meta-inference. They could not account for the disparity people showed between conditions as a function of suspicion. In the case of the ONE STEP and RECIPROCAL models, which were specified with context-specific assumptions in mind, this represents a challenge to the sampling assumptions account.

It is instructive at this point to recall the intuition behind the sender's decision. The virtue of a misleading utterance is that it appears (to a trusting receiver) to conform closely to communicative norms. Consequently, the intuition goes, the receiver will accord it a stronger inferential boost than they would a less informative utterance, promoting a strong yet misleading conclusion. But the ultimate goal for the sender (as we have framed it) is not to maximise the receiver's belief in one of the false hypotheses, but to minimise her belief in the true hypothesis. Under a weak sampling assumption ($\alpha \ll 0$), the evidence receives no inferential boost and so of course the sender should prefer to use uninformative evidence. Similarly the so-called strong sampling assumption ($\alpha = 0$) is not sufficiently strong as to warrant a change in preference. The intuition behind reciprocal and recursive meta-inference however, is that each layer of additional "he thinks, she thinks..." reasoning acts to increase the inferential boost that informative evidence receives relative to less informative evidence. Thus, a sufficiently strong (recursive) assumption on the part of the receiver should justify a reversal in the sender's preference so that he prefers to mislead. Yet what our analysis has shown is a limitation of our framework in this regard. As it stands, our framework fails to predict the necessary *interaction* between the information content of evidence and the strength of assumption in determining the strength (or rather weakness) of inference. At least not in the way that matters — the cumulative ratcheting effect of progressively stronger assumptions preserves

and never reverses the relative superiority of uninformative evidence in limiting receiver accuracy. Thus we have essentially demonstrated that two key (intuitively reasonable) meta-inferential assumptions — that trusting receivers make stronger assumptions (of a positive information bias) than suspicious ones, and that strong assumptions more strongly benefit informative content — are insufficient to explain sender behaviour in this situation. What additional or alternative assumptions might senders be making when deciding whether to conceal information or to actively mislead?

The receiver model fits shown in Fig. 7 reveal a potential clue. The figure shows that while the STRONG model provides a better fit to people's responses to the *Uninformative* evidence, the ONE STEP model provides a better account of the *Misleading* evidence. This suggests that receivers may make stronger assumptions on the basis of more informative evidence. This is an idea with some intuitive appeal; for example, if on hearing your words I believe you have chosen them carefully, I may be more likely to infer what you have implied. If receivers' assumptions are sensitive in this way, or the sender believes them to be, it should change the nature of the sender's evaluation. Instead of comparing what a receiver making a fixed assumption would infer from two alternative messages, the problem becomes one of comparing the alternatives under the different assumptions they would induce. For example, if senders assume that receivers reason beyond the data only when norms of relevance are upheld, then this might increase the incentive for the sender to mislead a trusting receiver. In the following section we extend our computational framework to accommodate these kind of "content-sensitive" sampling assumptions.

5. Modelling content-sensitive sampling assumptions

The computational models we have considered are based on a simple premise: namely, that people (as receivers) use information to rule out competing hypotheses and are therefore sensitive (as senders) to its evidentiary value when choosing information to convey. A given sampling assumption reflects a particular estimate about the degree to which evidence selection is biased in favour of the informative ($\alpha > 0$) or the uninformative ($\alpha < 0$). In the models we have examined, this estimate has been pre-determined solely on the basis of whether cooperation or competition is expected: that is, $a = +1$ (cooperation), or $\alpha = -1$ (competition). Although this approach has the virtue of simplicity, it fails to account for the *ostensive* nature of cooperative communication, in which the goal is not merely to produce utterances that are informative, but also ones that are easily recognised as such. In order to clarify how sampling assumptions might account for our experimental results, it makes sense to consider the notion of a receiver whose assumptions are sensitive to message content, and the implications for sender behaviour that this might have.

To see how a receiver might adjust her sampling assumption after observing the evidence provided, imagine that we (as an observer) already know which is the true hypothesis and are aware of the possibilities that the receiver is considering. If the aim of a cooperative sender is

Condition	Schema	Model
(Receiver)		
TEAMMATE	$\langle Rec_T \rangle$	$Ostensive \equiv \frac{1}{3}(Strong) + \frac{2}{3}(Help_{\supset}(\ldots(Weak)))$
OPPONENT	$\langle Rec_O \rangle$	$Hinder\ (Ostensive)$
(Sender)		
CONTROL	$Help\ (\langle Rec_T \rangle)$	$Help\ (Ostensive)$
LOW SUSPICION	$Hinder\ (\langle Rec_T \rangle)$	$Hinder\ (Ostensive)$
HIGH SUSPICION	$Hinder\ (\langle Rec_O \rangle)$	$Hinder\ (Hinder\ (Ostensive))$

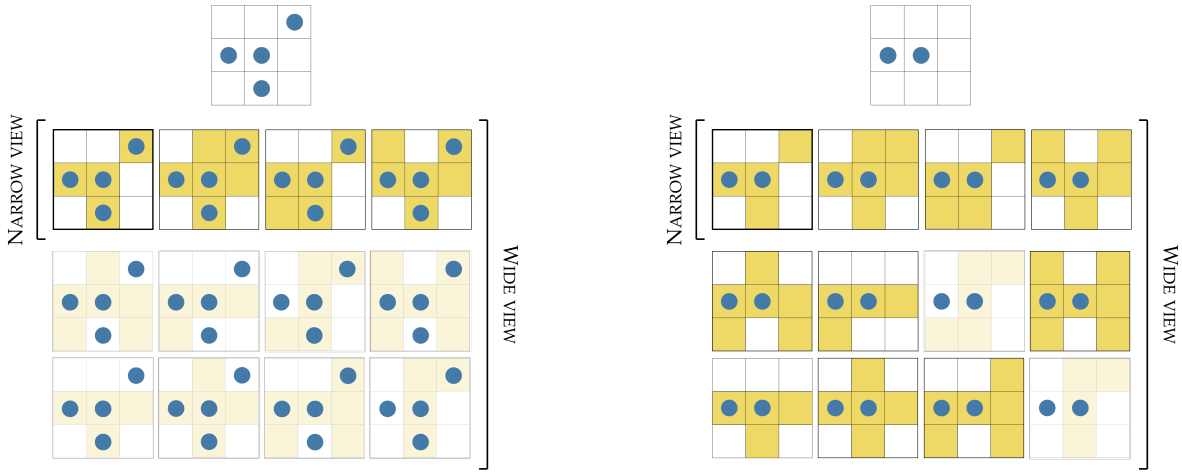
Table 2: **The OSTENSIVE model of sender and receiver behaviour**. The sender and receiver models for each condition follow the same model schema as the models introduced previously (see Table 1). The core *Ostensive* assumption corresponds to a reasoner who believes (with probability $p = \frac{1}{3}$) that the data is strongly sampled or (with probability $p = \frac{2}{3}$) that the data is helpfully sampled. The prior probabilities for the two assumptions were chosen to match the proportion of uninformative and informative stimuli used in the experiment (1:2). $Help_{\supset}(\ldots(Weak))$ denotes a recursive and reciprocal assumption, based on a general prior distribution (over a superset of the hypotheses currently under consideration). See main text for further details.

to select evidence that reduces the receiver's uncertainty about the matter at hand, then we can estimate his selection bias after seeing the evidence he selects, in the same sense that we might estimate the bias of a coin after seeing only a single toss. Of course, the receiver does not know the true hypothesis, but she may nonetheless form an estimate by considering all possibilities in order to determine the "likely helpfulness" of the information provided. We may model the receiver's assessment of the sender's likely helpfulness via the following straightforward extension of Eq. (1):

$$P_{\text{RECEIVER}}(h \,|\, x) \propto \sum_{s \in \mathcal{S}} P_{\text{SENDER}}(x \,|\, h, s) P(h) P(s) \tag{5}$$

where s represents an assumption that the receiver makes about the sender's sampling strategy, and $\mathcal{S}$ denotes the set of alternative strategies considered. As a simplifying assumption which should reasonably hold in the context of our experiments, we assume that the receiver considers the sender's sampling strategy to be independent of the true hypothesis.

If receivers are vigilant for ostensive signs of cooperation, then there are implications for the sender. In practical terms, the cooperative sender might select information to reduce the receiver's uncertainty not only about the hypotheses under consideration but also about the way in which the information was sampled. Because senders and receivers will not in general have perfect mutual information about each other's knowledge state, it makes sense for helpful senders to provide information that would be judged as being helpfully sampled, independent of any particular reciprocal assumption about prior knowledge. Intuitively, for example, the evidence sample shown in Fig. 11(a) feels more likely to have been provided by a competent and helpful sender than does



(a) Ostensively clear evidence

(b) Ambiguous evidence

Figure 11: **Examples of ostensively clear and ambiguous evidence.** In an example scenario drawn from the deception game, the receiver must use the evidence sample given (top) to distinguish amongst four hypotheses (narrow view) drawn from a larger set of possibilities (wide view). In interpreting the weight of a given piece of evidence, senders and receivers must take a stance regarding what constitues good evidence. Under a narrow view, evidence is informative only to the degree that it distinguishes amongst those hypotheses being directly considered. In this case both evidence samples are equivalent – each is equally ambiguous under a weak sampling assumption (because they rule out none of the four hypotheses in the narrow view), or equally helpful under a strong sampling assumption (identifying the target (dark border) due to the size principle). If instead, the receiver interprets informativity in the broader sense then the picture changes. In this case, evidence sample (a) is helpful in distinguishing the target from a considerably wide range of alternatives (pale yellow), whereas sample (b) is relatively unhelpful even in the wider context. Because the meaning of sample (a) is ostensively clear in this context it licences a stronger sampling assumption than sample (b), which remains relatively ambiguous.

the evidence sample shown in Fig. 11(b), despite the fact that each sample is equally ambiguous in the narrow sense. In terms of our model, a sender who wishes not only to be informative but also to be seen to be informative, selects information consistent with a strong selection bias ($\alpha \gg 0$, for example) under an appropriately general prior distribution. The deceptive sender may choose to mimic the helpful sender by making his informative intention clear, all the while selecting evidence intended to misinform. By being uninformative he may leave his sampling method (and his meaning) ambiguous.

To complete our content-sensitive model (which we shall refer to as the OSTENSIVE model), we need to specify the set of strategies $\mathcal{S}$ that the receiver considers in the typical course of cooperative communication. For simplicity we consider two strategies only, but in general we could integrate over any aspect of the model specification, such as the depth of recursion, the value of α and so on. To reflect the possibility that the evidence was selected by a helpful sender we adopt the sampling assumption from the RECIPROCAL model (see Table 1 – TEAMMATE condition). The alternative assumption that the receiver considers is that an indifferent sender selected the information at random ($\alpha = 0$), which is equivalent to a strong sampling assumption.This *Ostensive* sampling assumption is intended to describe the receiver's inference in the TEAMMATE condition. The sampling assumptions that complete the OSTENSIVE model are shown in Table 2.

In the next section, we apply this new model (in addition to the previous ones) to the data from Experiments 1 and 2. In order to investigate the presence of individual differences in reasoning, we focus on two sub-groups of participants that we identified in a *post hoc* fashion upon visual inspection of the data, as described below. Although this grouping is *post hoc* and the corresponding analysis should be taken with caution, we find that it (and the associated model fits) is revealing about the different kinds of reasoning that occur in deceptive communication.

5.1. Experimental results: Individual differences in content sensitivity

The results of our two experiments demonstrate that people's communicative inferences take into account the context in which communication takes place and whether cooperative norms can be taken for granted. Moreover, the data so far suggest that for receivers at least, people's reasoning is sensitive to context (suspicion level). However, context sensitive tailoring of deceptive strategy on the part of the sender is less evident. In order to investigate whether people are sensitive to the possibility that message content may signal the sender's intent, we now take a closer look at the response distributions of both receivers and senders.

Turning first to our receiver participants, upon visual examination of the data it appeared that there were two qualitatively distinct patterns of behaviour based on how people responded to the *Misleading* evidence in the TEAMMATE condition. As Fig. 12(a) reveals, the relevant response distribution is bi-modal. In addition, we used a Bayesian model to infer two independent binomial response rate parameters from the given response distribution. The model favours the same division that we identified by visual inspection. Further a Bayes' factor analysis revealed strong support for a model with two independent response rates over a model assuming only one ($BF_{10} > 1,000$).

We therefore defined, in a *post hoc* fashion, two qualitatively distinct groups. The Adaptive group, consisting of all participants who were consistently misled (choosing the *Lure* on five or six out of six relevant trials), appeared to be sensitive to the perceived intent of the sender and to adapt their assumptions accordingly. In contrast, the other participants, which we have labelled Conservative, appear to be largely *insensitive* to the sender's likely goal, displaying comparable conclusions in either condition.

The responses of receiver participants aggregated according to these groups are shown in Fig. 13. The Adaptive receivers drew stronger conclusions when evidence was *Uninformative* as well as *Misleading* in the TEAMMATE condition but showed a very different pattern in the OPPONENT condition, suggesting they were sensitive to the sender's intent. In contrast, the Conservative receivers responded similarly regardless of the nature of the sender or whether the evidence was *Misleading* or *Uninformative*.

We can apply a similar analysis to the sender data from Experiment 2. For the sender, the essential decision in each trial is whether to attempt to actively mislead the receiver or instead to just be as uninformative as possible. Where the sender stands in this regard should be influenced by their assumptions about the receiver – there is little point in revealing more than is necessary to a receiver who is unlikely to take the bait. We therefore divided people in two groups based on how frequently they chose to provide the *Misleading* evidence in the LOW SUSPICION condition. Although the relevant response distribution, shown in Fig. 12(b), is not as clearly bi-modal as was the case in our receiver analysis, the division into two groups is supported by the same Bayesian analysis used previously, applied in this case to the relevant sender response data ($BF_{10} > 1,000$). Thus, Adaptive senders are those who chose to mislead on three or more of the six relevant trials. In analogy with the receiver groups, the remaining participants comprise the Conservative group.

Sender choices for Adaptive and Conservative people are shown in Fig. 13(c) and (d) respectively. The figure shows that Adaptive senders, defined on the basis of their preference to actively mislead an unsuspecting receiver in the TEAMMATE condition, reverse this preference when the receiver is likely to be alert to the deception in the OPPONENT condition. In contrast, Conservative senders appear insensitive to the presence or absence of trust on the part of the receiver, strongly favouring the *Uninformative* option in both the LOW SUSPICION and HIGH SUSPICION conditions.

Taken together, the above analyses suggest that there may be a meaningful link between Adaptive senders and Adaptive receivers, and between Conservative senders and receivers as well. When Adaptive receivers believe that the sender can be trusted they are readily deceived by *Misleading* evidence. As Fig. 13 reveals, the proportion of Adaptive receivers who correctly infer the truth is lowest in this case. By favouring the use of *Misleading* evidence when facing a trusting receiver, Adaptive senders appear to target Adaptive receivers. However, Adaptive receivers appear to benefit from their strategy by being able to draw stronger conclusions when their inferences about

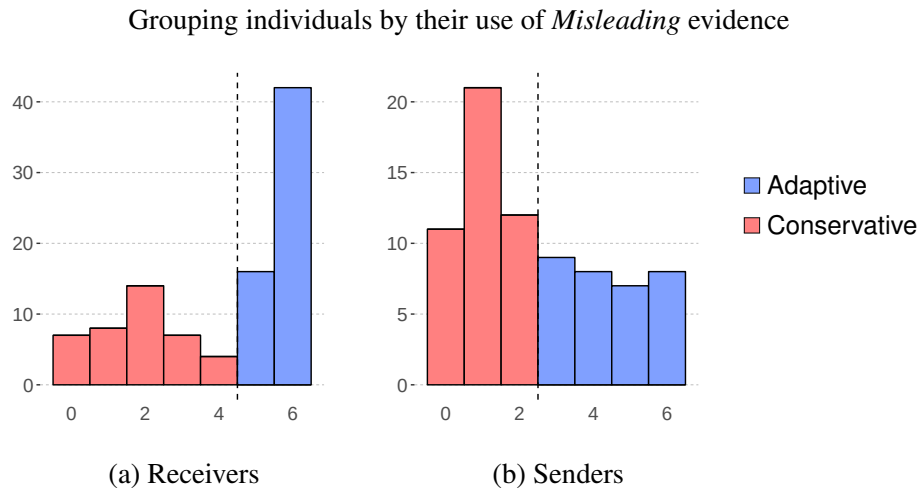


Figure 12: **Use of *Misleading* evidence by receivers and senders.** The histograms show the number of trials (out of six) that (a) receivers in the TEAMMATE condition chose the *Lure* item in response to *Misleading* evidence, and (b) senders in the LOW SUSPICION condition chose to provide the *Misleading* evidence. The vertical axis indicates the number of people responding with the frequency given on the horizontal axis. For a post hoc analysis, participants were separated into two groups on the basis of visual inspection of the response distributions. The dashed line separates Conservative participants (red bars) to the left, and Adaptive participants (blue bars) to the right. The receiver distribution is clearly bi-modal, while the sender distribution is less so. Nonetheless, a Bayesian analysis revealed strong evidence in favour of the partitioning illustrated (see main text for detail).

sender intent are correct. In contrast, the figure reveals that *Uninformative* evidence is most effective at concealing the truth from Conservative receivers, and that this strategy is the one favoured by Conservative senders.

To summarise, we have grouped our participants on the basis of how they reason about the effect of *Misleading* evidence in the TEAMMATE condition. In doing so, we have isolated those participants (the Adaptive ones) whose responses have driven the context sensitive behaviour we observed and modelled in aggregate in the first part of this paper. In what follows, we revisit our computational model to determine whether a sampling assumptions account can explain the behaviour of these two distinct groups.

5.2. Model-based analyses: Individual differences in content sensitivity

We now use the extended version of our model developed above in order to address two important questions that arose from the original analysis. Firstly, to what degree does the behaviour of our receiver participants indicate that they are adopting content-sensitive sampling assumptions? Specifically, do people appear to draw stronger conclusions (based on a stronger sampling assump-

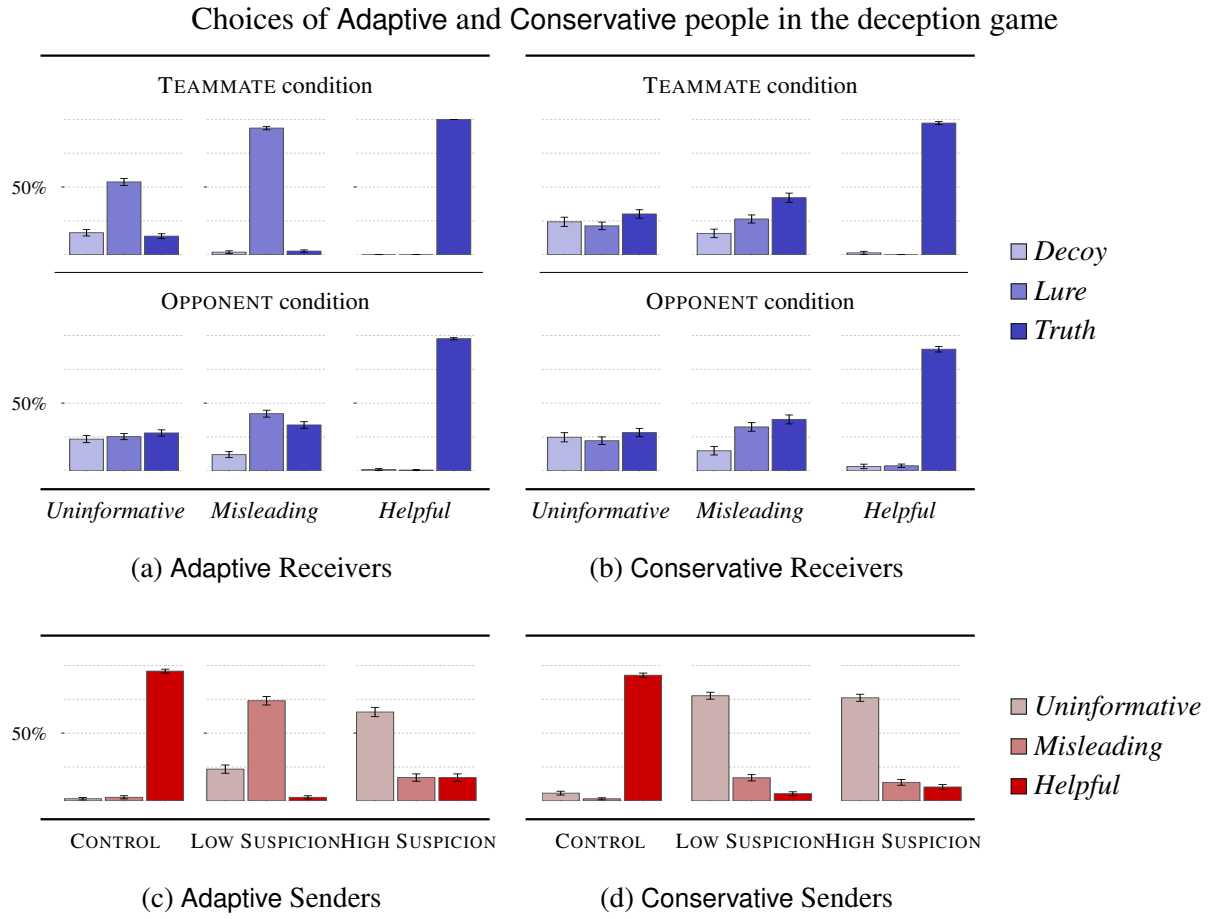


Figure 13: **Upper panel:** Receiver choices for two different types of participants: (a) Adaptive and (b) Conservative. Adaptive receivers (N=58) were defined as those more likely to select the *Lure* when faced with *Misleading* evidence in the TEAMMATE condition. Adaptive people also drew stronger conclusions from the seemingly uninformative evidence, but only when the sender's cooperation was expected. The difference in their inferences between the OPPONENT and TEAMMATE condition suggests that they were sensitive to the sender intent when deciding what conclusions to draw. In contrast, Conservative receivers (N=40) responded in the same manner regardless of whether the evidence was *Misleading* or *Uninformative*, as well as irrespective of the sender's intent. **Lower panel:** Sender choices for (c) Adaptive and (d) Conservative participants. Conservative senders (N=44) were defined as those more likely to favour *Uninformative* evidence in the LOW SUSPICION condition. This preference is reversed in favour of *Misleading* evidence for Adaptive senders (N=32). But while Conservative senders prefer to be uninformative without regard to receivers' suspicions, Adaptive senders adapt their strategy accordingly, providing *Misleading* evidence when the receiver is likely to be low in suspicion but *Uninformative* evidence when the receiver suspects them already.

tion) when presented with *Misleading* evidence compared to *Uninformative* evidence? Secondly, if the sender assumes that the receiver makes a content-sensitive sampling assumption, how does this impact his choices? Can this type of reasoning account for the pattern of deceptive behaviour observed in our sender experiment?

To address these questions, we compared model predictions of the OSTENSIVE model (as well as the four original models) to the choices of Adaptive and Conservative participants separately. Model predictions and associated fits are shown in Figs. 14, 15 and 17.[11] Because our group-level analysis indicated that the pattern of context sensitive behaviour is driven primarily by Adaptive participants, we focus our discussion on those participants first, returning subsequently to consider what sampling assumptions best account for Conservative participants.

5.2.1. Adaptive participants

Fig. 14 illustrates that for the Adaptive receivers, the OSTENSIVE model best captures their behaviour, suggesting that they are indeed drawing inferences about the way that the sender sampled the data based on how helpful the data appears to be. Because the *Misleading* evidence appears to be consistent with what might reasonably be chosen by a helpful sender, the model predicts that a trusting receiver will draw strong conclusions from it, in line with the predictions of the RECIPROCAL model. In contrast, the *Uninformative* evidence is inconsistent with helpful sampling and is therefore more likely to have been sampled at random. In this case, the OSTENSIVE model predicts weaker conclusions, more in line with the STRONG model. We find that this tendency to treat misleading and uninformative evidence in a qualitatively different way has the expected consequence: Adaptive receivers in the TEAMMATE condition were less likely to uncover the truth when given *Misleading* evidence (see Fig. 16(a)).

It is important to note that the predictions of the OSTENSIVE model were not reflected by our participants in the OPPONENT condition. Under the OSTENSIVE model, a suspicious receiver can discount the ostensive implication of the *Misleading* evidence, that way ruling out the *Lure* hypothesis and improving their chances of uncovering the truth (see the OSTENSIVE model prediction in Fig. 16(b)). Instead, for all participants, it seems more likely that they adopted a weak sampling assumption across the board. Nonetheless, the qualitative reversal predicted still offers a possible explanation of sender behaviour. If the sender does assume, as the OSTENSIVE model predicts, that misleading pays off when the receiver is trusting and backfires when she is suspicious, then a qualitative reversal of deceptive strategy as a function of receiver suspicion is justified. Indeed, as Fig. 17 shows, the OSTENSIVE model best fits the behaviour of Adaptive senders; it is the only one that predicts a significant change in sender behaviour between the LOW SUSPICION and HIGH SUSPICION conditions. It therefore captures the strong preference to mislead a trusting re-

[11] Predictions and fits were calculated for all models and conditions, but those not relevant to he present analyses have been dropped from the figures.

Model fits to choices of Adaptive and Conservative Receivers: TEAMMATE condition

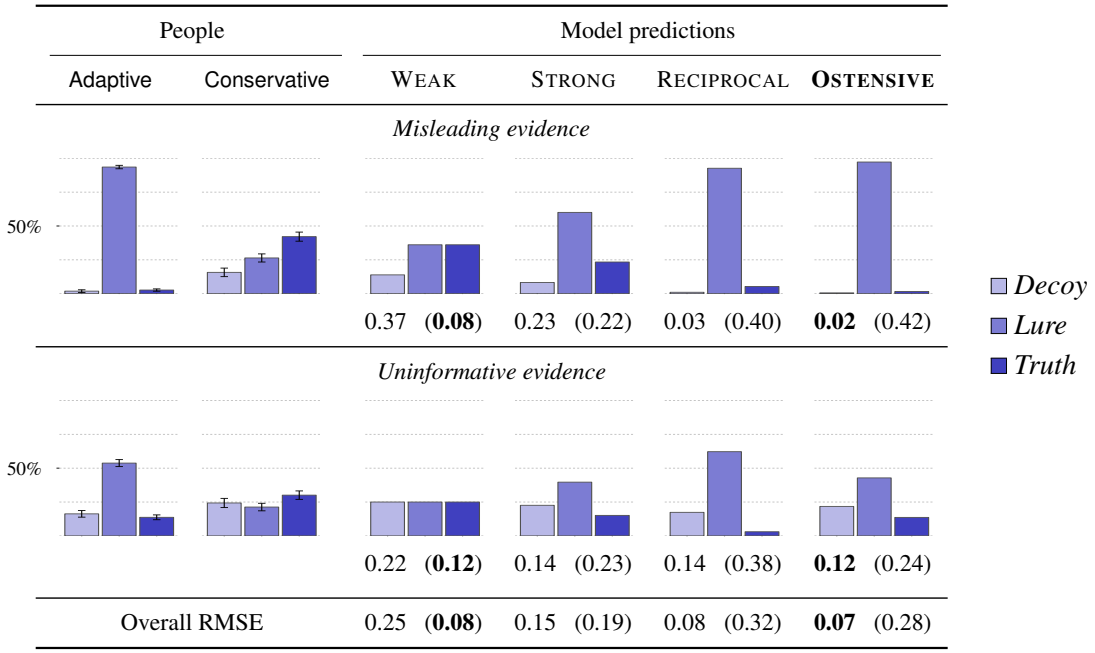


Figure 14: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants in the TEAMMATE condition. In the WEAK, STRONG, and RECIPROCAL models (described earlier), the way that the assumptions are arrived at in the first place, is left undefined. The OSTENSIVE model in contrast, describes the computational problem faced by the receiver as one of joint inference over sampling strategy and the hypotheses in question. Under this form of joint inference, certain scenarios are considered more likely than others: helpful (but misleading) content is more likely to have been helpfully selected, while uninformative content is more likely to have been selected randomly or without care. The closely matching predictions made by the OSTENSIVE and RECIPROCAL models (for *Misleading* content) and by the OSTENSIVE and STRONG models (for *Uninformative* content), follow as a consequence of the *content-sensitive* nature of the OSTENSIVE model. The numbers below each graph show the model fits for Adaptive and (Conservative) participants, as measured by RMSE. Once again, lower RMSE values represent better model fits. Adaptive receivers are best fit by the OSTENSIVE model, appearing to rely on a stronger assumption when given misleading evidence than when faced with something less informative. Conservative receivers in contrast, gain little leverage from their sampling assumptions irrespective of the content, and are best fit by the WEAK model.

Model fits to choices of Adaptive and Conservative Receivers: OPPONENT condition

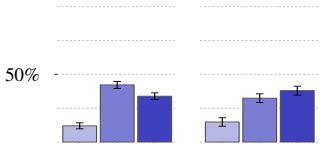
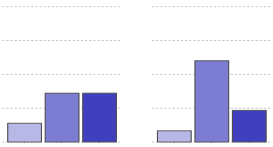
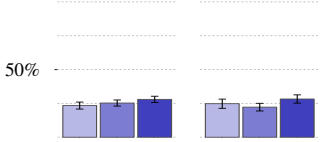
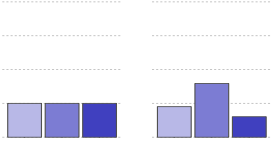
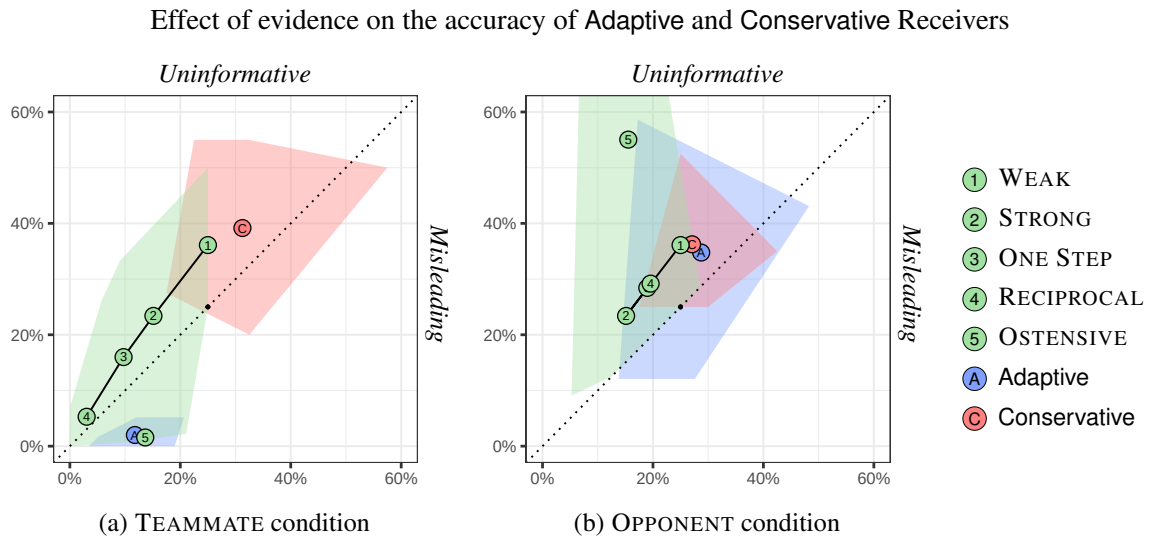
People		Model predictions			
Adaptive	Conservative	WEAK	STRONG	RECIPROCAL	**OSTENSIVE**
Misleading evidence					
					
		0.11 (**0.08**)	0.18 (0.20)	0.13 (0.13)	0.25 (0.23)
Uninformative evidence					
					
		0.10 (**0.10**)	0.20 (0.21)	0.15 (0.15)	0.17 (0.18)
Overall RMSE		**0.09** (**0.08**)	0.16 (0.17)	0.11 (0.12)	0.17 (0.17)

Figure 15: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants in the OPPONENT condition. The OSTENSIVE model is based on the intuition that message content may be informative both in the usual way and with regard to how it was sampled. Content that is informative and easily recognisable as such, the intuition goes, can be particularly misleading. The predictions of the model regarding *Misleading* evidence, show that a receiver alert to this form of deception, rather than be misled, could effectively leverage her suspicion to get closer to the truth. Yet, the plots clearly indicate that this is not what people did. Similarly, neither the STRONG nor the RECIPROCAL model represent a close match for either group of receivers, since both models embody a modest amount of meta-inferential leverage (due to the size principle), despite the receiver's suspicions. Only the WEAK model, which effectively discounts all evidence of a meta-inferential nature, provides a reasonable account of either group of participants. Model fits (RMSE) for Adaptive and (Conservative) participants are shown beneath each plot.



(a) TEAMMATE condition (b) OPPONENT condition

Figure 16: Accuracy of Adaptive and Conservative receivers based on the type of evidence provided. The plotted points represent model predictions (green circles) and people's performance (blue circles) aggregated across the six sets of stimuli, while the polygons illustrate the spread of predictions – each vertex corresponds to a single set. (a) Model predictions in the TEAMMATE condition highlight the qualitatively different predictions of the OSTENSIVE model which assumes that the receiver forms their sampling assumption based in part on the information content itself. Consequently, only the OSTENSIVE model predicts that *Misleading* evidence will have a greater negative impact on receiver accuracy than *Uninformative* evidece, and is able to capture the accuracy of Adaptive receivers as a result. (b) In the OPPONENT condition in contrast, the OSTENSIVE model predicts a backfire effect whereby the *Misleading* evidence improves rather than impairs the receiver's accuracy. Notably, this backfire effect did not occur. When faced with a potentially deceptive sender, both Adaptive and Conservative receivers favoured a literal interpretation of the evidence (in keeping with the "no lying" rule), as predicted by the WEAK model.

ceiver as well as the equally strong preference to be uninformative when the receiver is likely to be suspicious.

5.2.2. *Conservative participants*

Our Conservative receivers showed little difference in their behaviour between the TEAMMATE and OPPONENT conditions, preferring to avoid strong conclusions in both situations. Accordingly, we find that the WEAK model captures recipient behaviour consistently for both conditions and for both *Misleading* and *Uninformative* evidence (see Figs. 14 and 15).

What about the senders? Under a weak sampling assumption, the receiver uses evidence solely to disconfirm incompatible hypotheses. It logically follows then that the less information the sender reveals the less chance the receiver has of inferring the truth. But, as Fig. 16 shows, if the receiver adopts a weak sampling assumption, then the advantage (from the sender perspective)

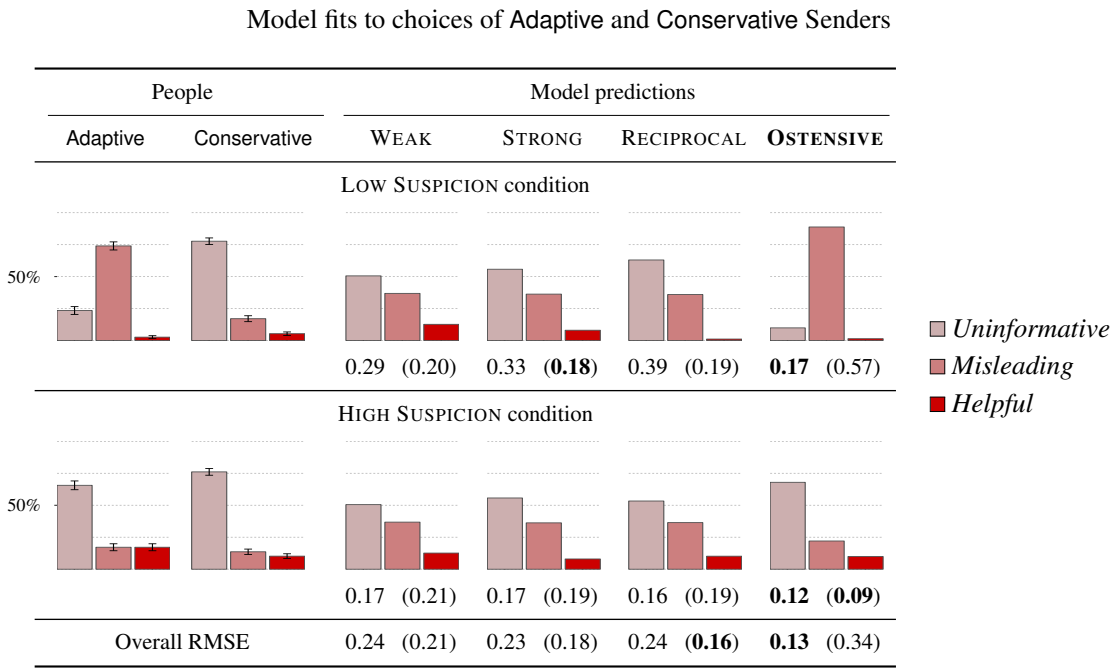


Figure 17: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants playing the role of the Sender. The OSTENSIVE model is based on the idea that data, in addition to being informative about some matter at hand, may also be informative regarding how it was selected in the first place. The remaining models describe meta-inference that is not sensitive to content in this way. Consequently, only the OSTENSIVE model is able to capture the strong preference for *Misleading* content exhibited by Adaptive people in the LOW SUSPICION condition. Furthermore, because the model predicts that attempts to actively mislead will backfire when the receiver is suspicious, it is the only model that captures the qualitative reversal of preference that Adaptive people show between conditions. Model fits (RMSE) are shown beneath each plot, and averaged across conditions in the bottom row of the table. Lower numbers represent better fits. The fits reveal that only the OSTENSIVE model, which heavily penalizes *Misleading* evidence when the receiver is suspicious, comes close to capturing the strength of people's preference for being uninformative in the HIGH SUSPICION condition. The fits also reveal that the behaviour of Conservative people in the LOW SUSPICION condition is not well explained by the models.

of offering *Uninformative* evidence over *Misleading* evidence is small. If senders choose their strategy according to this small relative difference, then we might expect to see senders exhibit a correspondingly small relative preference for *Uninformative* evidence. Yet, as the poor fits of the WEAK model for the senders indicate (see Fig. 17), this is not how senders behave. Rather, regardless of condition, Conservative senders show the same strong preference to be uninformative. This inclination to avoid the misleading option suggests that Conservative senders, like Adaptive

senders anticipating suspicious receivers, believe that attempts to mislead the receiver will backfire. The predictions of the OSTENSIVE model in the HIGH SUSPICION condition which capture this "backfire" concept provide the best (and only reasonable) fit to the behaviour of conservative participants. Because Conservative senders responded similarly in the LOW SUSPICION and HIGH SUSPICION conditions, we also assessed the degree to which the *Ostensive* assumption from the HIGH SUSPICION condition captured behaviour in the LOW SUSPICION condition. This yielded a better fit than the next best fitting assumption (*Strong*: 0.18).

5.3. Discussion

The goal of the present study was to examine how people reasoned about evidence in situations where deception was possible but lying was not an option. People played the deception game in two related experiments: both as "receivers" and "senders" of messages (evidence). When viewed as an homogenous sample, people as receivers were sensitive to the context in which communication took place. They drew strong conclusions from evidence, but only when they thought the sender could be trusted. But evidence for context sensitive behaviour amongst senders was weaker overall. Though people were more willing to mislead when they thought deception was not expected, they favoured uninformative evidence regardless of context. Using a computational framework to predict responses on the basis of meta-inferential sampling assumptions, our original analysis found that while the behaviour of receivers as a whole was consistent with a sampling assumptions account, the behaviour of senders in the aggregate was not so easily accounted for by standard sampling assumption discussed in the literature.

Our more detailed analysis was thus prompted by two issues. Firstly, the surprising result that sender behaviour appeared to be somewhat insensitive to context. And secondly, the fact that our original model which builds on standard sampling assumptions cannot account for even the modest change in sender strategy that was observed. With respect to the first issue, subsequent analyses of participant's choices revealed a plausible explanation. For both senders and receivers there appear to be two qualitatively distinct groups of people: Adaptive participants who tailored their behaviour according to context, and Conservative participants who maintained a single consistent approach. Adaptive receivers reasoned well beyond a literal interpretation of the evidence, but only when the sender's cooperation was implied. Adaptive senders demonstrated a clear reversal in preference for misleading evidence between conditions.

To address the limitations of our original model, we explored the predictions of the OSTENSIVE model. Table 3 summarises the findings of our original and revised analyses. The analysis indicates that Adaptive participants acted consistent with an ostensive-inferential view of communication. When the context dictated that cooperative norms should apply, full cooperation was not taken for granted. Instead, Adaptive receivers leveraged ostensive signs of helpfulness. Adaptive senders overwhelmingly preferred to provide ostensively helpful yet misleading evidence when seeking to mislead.

	All People		Adaptive People		Conservative People	
Condition	Assumption	Fit	Assumption	Fit	Assumption	Fit
(Receiver)						
TEAMMATE	*Strong*	0.09	*Ostensive*	0.07	*Weak*	0.08
OPPONENT	*Weak*	0.07	*Weak*	0.09	*Weak*	0.08
(Sender)						
LOW SUSPICION	*Weak*	0.07	*Hinder* (*Ostensive*)	0.17	*Hinder* (*Hinder* (*Ostensive*))	0.09
HIGH SUSPICION	*Strong*	0.17	*Hinder* (*Hinder* (*Ostensive*))	0.12	*Hinder* (*Hinder* (*Ostensive*))	0.09

Table 3: **Best fitting models of sender and receiver behaviour in the deception game**. See tables 1 and 2 and main text for model descriptions. For model fits (RMSE), lower numbers represent better fits. Overall the fits indicated by our *post hoc* group-level analysis suggest a more nuanced picture than the aggregate analysis revealed. In particular, the revised analyses suggests a role for the kind of content-sensitive sampling assumption captured by the OSTENSIVE model. Further it suggests that Conservative senders and receivers may have adopted a form of "worst case" assumption whether or not suspicion was warranted.

In contrast, when the context suggests that cooperative norms may not apply, the analysis indicates a disconnect between sender and receiver assumptions. Anticipating that ostensive signals can backfire when the receiver is suspicious, Adaptive senders declined to offer them. Yet absent trust, Adaptive receivers ignored ostensive signals and took a literal view of data. Despite the disconnect, both sets of assumptions represent sensible defensive positions. For receivers, a weak sampling assumption means that they are immune to strategic exploitation. For senders, the extra caution is largely without cost (unless there are a sufficiently large population of receivers who are trusting despite obvious cause for suspicion). The concept of a "defensive" assumption may explain the behaviour of Conservative participants. Our analysis revealed that the best fitting assumption for both senders and receivers in this group matched the assumption of Adaptive participants in an adversarial context. We comment further on the Conservative stance, and other broader issues in the following general discussion.

Despite the limitations of our group-level analyses, due to its post hoc nature and the limited amount of data collected per individual, we find that the revised analysis gives a more compelling account of behaviour overall, and for senders in particular. Importantly, by isolating the group of participants responsible for driving context-sensitive behaviour, and introducing the OSTENSIVE model to capture content-sensitive sampling assumptions, we have better captured the way that qualitative patterns of responding were driven by both content and context.

6. General Discussion

Using a non-verbal communication game where deception was motivated but outright lying was not an option, we investigated how the spectre of deception changes the way that people reason about evidence. Across two experiments, in both production and comprehension tasks, we found that people's behaviour was guided by their inferences about how others reason from evidence. When selecting evidence to provide, people reasoned about the way that suspicion affects the comprehension process. And when interpreting evidence provided, people considered the ways that evidence may be used deceptively. Support for this conclusion in its most basic sense can be seen in the context sensitive pattern of responses we observed in both experiments.

On the comprehension side, people were sensitive to the (presumed) intent of the sender. They drew strong conclusions from evidence when the context dictated that it made sense to trust the sender, but reached guarded conclusions otherwise. This behaviour is consistent with the findings of comparable studies investigating pragmatic implicature in non-cooperative contexts. For example, using a picture selection task Pryslopska (2013) found that the pragmatic interpretation of "some" as meaning "some but not all" was more likely when the context emphasised cooperation over competition. In a similar vein, Dulcinatti (2018) demonstrated (via a picture selection task, as well as a task involving purely verbal reasoning) that a range of scalar and ad hoc implicatures drove people's conclusions in cooperative but not competitive scenarios.

On the production side, people's selective presentation of evidence was also sensitive to their communicative goal, even without recourse to outright lying. While comparable studies are somewhat rare, people's ability to selectively employ seemingly helpful yet ultimately misleading information has been demonstrated in verbal reasoning tasks in adults (Franke, Dulcinati & Pouscoulous, 2019) and in concept teaching tasks in children as young as 4 years old (Rhodes, Bonawitz, Shafto, Chen & Caglar, 2015). Our experiment extends these results by using the same experimental task to examine how people adjust their strategy in the face of low and high suspicion. When people believed their intentions would not be viewed with suspicion many preferred to make misleading implications, but when their motive to deceive was made plain in context, people strongly preferred to give nothing away.

6.1. Context-sensitive and content-sensitive sampling assumptions

Our empirical results established the basis for our computational analyses of people's assumptions which reveal useful insights into how people think that others reason from evidence. Our modelling suggests that people (both as senders and receivers) reasoned probabilistically about the generative process underlying communication within the game, and interpreted evidence flexibly in light of those assumptions. The changing nature and strength of people's assumptions brought about by the cooperative or competitive context drove the corresponding change in responding observed. Our findings replicate and extend core findings in the sampling assumptions

literature. From our analysis of receiver behaviour, we find evidence of the size principle in operation. Following this principle people tended to generalise from evidence to the smallest compatible hypothesis even when that evidence was otherwise uninformative. This principle, and assumptions which build upon it have been shown to shape inductive reasoning in a variety of tasks including learning abstract concepts (Tenenbaum, 2000), word learning (Xu & Tenenbaum, 2007), category learning (Hendrickson, Perfors, Navarro & Ransom, 2019), property induction (Sanjana & Tenenbaum, 2003; Fernbach, 2006), and similarity judgments (Tenenbaum & Griffiths, 2001; Navarro & Perfors, 2010). Receiver behaviour in the OPPONENT condition was well captured by a weak sampling assumption. Frequently, this assumption has been used in the literature to model aleatory uncertainty in the generative process – when observations are sampled at random and independently of the concept of interest, for example (Kemp & Tenenbaum, 2009; Heit, 1998; Shepard, 1987). Our results highlight that it applies in situations of epistemic uncertainty also, even when observations are clearly restricted to the true concept.

Despite the fact that trusting receivers reasoned beyond the evidence even when it was seemingly uninformative, our analysis suggests that people may have adjusted their sampling assumptions based on the data observed. Although our evidence in support of this is modest at best, it is nonetheless consistent with previous findings that people's sampling assumptions may be shaped by the data (Hendrickson et al., 2019; Ransom, Perfors & Navarro, 2016) and that people perform joint inference over the knowledge and intent of their informant and the truth of the matter at hand (e.g., Gweon, Tenenbaum & Schulz, 2010; Shafto, Eaves, Navarro & Perfors, 2012a; Goodman & Frank, 2016). The data from the OPPONENT condition is less ambiguous. There are no signs that suspicious receivers were drawn into content-based second guessing of strategy whether the given evidence appeared purposefully or haphazardly sampled. Any joint inference (and the function of epistemic vigilance that it supports) was effectively suspended given the high prior probability that the sender was uncooperative. Our computational model cannot speak directly to the question of whether joint inference regarding sender intent is *actually suspended* in this case, or whether such inferences are drawn and over-ruled. Nonetheless, such questions are an interesting avenue for future investigation. This issue potentially connects with a debate in the pragmatics literature regarding whether implicatures are drawn as the context demands (e.g., Russell, 2006) or are always computed by default but sometimes discarded (e.g., Levinson, 2000).

Our analysis of the problem facing the would-be deceptive sender reveals that what may seem like an obvious heuristic – mislead the trusting, conceal from the suspicious – is not so readily justified. Setting aside the fact that the apparently obvious intuition was shared by only half our participants, our simulations revealed two important disconnects with standard (content-insensitive) sampling assumptions like strong and weak sampling. The first of these is when the receiver is not suspicious. Like the rising tide that lifts all boats, a content-insensitive sampling assumption based on the size principle supports reasoning beyond the data no matter what the data. Under such as-

sumptions, the mathematics of Bayesian inference suggests that however misleading a given piece of evidence may appear, a subset of that same evidence is always a better option.[12] When the receiver instead believes that the sender's goal is opposed to her own, her best bet is to ignore those aspects of the signal that the sender controls (e.g., Hespanha et al., 2000). In the deception game, this means adopting a weak (uninformative) sampling assumption – a tactic overwhelmingly followed by our receiver participants. Thus the second disconnect is that senders did not appear to behave in line with the assumption that the receiver would ignore all unreliable aspects of the evidence. Instead, senders showed a bias against misleading evidence to an extent not justified by the information penalty alone.[13]

Our analysis offers a plausible (albeit speculative) explanation for the senders' bias: that is, senders assumed that receivers would reason further beyond some data than others. Whether people arrived at this assumption through some form of mental simulation or via an intuitive theory derived from experience, the fact that people assumed that receivers would act in this way complements the modest evidence from our receiver experiment that this is the case. Alternatively, people might simply be mistaken – in itself this would represent an intersting disconnect between production and comprehension that would be worth pursuing. Regardless, to the best of our knowledge, our finding that such an assumption is operating on the production side is a novel one, and one with interesting implications if it can be replicated.

6.2. Ostensive meta-inference

If people do have an (implicit) awareness that comprehension may be affected by content-sensitive sampling assumptions, it is interesting to consider whether and how this effects communication on the production side. For instance, do senders attempt to increase the chances that a particular sampling assumption will be adopted by their counterpart by signalling it in some way? The use of ostensive signals such as eye-gaze, pointing and tone modulation have been shown to play an important role in infant learning. Such signals help the infant to understand that they are being addressed, to make clear the referent when teaching object labels, and even to indicate that

[12] This is certainly the case in the deception game where only positive (and truthful) evidence is allowed, and applies regardless of the strength of the informativity bias or the number of recursive layers of "he thinks, she thinks reasoning". Additionally, initial simulations show that this may be a robust result that applies to any discrete likelihood function obeying reasonable constraints related to the size principle: namely, any given observation should be more (or equally) likely under the smaller of any two hypotheses with which it is compatible; and for any two observations compatible with the same hypothesis, the one that is compatible with fewer alternatives should be more (or equally) likely. A formal proof and further investigation of the generality of this property are an area for future work.

[13] Under the (somewhat standard) assumption that $\alpha = 1$, as used in our model. One might alternatively account for the bias observed in the HIGH SUSPICION condition by using a higher value to reflect more optimal choosing on the part of the sender. But doing so consistently would also predict more extreme values in the LOW SUSPICION condition which were not observed.

that the information being conveyed is of a generalizable nature (Csibra & Gergely, 2009; Topál, Gergely, Miklósi, Erdőhegyi & Csibra, 2008).

More broadly, a central idea of Relevance Theory (Wilson & Sperber, 2004) is that of *ostensive-inferential* communication, the purpose of which is not only to inform one's interlocutor, but also to inform them of your intention to inform them. The idea of what we might call *ostensive meta-inferential* communication is closely related. A simple example can be found in everyday discourse. Replying "It's after 5." when a colleague asks you the time suggests not that the time is "5:01" as it might under a strongly informative assumption, but more likely that it is some time after 5 o'clock (and presumably before 6 o'clock). The use of the modifier "after" may signal that the recipient should not generalise too narrowly from the data. Using our computational model we analysed one particular form that ostensive meta-inference might take. The OSTENSIVE model captured the notion that although two stimuli might license the same inference in a particular context, the more ostensive one would licence stronger inference in a broader range of contexts. The results of our sender experiment suggests that senders considered such implications when weighing up their options. This was evident in their avoidance of the *Misleading* option in the HIGH SUSPICION condition, even when it was technically no more informative than the *Uninformative* option (see Fig. 11 for an example of such).

In experiments investigating the generation of referential expressions, the production of contextually redundant information (so-called *over-specification*) has been frequently observed, while under-specification is comparatively rare (Pogue, Kurumada & Tanenhaus, 2016). And while under-specification is consistently rated as unhelpful by receivers, over-specification is not viewed in this way (Engelhardt, Bailey & Ferreira, 2006). Indeed, by making communication more robust, over-specification can facilitate faster object identification (Arts, Maes, Noordman & Jansen, 2011). In our experiment, misleading but uninformative stimuli can be considered "over-specified", at least in relation to the purely uninformative stimuli. Thus, these findings lend support to the idea that people in our experiment would consider the ostensive properties of stimuli when reasoning about evidence. If over-specification is common and helpful, then for some senders it will make sense to favour it when the receiver has no reason to be suspicious and to avoid doing so otherwise (for fear of the strategy back-firing).

There is some evidence to suggest that a complementary tactic of ostensive under-specification may too play a role in deceptive communication. In a study of non-verbal deception with parallels to our own, Montague et al. (2011) used a "rectangle game" to investigate the use of deceptive strategies and their impact on learners. Participants played the part of informants who indicated points within or outside of a rectangle, or learners who had to infer the true boundary from the evidence provided. The cover story and instructions provided to learners left the helpfulness of informant testimony in question. Although informants were allowed to lie outright, it was not the preferred strategy in the competitive condition, presumably because learners were allowed to

verify information. Instead, informants in that condition favoured points which were relatively uninformative (and had no significant correlation with learner error - a measure of deceptive success in this case). Because informants in the cooperative condition were required to provide more points than the two strictly required to mark the opposite corners of a rectangle, they too provided uninformative points (which also had no significant correlation with learner error). Nonetheless, informants displayed context sensitivity in the choice of uninformative points. Uninformative evidence provided by cooperative informants was mostly positive (within the rectangle), while competitive informants favoured negative evidence (exterior points). In information theoretic terms, whether negative evidence is more or less informative than positive evidence depends upon the structure of the hypothesis space and the size of the hypothesis in question (Navarro & Perfors, 2011). But given the lack of correlation in Montague et al.'s data between learner error and uninformative evidence of either kind, the qualitative reversal of strategy observed between cooperative and competitive informants is intriguing.

A plausible connection with ostensive signalling arises as a consequence of the frequently sparse nature of the hypotheses with which learners are concerned (Navarro & Perfors, 2011). In an environment where hypotheses are sparse the expected information value of negative evidence (in advance of actually determining it) is less than that of positive evidence. Deceptive informants sensitive to the average uninformativeness of negative evidence (rather than its context-specific value) may thus prefer it over positive (yet uninformative) evidence without any further inference required. There is evidence to suggest that this ostensive use of negative evidence may impact people's sampling assumptions. For example, it has been noted that negative evidence or evidence from a second concept can induce a weaker sampling assumption on the part of the learner (Hendrickson et al., 2019; Ransom et al., 2016).

Taken together, our own results and those of Montague et al. (2011) support the idea that deceptive informants are sensitive to the ostensive qualities of data as well as its context-specific information content. An interesting avenue for future research would be to investigate whether any such sensitivity is heightened in deceptive contexts or representative of communication more broadly. An awareness of such differential sensitivity has the potential to benefit verbal deception detection techniques such as *forced choice tests* (Frederick & Speed, 2007) and *model statements* (Vrij, Leal & Fisher, 2018).

6.3. Individual differences in meta-inferential stance

Responses across both experiments were subject to important qualitative differences amongst individuals. On the comprehension side, only Adaptive receivers were sensitive to a difference in the evidentiary value of data in cooperative and competitive contexts. Likewise on the production side, only Adaptive senders were sensitive to suspicion in forming meta-inferential assumptions. Similar patterns of individual differences have been noted elsewhere in the literature. In a related

study, Franke & Degen (2016) used a similar Bayesian modelling framework to analyse production and comprehension behaviour in a (cooperative) reference game. They found that while listener behaviour appeared consistent with *Gricean* reasoning (analogous to our ONE STEP model) in the aggregate, closer analysis revealed that the majority of listeners used so-called *exhaustive* reasoning (analogous to our STRONG model), with the average being skewed by a smaller number of highly pragmatic participants. On the back of their analysis Franke & Degen (2016) highlight the importance of considering individual differences in computational level analysis, lest averaging effects obscure the different computational strategies being employed. Based on our own analysis we echo these sentiments.

Given our analyses, how should we interpret the differences in assumptions between Adaptive and Conservative participants? One obvious answer relates these differences to differences in the depth of reasoning in which people engaged. Such differences have been observed in experimental studies employing strategic reasoning games (e.g., Stahl & Wilson, 1995; Hedden & Zhang, 2002; Ohtsubo & Rapoport, 2006). Stahl & Wilson (1995) for example, analysed people's responses across twelve 3×3 symmetric games. Comparing various models of player behaviour, they found that most people could be grouped into one of four major categories: *level 0* types who choose randomly, *level 1* types who reasoned as if their opponent was a level 0 type, naive Nash types who used an equilibrium strategy (analogous to our RECIPROCAL model), and *worldly* types (the largest group) who reasoned that their opponent might be any one of the preceding types. Stahl & Wilson's finding that a significant proportion of people were sensitive to individual differences in reasoning styles connects with our own finding regarding Adaptive participants. If people expect a resonable amount of variation between (or within) individuals then the cognitive effort required to infer content-sensitive sampling assumptions may be justified. And given a sufficient population of Adaptive receivers, sensitivity to the meta-inferential implications of content makes sense for senders motivated to deceive.

But what about our Conservative participants – what might explain their behaviour? A simple explanation is that Conservative receivers failed to engage in meta-inferential reasoning at all. But given that the deception game explicitly entails the use of positive evidence only, an assumption that evidence was selected at random should justify a strong sampling assumption, not the weak assumption that Conservative receivers adopted. This does not rule out the *no meta-inference* explanation of course. The tendency of experimental participants to underweight the value of evidence has long been noted (e.g., Phillips & Edwards, 1966; Edwards, 1968), and a variety of explanations have been offered (for a review, see Corner, Harris & Hahn, 2010). Navarro, Dry & Lee (2012), found evidence that people adopt conservative sampling assumptions across a range of simple generalisation tasks. By modelling the strength of assumptions drawn (as a linear combination of strong and weak sampling), they found considerable variation amongst individuals. However, the "no meta-inference" explanation cannot account for Conservative senders – such

behaviour among senders would have meant choosing information at random, for which there was no evidence.

An alternative explanation for the behaviour of some Conservative receivers at least is not that they didn't (or couldn't) engage in the kind of meta-inference required, rather that they drew strong inferences but rejected them in favour of a more literal/logical interpretation. Feeney, Scrafton, Duckworth & Handley (2004) found evidence of comparable pragmatic inhibition. Their study looked at how people respond to uses of "some" that are felicitous (e.g. *some cars are red*) or infelicitous (e.g. *some birds have wings*). Reaction time data indicated that people took longer to endorse the literal meaning of infelicitous examples, suggesting extra cognitive effort was required to reject a misleading implication (for example, that *some but not all birds have wings*). The idea that some receivers draw but reject misleading inferences would help to explain the presence of conservative senders who avoid making such implications in the first place. However, given that our experiment was not designed to distinguish between the "no meta-inference" and "rejected meta-inference" explanations, this remains an area for future investigation.

7. Conclusion

We presented a computational framework for modelling the production and comprehension of information in a combined experimental and computational study of deception without lying. Our work makes two main contributions. First, we have provided an empirical demonstration that by formalising the production of messages as the computational inverse of comprehension it is possible to capture the behaviour of people seeking to mislead or conceal information from suspicious or naive targets. On the flip side, we have shown that by casting people's beliefs about the contingent nature of message production as probabilistic sampling assumptions, the same model can capture people's inferences when they are knowingly or unknowingly the target of deception. Reflecting on the findings of decades of deception research, Levine & McCornack (2014) argue that the principle drivers of deceptive behaviour are rational and utilitarian. People deceive when they need to, making the best of the information they possess given the contextual constraints. Further, they argue that the practical concerns of deception detection would be better served by an understanding of message content and the context in which it is produced, than by the myriad non-verbal cues which have proved relatively ineffective (see for example, Bond & DePaulo, 2006). By showing that the framework can capture a diversity of behaviour — that is, production and comprehension tasks in both cooperative and non-cooperative scenarios and across contexts where suspicion does and does not naturally arise — we hope to have demonstrated its applicability for further deception research.

Importantly, by using the framework to examine the predictions that particular models *cannot* make, we have been able to test alternative hypotheses concerning the ways that content and context combine to drive inference beyond the data provided. Our second contribution is thus

an empirical demonstration and analysis of the context- and content-sensitive nature of meta-inference.

The process of reasoning about the inferences of another, has been studied in a variety of settings, including concept learning and teaching (Shafto et al., 2014), learning from goal directed actions (Baker, Saxe & Tenenbaum, 2009; Shafto, Goodman & Frank, 2012b; Ullman, Baker, Macindoe, Evans, Goodman & Tenenbaum, 2009), intentional selection (Durkin, Caglar, Bonawitz & Shafto, 2015; Shafto & Bonawitz, 2015), preference learning (Jern, Lucas & Kemp, 2017), attitude attribution (Hawthorne-Madell & Goodman, 2015; Walker, Smith & Vul, 2015), and pragmatic language understanding (Frank & Goodman, 2012; Harris, Corner & Hahn, 2013; Hawkins, Stuhlmüller, Degen & Goodman, 2015; Goodman & Stuhlmüller, 2013; Franke & Degen, 2016). These studies share a common view that people make probabilistic assumptions about the way that others reason and act, and that they take this into account when drawing conclusions and communicating. Our work adds to this growing body of literature demonstrating that people enjoy the benefits of such meta-inference, learning more from less when interlocutors cooperate, while guarding against those seeking to exploit such tendencies in order to mislead.

8. Acknowledgments

KR was supported by an Australian Government Research Training Program Scholarship. AP received salary support from ARC grants DP110104949 and DP150103280 and DJN from ARC grant FT110100431. Preliminary versions of this work were published in the proceedings of the 39th Annual Meeting of the Cognitive Science Society.

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Bond, C., & DePaulo, B. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234.

Coleman, L., & Kay, P. (1981). Prototype semantics: The english word lie. *Language*, *57*, 26–44.

Colman, A. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, *7*, 2–4.

Corner, A., Harris, A., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1625–1630).

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*, 148–153.

Dulcinatti, G. (2018). *Cooperation and pragmatic inferences*. Ph.D. thesis University College London.

Durkin, K., Caglar, L. R., Bonawitz, E., & Shafto, P. (2015). Explaining choice behavior: The intentional selection assumption. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 614–619).

Dynel, M. (2011). A web of deceit: A neo-gricean view on types of verbal deception. *International Review of Pragmatics*, *3*, 531–538.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment* (pp. 17–52)). New York, NY: Wiley.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*, 554–573.

Eskritt, M., Whalen, J., & Lee, K. (2008). Preschoolers can recognize violations of the gricean maxims. *British Journal of Developmental Psychology*, *26*, 435–443.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *58*, 121–132.

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. In R. Sun, G. W. Cottrell, & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 1287–1293).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*, e0154854.

Franke, M., Dulcinati, G., & Pouscoulous, N. (2019). Strategies of deception: Under-informativity, uninformativity, and lies—misleading with different kinds of implicature. *Topics in Cognitive Science*, *early access*. doi:10.1111/tops.12456.

Frederick, R. I., & Speed, F. M. (2007). On the interpretation of below-chance responding in forced-choice tests. *Assessment*, *14*, 3–11.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*, 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*, 173–184.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*, 9066–9071.

Hardin, K. J. (2010). The Spanish notion of *lie*: Revisiting Coleman and Kay. *Journal of Pragmatics*, *42*, 3199–3213.

Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*, 414–429.

Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 878–883).

Hawthorne-Madell, D., & Goodman, N. D. (2015). So good it has to be true: Wishful thinking in theory of mind. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 884–889).

Hedden, T., & Zhang, J. (2002). What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, *85*, 1–36.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford, & N. Chater (Eds.), *Rational Models of Cognition* (pp. 248–274). Oxford: Oxford University Press.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sam-

pling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Hespanha, J. P., Ateskan, Y. S., & Kizilocak, H. (2000). Deception in non-cooperative games with partial information. In *Proceedings of the 2nd DARPA-JFACC Symposium on Advances in Enterprise Control* (pp. 1–9).

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20–58.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*, 785–807.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology*, *33*, 431–440.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cmabridge, MA: MIT press.

Montague, R., Navarro, D., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*, 187–223.

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery and the size principle. *Acta Psychologica*, *133*, 256–268.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*, 120–134.

Ohtsubo, Y., & Rapoport, A. (2006). Depth of reasoning in strategic form games. *The Journal of Socio-Economics*, *35*, 31–47.

Okanda, M., Asada, K., Moriguchi, Y., & Itakura, S. (2015). Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in Psychology*, *6*, 901. doi:10.3389/fpsyg.2015.00901.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*, 833–838. doi:10.1073/pnas.0707192105.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 2035.

Pryslopska, A. (2013). *Implicatures in uncooperative contexts*. Master's thesis University of Tübingen.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*, 1775–1796.

Reboul, A. (2017). Is implicit communication a way to escape epistemic vigilance? In S. Assimakopoulos (Ed.), *Pragmatics at its Interfaces* (pp. 91–112). Walter de Gruyter GmbH volume 17.

Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in psychology*, *6*, 867.

Rogers, T., Zeckhauser, R. J., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of personality and social psychology*, *112*, 456–473.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of semantics*, *23*, 361–382.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–66). MIT Press.

Schauer, F., & Zeckhauser, R. (2009). Paltering. In B. Harrington (Ed.), *Deception: From ancient empires to internet dating* (pp. 38–54). Stanford, CA: Stanford University Press.

Shafto, P., & Bonawitz, E. (2015). Choice from intentionally selected options. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (pp. 115–139). Academic Press volume 63.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012a). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012b). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*, 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Skarakis-Doyle, E., Izaryk, K., Campbell, W., & Terry, A. (2014). Preschoolers' sensitivity to the maxims of the cooperative principle: Scaffolds and developmental trends. *Discourse Processes*, *51*, 333–356.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*, 359–393.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*, 218–254.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*, 176–190.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–65). MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioural and Brain Sciences*, *24*, 629–640.

Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*, 1831–1834.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

Vogel, A., Potts, C., & Jurafsky, D. (2013). Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 74–80).

Vrij, A., Leal, S., & Fisher, R. P. (2018). Verbal deception and the model statement as a lie detection tool. *Frontiers in Psychiatry*, *9*, 492.

Walker, D., Smith, K. A., & Vul, E. (2015). The "fundamental attribution error" is rational in an uncertain world. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2547–2552).

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. R. Horn, & G. Ward (Eds.), *Handbook of pragmatics* (pp. 607–632). Oxford: Blackwell Publishing.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*, 288–297.