CESifo WORKING PAPERS

9286
2021

Original Version: September 2021
This Version: September 2023

Deceptive Communication: Direct Lies vs. Ignorance, Partial-Truth and Silence

Despoina Alempaki, Valeria Burdea, Daniel Read

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9286*

Deceptive Communication: Direct Lies vs. Ignorance, Partial-Truth and Silence

Abstract

In cases of conflict of interest, people can lie directly about payoff relevant private information, or they can evade the truth without lying directly. We analyse this situation theoretically and test the key behavioural predictions due to differences in psychological costs in a novel experimental sender-receiver setting. We find senders prefer to deceive through evasion rather than direct lying, more so when evasion takes the form of partial-truth. This is because they do not want to deceive others, and they do not want to be seen as deceptive. Receivers are highly sensitive to the language used to deceive and are more likely to act in the sender's favour when the sender lies directly. Our findings suggest dishonesty is more prevalent and potentially costlier than its previous best estimates focusing on direct lies.

JEL-Codes: C910, D820, D910.

Despoina Alempaki
Behavioural Science Group
Warwick Business School
Coventry / United Kingdom
despoina.alempaki@warwick.ac.uk

Valeria Burdea
LMU Munich
Faculty of Economics
Munich / Germany
valeria.burdea@econ.lmu.de

Daniel Read
Behavioural Science Group
Warwick Business School
Coventry / United Kingdom
daniel.read@warwick.ac.uk

September 1, 2023
We are grateful to Daniele Nosenzo, Collin Raymond, Friederike Reichel and Daniel Seidmann for valuable feedback. We also thank participants at the ESA 2020 Global meeting, the Winter Summit on (Un)ethical Behavior in Markets in Innsbruck, the Social Image and Moral Behavior workshop in Cologne, and at seminars at the University of Nottingham, WZB Berlin, the University of Portsmouth, the University of Vienna, the University of Warwick, Aarhus University, the Institute for Advanced Studies in Vienna, GATE-Lyon and the Birmingham-Warwick Strategic Information Network for excellent comments and suggestions. This work was supported by the Economic and Social Research Council [grant number ES/K002201/1, ES/P008976/1] via the Network of Integrated Behavioural Science. The Humanities and Social Sciences Research Ethics Committee at the University of Warwick reviewed and approved the procedures.

Do I agree that I lied? I don't know of times when I lied. Look, there are times when I, certainly times when I was acting as a representative, as a marketer for FTX and when I was looking for how can I — in a way which is truthful — paint FTX in as a compelling way as possible. Sam Bankman-Fried, quoted in New York Times (Dec 1, 2022).

1. Introduction

Most news sources would be impoverished indeed if we were to remove all cases of lying, deception and fraud from their front pages. Yet this abundance of real-world deception is apparently at variance with experimental research which finds that people are surprisingly reluctant to lie (e.g., Abeler et al., 2019), even when by lying they would obtain material benefits from others being deceived, and even when their lies cannot be punished or even detected (e.g., Bucciol and Piovesan, 2011; Fischbacher and Föllmi-Heusi, 2013; Gneezy, 2005; Mazar et al., 2008). This reluctance to lie has been ascribed to a psychological cost of lying that is primarily driven by two components: a preference for being honest, which produces an intrinsic cost of lying, and a preference for being *seen* as honest, which produces a social image cost (e.g., Abeler et al., 2019; Dufwenberg and Dufwenberg, 2018; Gneezy et al., 2018; Khalmetski and Sliwka, 2019 for recent evidence on the structure of lying costs).

While any such aversion to lying would undoubtedly be a helpful check on the tendency to deceive, the prevalence of deception suggests it is far from 100% effective. One reason, we propose, is that the deceptions which previous research has largely focused on are what we call *direct lies*, meaning direct falsehoods about instrumental information. For instance, in a typical study participants might roll a die and report the number that came up to determine the payment they will receive (Fischbacher and Föllmi-Heusi, 2013). Reporting a higher number than the one actually observed would constitute a direct lie. Yet in the wider world, even the wider experimental world, direct lying is not the only way to attempt deception, and what is true for direct lies may not be true for other forms of deception. In particular, some of the psychological costs of lying may be avoided by deceiving without lying.

Consistent with this view, we have observed it is commonplace for people caught in seemingly quite egregious falsehoods to argue that because they did not (exactly) lie, their bending of the truth is not all that bad. This is seen in the epigraph to this paper, from disgraced FTX CEO Sam Bankman-Fried, which also illustrates how important it is even for liars not to be seen as such. It is also common for those with something to hide to carefully choose their words so that their statements might be interpretable as non-lies. Former President Bill Clinton made a specialty of this, perhaps most notoriously when he denied his relationship with Monica Lewinsky by stating that "there *is* no improper relationship" when, in fact, there most certainly *had* been one -- but it was now over. He later said that he had chosen these words because he "didn't want to lie." (PBS, 2004). Apparently, people choose their words as if seeking to avoid paying the full psychological costs of lying.

In this paper we investigate *evasions*, messages which bend, withhold or distort the truth, but do not necessarily involve direct lies. Evasions are a diverse species, defined largely as attempts to convey something other than the (entire) truth through some means other than a direct lie.

To illustrate the distinction between the range of possible evasions and direct lies, consider a manager who receives a promotion request from an employee and is now asked by that employee to report on the progress of their request. Imagine the manager knows that, in fact, this request has already been denied and that it will be a year before another request can be made, but the manager also wants to postpone giving their employee the news, especially since the employee is considering an attractive outside option. The manager could lie directly, by stating that the promotion case is currently being given very favourable consideration by the board. The manager might, however, prefer less extreme deception and so choose to evade, perhaps by feigning ignorance through saying that "I do not know what the board intends to do," or by providing a partial truth, such as "promotions will be discussed at the next board meeting" (true, except *this* promotion will not be), or they may simply remain silent on the issue altogether by changing the subject to the employees' family or their vacation plans.

In this paper we compare direct lies to evasions in terms of their psychological costs, in terms of how much people are inclined to use them, and in terms of how effective they are. We do this through both theory and a novel experimental design that allow us to cleanly identify psychological factors as an important driver of differences between the various deceptive communications. Our study provides insights into the pervasiveness and consequences of deception, and how and why it may be more widespread than current best estimates which have primarily focused on direct lies (e.g., Abeler et al., 2019; Egan et al., 2019; Gerlach et al., 2019; Gurun et al., 2018; Johnson et al., 2019).

Applying key ideas from Sobel (2020), we extend the concept of lying cost by distinguishing four psychological costs that can be incurred by those who attempt to deceive, and which may differ depending on how that deception is carried out.[1] These costs include: 1) a *deception* cost, incurred when acting on the intention to create or maintain a false belief. All lies and evasions, in our view, will give rise to a deception cost; 2) a *falsehood* cost, incurred when making a statement believed to be false (note that deception and falsehood combined are what is conventionally called "lying"). Not all evasions will incur a falsehood cost because many evasions are truthful even if incomplete or irrelevant; 3) an *influence* cost, which increases in the perceived likelihood that the message will lead its recipient to adopt a course of action not in their interest; and 4) a *social image* cost, which increases to the degree that the recipient of the message judges its sender to be dishonest. We apply this analysis to the three representative classes of evasion already introduced in the story of the manager and the luckless employee: feigning ignorance, telling partial truths and remaining silent. These classes of evasion are

[1] Braghieri (2023) is also relevant when defining the concepts of deception and lies, but that paper focuses on the listener side, whereas Sobel (2020) focuses on the speaker's perspective. Given our focus on the psychological costs that speakers incur in strategic communication settings, we follow Sobel (2020) when describing our theoretical framework.

easily identifiable and commonplace, as indicated by our own pilot research (see Appendix C). They also cover a broad range of possible degrees of falsehood and deception.

We investigate the role of the different psychological costs in evasions and direct lies by means of experimental investigations of an asymmetric information game between an informed sender and an uninformed receiver where the sender has a material incentive to deceive and messages are cheap talk. The experimental game is a new variation of the widely studied cheap-talk sender-receiver game (e.g., Blume et al., 2020; Crawford, 1998; Crawford and Sobel, 1982; Gneezy, 2005; Khalmetski et al., 2017; Sobel, 2020) that allows us to isolate the channel of psychological costs. Within this game we provide the same classes of deceptive communications (or "messages") as were available to the manager in our fictional scenario: to lie outright, to feign ignorance, to provide partial truth, and to remain silent -- as well, of course, as to simply tell the truth.

Although evasions are intended to deceive, we hypothesised that at least some and sometimes all of the psychological costs just identified are lower for evasions than for direct lies. Consequently, message senders will be more likely to evade than to lie, holding the benefits from the deceptive communication constant. We can illustrate this with the evasive manager. Take as an example the partial-truth that "*promotions* will be discussed at the next board meeting." It is likely the manager hopes the employee will interpret this as "*your promotion* will be discussed …". Indeed, that would be the only circumstance in which the next board meeting is a relevant response to the employee's inquiry. However, unlike a direct lie, the manager's statement is strictly true, and so, while it will incur the deception cost, it will not incur the falsehood cost, making the partial truth a "cheaper" deception and so, more likely to be chosen. Similarly, the manager may want the employee to stay at the firm and not look for a new position – something that is more likely to happen if the employee is told the truth. The influence cost is incurred to the degree that the manager's evasive message satisfies the employee, and keeps them happily on staff a few months longer, when the employee would be better off sending out their resume. The direct lie is more likely than any of the three evasions to, at least temporarily, keep the employee at the firm, and so the influence cost of the direct lie is greater than that of these evasions, making the evasions more likely to be chosen based on influence cost alone. Finally, through evasion the manager may be able to avoid incurring a social image cost, because the employee may never be certain that the manager deceived them (i.e., they will not know in which meeting the promotion decision was discussed), and perhaps will not even suspect they were deceived. Again, the evasion is more likely to be chosen if social image costs matter. In sum, evasions generally incur lower costs than direct lies, and will often be preferred to lies whenever these costs matter. Consequently, we predicted that senders would be more likely to evade than to lie directly.

We also hypothesized that evasions differ amongst themselves in the psychological costs they incur. These detailed hypotheses are presented fully in Section 4, but here we summarise. First, we hypothesised that silence would have a lower influence cost than partial truth, and consequently would be chosen more frequently. This means that staying silent would be less likely to persuade the employee

that the promotion is being given full consideration than would partial truth ("promotions will be discussed …"). In addition, we hypothesised that partial truth would incur a lower falsehood cost than feigning ignorance (which is a lie even if not a direct lie). Consequently, partial truth would be chosen more often than feigning ignorance.

Moreover, we hypothesised that evasions incur a lower social image cost than direct lies because the receiver cannot be sure they were deceived. We therefore predicted the difference between the likelihood of choosing evasion and that of choosing a direct lie would be reduced if this credible deniability were eliminated by informing the sender that the receiver will learn they were evasive. Finally, with respect to the persuasiveness of the different communications, we hypothesised that receivers would be more likely to act in the sender's favour (i.e., be taken in) when the sender lies directly. This is because of receivers' naivety, inducing them to take messages at face value and act according to the recommendation implied by the message.

We conducted three pre-registered incentivised experiments (N = 3,615), two examining the actions of senders and one those of receivers. In our experimental game, there are two possible states of the world, *Red* and *Blue*. The sender views a private signal that either fully specifies the state (definitely Red, or definitely Blue) or leaves it unknown (it could be Red or Blue, with each possibility having a known probability). The sender gains a material advantage if the message receiver always believes the state is Red. When the state is Blue, therefore, the sender has an incentive to deceive. In our game it is only then, when the state is Blue, that the sender must choose a cheap-talk message to send to the receiver. This message can be either truthful or deceptive: each sender can choose between only two options, the truth or a single, specific deceptive option, drawn from the four deceptions described above.

In the direct lie treatment (DIRECT -- we use all caps to denote these experimental treatments), the sender chooses between telling the truth and a direct lie. In three evasion treatments, the choice is between telling the truth and evading by feigning ignorance (IGNORANCE), by telling partial truths (PARTIAL), or by remaining silent (SILENCE). Upon receiving a message, the receiver chooses an action which determines the payoff for both players. As noted already, there is a conflict of interests: the sender always wants the receiver to choose Red, whereas the receiver wants to choose the correct colour whether it is Red or Blue.

Experiments 1 (*Sender-Hidden*) and 2 (*Sender-Open*) focused on senders. In each experiment we compared senders' choices across four treatments that differed only in the type of deceptive communication available to them, with the three evasions being those discussed already. As we explain in Section 3, the evasions differed from direct lies in that they allowed for plausible deniability on the part of the sender, since it could never be known if the sender was evasive or truly uninformed. Sender-Hidden allowed for this plausible deniability, since the sender's decision was not revealed to receivers. Sender-Open, however, ruled out plausible deniability by explicitly revealing the sender's decision to the receiver at the end of the game, and letting the sender know this would be done before they chose

their message. By comparing the Sender-Hidden and Sender-Open experiments, we could therefore test for the role that social image plays in choices to deceive.

We also obtained senders' incentivised beliefs about how receivers would respond to each message and, in Experiment 3 (*Receiver-Hidden*) the actual responses of a large number of receivers to each of the four deceptive communications. As hypothesised, direct lies were chosen less frequently than evasions in the two sender experiments, especially in Sender-Hidden. In that experiment, DIRECT had a lower deception rate than all the evasion treatments, significantly lower than PARTIAL and SILENCE. We also find that social image costs play a large role, as the difference between DIRECT and the three evasion treatments was substantially reduced in Sender-Open, and the DIRECT versus SILENCE comparison ceased to be significant. However, social image costs are not the only driver of the difference between direct lies and evasion, since even in Sender-Open there remained significantly less deception in DIRECT than PARTIAL.

With respect to whether the language of evasion matters, we find that in both Sender experiments individual heterogeneity led to statistically indistinguishable rates of deception across the three evasion treatments. However, when controlling for this heterogeneity, senders in Sender-Hidden engaged in significantly more deception in both PARTIAL and SILENCE compared to IGNORANCE. The difference between PARTIAL and IGNORANCE remained significant even after increasing the social image costs in Sender-Open. This suggests that the falsehood cost is a key determinant of differences among evasions and potentially more important than the influence cost. Moreover, the remaining differences observed in Sender-Open highlight that the variety of non-falsehoods were associated with different image costs and, in particular, that active silence was seen as more costly for one's social image than a partial truth.

After showing that deception rates differed between direct lies and evasion, we examined whether this might be due to senders' expectations about the potential benefits from deception. Perhaps, for instance, senders believe an evasive message is more likely than a direct lie to elicit the desired "Red" response. However, the incentivized elicitation of senders' beliefs about receivers' actions suggests this is not the case. If anything, senders believe that receivers are *more* likely to choose Red after the direct lie. That is, even though a direct lie is more likely to elicit the highest payoff for senders, they are less likely to choose it. This strengthens our view that evasion is less psychologically costly than a direct lie, because it is chosen despite offering a lower material benefit.

We analysed the Receiver-Hidden experiment to learn which deceptive communications were most persuasive as well as the monetary implications of this persuasiveness. Direct lies were more convincing than all evasions. We also found a striking pattern, indicating that partial truths were significantly more persuasive than feigning ignorance or remaining silent. An analysis of receivers' beliefs suggested that their choices were not driven by beliefs about senders' decisions or, by implication, anticipated differences in senders' psychological costs, since the receivers believed that senders were equally likely to deceive in all treatments. Rather, our data suggests receivers are naive

and take messages at face value. The most persuasive messages are those that most strongly indicate which action the receiver should take. This result is consistent with related research in cheap-talk sender-receiver games showing that receivers' largely follow the senders' recommendation (e.g., Gneezy, 2005).

By combining the data from the Sender-Hidden and Receiver-Hidden experiments, we obtained important new insights regarding the welfare implications of deception through evasion. In particular, all forms of evasion can be materially harmful for both senders and receivers, and sometimes even more so than direct lies. This is true not only when interests are misaligned but even when they are aligned, meaning that policies properly targeted to reduce the various forms of deception can be Pareto improving.

We contribute to previous literature investigating deception when evasion is possible in addition to (or instead of) direct lies. Serra-Garcia et al. (2011) show that senders sometimes use vague messages instead of precise but untruthful ones to disguise the truth. Similarly, senders frequently stay silent (e.g., Leibbrandt et al., 2017; Sánchez-Pagés and Vorsatz, 2009) or pretend ignorance (e.g., Khalmetski et al., 2017; Khalmetski and Tirosh, 2012) instead of telling a direct lie in cheap-talk games. Also related is the study by Turmunkh et al. (2019), who analyse data from a TV game show where players make non-binding pre-play statements about their willingness to cooperate in a prisoners' dilemma and argue that many players who plan to defect use indirect statements or evasions to disguise their intentions rather than direct lies claiming to cooperate.

Our work is distinguished from previous research in that we do not investigate whether people prefer evasion over direct lying when both are possible. Khalmetski et al. (2017), illustrates this other approach within economics with a study in which the sender has three options (tell the truth, tell a direct lie, or declare ignorance) and the expected payoff of ignorance is higher than the expected payoff of direct lying. Instead, we seek to understand whether any preference for evasion is due to differences in the psychological costs of each communication in isolation, and not due to differences in perceived (or actual) relative benefits which might arise when all options appear side by side.

A key contribution of our experiments is therefore that they provide a direct test of whether evasion is less psychologically costly than outright lying by ruling out "menu effects," since the sender has only two options, either to tell the truth or deceive, with some being able to deceive by direct lying and others by evasion. In addition, we are the first to systematically contrast multiple commonplace types of deception in a unified framework, to isolate the role of the social image cost in making evasion a more attractive means of deception, and to compare senders' beliefs about receivers' scepticism toward different forms of deceptive communication. In combination, these design elements allow us to better understand the limits of deception-reduction mechanisms that focusing on material or reputational costs (e.g., increasing detection probability, subsequent punishment value or visibility of such actions), while providing the grounds for developing and testing new solutions for tackling deception at various organizational levels.

Also relevant to this research are studies from outside economics that similarly distinguish between varieties of deception. Schauer and Zeckhauser (2007) use the term "paltering" much as we do evasion, arguing that the possibility of credible deniability means it should be responded to with appropriately large sanctions. Rogers and Norton (2011) discuss "dodging" or answering a different question than the one being asked. Bickart et al. (2015) describe "obfuscation", or providing answers to questions that are irrelevant and tangential but might appear pertinent at first glance. Kang et al. (2020) distinguish, as we do, between lying and evading and argue that for self-presentational and emotional resons consumers often prefer the latter. Another important distinction is between lies of omission and commission (e.g., Bok, 1978; Gaspar et al., 2019; Levine et al., 2018; O'Connor and Carnevale, 1997; Pitarello et al., 2016; Spranca et al., 1991; Schweitzer and Croson, 1999 - see also the review by Fallis, 2018 and the references therein). We contribute to this literature by studying multiple forms of deception in a single overarching framework which allows for different deceptions to be studied together and compared to each other by means of incentive compatible tasks.

This work also brings important nuances to the study of receivers' naivety. A well-documented result in cheap-talk sender-receiver games with conflicting interests is that a significant proportion of receivers are too trusting, placing undue faith in senders' messages (e.g., Cai and Wang, 2006; Forsythe et al., 1999; Hurkens and Kartik, 2009; Sanchez-Pages and Vorsatz, 2007; 2009; Sheremeta and Shields, 2013). Most of these studies, however, examined settings where messages are direct, with one exception we know of: Sanchez-Pages and Vorsatz (2009). The communication game in their study differed from ours in several dimensions. First, senders were allowed to choose between truth, direct lies *and* silence. Second, the two deceptive communications (lies and silence) were associated with different ex-ante credibility. This is because there were no "uninformed" senders and so, remaining silent was a clear signal of avoiding telling the truth whereas a direct message could be sent by a truthful sender. Moreover, if senders chose to stay silent, receivers did not know what the senders' preferred action is, and so, their response to silence cannot be interpreted in the framework of persuasion. All of these differences make it impossible to pin down how receivers respond to *the language* of direct lies as compared to evasion in the form of silence. Our study allows us to make this comparison across three types of evasion, expanding our understanding of the mechanisms that lead receivers to "take messages at face value." Our findings suggest that how receivers interpret and act on senders' potentially deceptive communication is a function of both beliefs about the likelihood that the sender is deceitful and of the precision ("directness") of the language used in the message.

The remainder of the paper is organized as follows. Section 2 introduces the deception game and our theoretical framework. Section 3 presents our experimental design, and Section 4 our main hypotheses. Section 5 discusses the experimental results, while Section 6 discusses welfare consequences of evasion. Section 7 concludes with potential policy implications.

2. The Deception Game

We study a game with two players: a sender (S, she) and a receiver (R, he). The sender may have private information about the state. She can communicate with the receiver, but she cannot directly influence either player's payoffs. The receiver does not have private information about the state, but his actions determine the payoffs of both parties.

The game begins with nature determining if the state is Red or $Blue$. Red is more likely than $Blue$. In particular, the probability of Red is $\frac{11}{20}$ and that of $Blue$ is $\frac{9}{20}$. Nature also determines with probability $\frac{7}{10}$ whether the sender is informed about the state. The probability of Red is $\frac{3}{7}$ if the sender is informed and $\frac{5}{6}$ otherwise. The state is therefore more likely to be Red if the sender is uninformed $\left(\frac{5}{6}\right)$, and more likely to be $Blue$ if she is informed $\left(\frac{4}{7}\right)$.[2] A sender who is informed that the state is $Blue$ then chooses a message, either the truth or a deception, from a set of possible messages that depend on the experimental treatment. A sender who is informed that the state is Red always tells the truth to the receiver, and a sender who is uninformed always sends a specific message drawn from a set of evasive messages as described in the next paragraph. The receiver observes the message, guesses the colour of the state (Red or $Blue$), and the payoffs are realised.[3] All these details are common knowledge.

The deceptions we consider can be grouped in two broad categories: (i) direct statements - about the colour of the state (e.g., "The state is Red"), and (ii) evasive statements. We will refer to the following set of evasive statements as X, and to an element of this set as x_i.

x_1(IGNORANCE) = "I don't know the colour of the state"

x_2 (PARTIAL) = "The state was more likely to be Red than $Blue$"

x_3 (SILENCE) = $\emptyset$

When the sender is uninformed, one of these three statements is automatically sent. These messages are chosen such that they are applicable whenever the sender is genuinely uninformed, so that uninformed senders who use these messages cannot be construed as deceiving. When the sender is both informed and the state is $Blue$, the sender in the evasion treatments can choose between telling the truth ("The state is $Blue$") or sending one of these evasive messages. The corresponding sender in the direct lie treatment can either tell the truth ("The state is $Blue$") or tell a direct lie ("The state is Red").

Note that the message "The state is $Blue$" is perfectly informative, since it can only be sent when the state is $Blue$ and the sender is informed. As such, the setting does not allow for sophisticated deception via truth-telling (e.g., Sutter, 2009). On the other hand, when choosing the message "The

[2] These parameters are chosen such that in equilibrium the expected material benefit of evasion is not larger than that of a direct lie to ensure that a revealed preference for evasion cannot be due to higher expected material benefits.

state is Red" the senders in DIRECT pool with the truthful types, whereas when choosing one of the evasive x_i messages, senders in the evasion treatments pool with the uninformed types.

We allow senders to choose the message only when they have an incentive to disguise the truth (as will become clear when introducing the monetary payoffs in the next paragraph). This is both to attach natural meanings to messages, necessary for a literal interpretation of what constitutes a lie and to restrict the equilibrium strategies.

Payoffs. Table 1 summarizes the payoffs, where $h > l$ (the sender's payoff is listed first in each cell).

Table 1. Payoff matrix $(\boldsymbol{S}, \boldsymbol{R})$

		Receiver's choice	
		Red	$Blue$
State	Red	$(h,\ h)$	$(l,\ l)$
	$Blue$	$(h,\ l)$	(l, h)

Given the payoff structure, the sender maximizes her expected payoff if the receiver always chooses "Red." The receiver does so when he guesses the correct colour of the state. Note here that the sender's payoff depends only on the receiver's action while the receiver's payoff depends both on his action and on the colour of the state. Hence, when the game is finished and the receiver has observed his payoff, he will know the colour of the state. However, in case of an evasive statement, the receiver will not be able to infer with certainty whether the sender was informed, since both states can arise when the sender is uninformed.

Definitions. To structure the exposition, we introduce some definitions.

First, we define the *literal meaning* of a message as being what the message says. If the message states a fact, then its literal meaning is that fact. For example, the literal meaning of "The state is $Blue$" is that the state is, indeed, $Blue$. Throughout, we maintain the assumption that these literal meanings are understood. Denoting the set of messages as M, with members of the set denoted m_i and a chosen message as m, then we have the following definition:

Definition 1 (Literal meaning). The literal meaning of m is the a priori, common understanding that $m = m_i$ implies that some characteristic of the game takes the value m_i.

Next, we distinguish between direct and evasive messages. A direct message states the value of the state. For example, "The state is $Blue$" is a direct message. Such messages are, by construction, not probabilistic and so we call them direct because their literal meaning makes a clear and definite suggestion regarding the value of the state (and hence, such a message has a direct implication for the action the receiver should take).

Definition 2 (Direct message). A message $m = m_i$ is direct if $m_i \in \Theta$, where Θ is the set of all possible values of the state.

A message is evasive when it does *not* make a direct suggestion regarding the state *and* a direct truthful message is also available. For example, "The state might have been $Blue$" is evasive if the sender knows the truth about the state (i.e., whether it is $Blue$ or Red) and could have communicated it in a direct, non-probabilistic manner.

Definition 3 (Evasive message). A message $m = m_i$ is evasive if $m_i \neq \theta$ and the sender is informed about the state and $M(\theta) \supset \{\theta, x\}$, where $\theta \in \{Red, Blue\}$ and $x \in X$.

Next, we define truthful messages as those messages with a literal meaning equal to the value of the characteristic of the game the message refers to.

Definition 4 (Truth). A message $m = m_i$ is true if $m_i = j$, $\forall m_i$, where j is the value of a characteristic of the game.

Given this, we define lies as messages with a literal meaning that differs from the truth. For example, "The state is $Blue$" is a lie if, in fact, the state is Red. Similarly, "I don't know the colour of the state" is a lie if the sender does know the colour. Given our focus on strategic settings, this definition follows Sobel (2020) who defines lies strictly in terms of the relation between truth and the literal meaning of the message.

Definition 5.0 (Lie). A message $m = m_i$ is a lie if $m_i \neq j$, $\forall m_i$, where j is the value of a characteristic of the game.

We further distinguish between direct and evasive lies. In line with Khalmetski et al. (2017), a lie is direct if it concerns the value of the state. In the examples above, "The state is $Blue$" is a direct lie since the colour of the state is in fact Red. Formally:

Definition 5.1 (Direct Lie). A message $m = m_i$ is a direct lie if $m_i \in \Theta$ and $m_i \neq \theta$, where Θ is the set of all possible values of the state, and θ is the value nature drew.

A lie is evasive if it is about any characteristic of the game that is not the state – the only characteristic with direct payoff relevance. Saying, for instance, "It is Saturday" on a Sunday, when the day of the week is payoff irrelevant, is an evasive lie. Similarly, saying "I don't know the colour of the state" when one does know, is an evasive lie. Importantly, direct lies can be detected upon the payoff realization, whereas evasive lies cannot.

Definition 5.2 (Evasive Lie). A message $m = m_i$ is an evasive lie if $m_i \notin \Theta$ and $m_i \neq j$, where j is any characteristic of the game different than θ.

We follow Sobel (2020) and distinguish between lies and deceptions. Deception is defined relative to other available messages. Specifically, a message is deceptive if (a) the sender has a choice between which message to send, and (b) relative to other messages the sender could send, the message will lead the receiver further from an accurate belief about the state. For instance, saying "I don't know the colour of the state" is deceptive when one knows it is Red and could say instead "The colour is

Red." This is because the first statement is likely to lead the receiver farther from the truth than the second.

Definition 6 (Deception). Let $\mu(\theta)$ be the receiver's belief about the state. A message $m = m_i$ is deceptive if $\mu(\theta|m_i) - Pr(\theta) > 0$ and S has the option to send $m' = m'_i$ for which $\mu(\theta|m_i) - Pr(\theta) > \mu(\theta|m'_i) - Pr(\theta)$.

In other words, messages are deceptive when they induce more inaccurate beliefs than another available message would. A belief $\mu(\cdot\,|m_I)$ is inaccurate if, given θ, $\mu(\theta|m_i) \in [0,1)$; that is, whenever the receiver believes that, given a message, the state is not 100% likely to take its true value (similar to Sobel, 2020). The farther from 1 this belief is, the more inaccurate it is.

2.2. Analysis

We delegate the formal analysis to Appendix A and discuss here its key insights. If both senders and receivers care only about material payoffs, senders are indifferent between the direct and the evasive deception, and they will pool on one of them. Receivers will choose $Blue$, when receiving the message $Blue$ or senders' pooling message, and Red otherwise (see Lemma 1 in Appendix A for the proof). So, the expected benefit to senders of a direct or evasive lie is the same in both cases, and equal to the low payoff (l).

However, people are not perfectly rational and also care about non-material payoffs. As we describe next, given certain assumptions about these behavioural features of receivers and senders, the likelihood of choosing the deceptive message depends on the message set. First, we assume that receivers are one of two types: sophisticated (R^S) or naive (R^N) (similar to e.g., Kartik, 2009).[4] A sophisticated receiver chooses the action that maximizes his expected payoff given his beliefs about the state distribution which are updated in line with Bayes' rule upon observing the sender's message.

In contrast, a naive receiver interprets the message literally.[5] Specifically, if a message makes no statement about the state, the naive receiver's posterior belief about the distribution of the state remains equal to his prior (i.e., $\mu_{R^N}(\theta = Red) = \Pr(Red) = \frac{11}{20}$). If the message makes a statement about the payoff relevant state dimension, the naive receiver's posterior belief moves away from the prior in the direction implied by the message, more so depending on the precision of the message. That is, if $m = Red$, $\mu_{R^N}(\theta = Red|m) = 1$; if $m = Blue$, $\mu_{R^N}(\theta = Red|m) = 0$; if $m = x$ and the message implies a higher probability for the state taking the value Red, then $\mu_{R^N}(\theta = Red|m = x) > \frac{11}{20}$. The naive receiver then chooses $a = Red$ if their posterior belief suggests that $\theta = Red$ is at least equally likely as $\theta = Blue$, i.e., $\mu_{R^N}(\theta = Red|m) \geq \frac{1}{2}$.

---

[4] Kartik (2009) introduces naïve receivers in an alternative but equivalent way by assuming that receivers are likely to take a naïve action with a certain probability, e.g., η.

[5] We obtained strong empirical support for this assumption as discussed in detail in Section 5.

Furthermore, naive receivers do not draw inferences about the sender's message (i.e., whether it is deceptive or truthful) from comparing the realised and expected payoff. That is, if the sender sent $m = Red$ when they knew the state was $Blue$, and the receiver chooses $a = Red$ (or $a = Blue$) getting a payoff of l (or h), the naive receiver does not infer whether the message is deceptive by comparing the payoff they receive to the one they would have if the message was truthful. The sophisticated receiver, however, does go through this inference process. Therefore, the likelihood that a deceptive message (in particular, a direct lie) will be interpreted as such depends on the proportion of sophisticated receivers in the population. This proportion influences the magnitude of the social image cost described below.

Next, we assume that senders incur psychological communication costs. We consider four types of cost:

- a *deception* cost - incurred whenever the sender chooses a lie or an evasion (i.e., when $m \neq Blue$);
- a *falsehood* cost - incurred when the message is false (i.e., a lie);
- an *influence* cost - which increases with the difference between the sender's belief about $\mu(\theta)$, the receiver's belief about the state, and the realized probability of the state. That is, the influence cost increases the more inaccurate the beliefs induced by the sender's message are[6];
- a *social image* cost - incurred when the sender's message is not the truth and increasing with the probability the receiver can learn the sender was deceptive upon the realization of payoffs.

When the message is perfectly informative about the sender's type (i.e., the receiver can infer it from the message with certainty) or the sender does not have a choice regarding which message to send, we assume no communication cost. This happens when $m = Blue$ (a perfectly informative message that is only available to the informed sender when $\theta = Blue$) or when a message is sent automatically (i.e., either when $\theta = Red$ or the sender is uninformed).

Communication costs may vary across senders and across situations. If these costs are sufficiently high, the sender will always tell the truth (truthful type). If they are sufficiently low, the sender will deceive when it is beneficial to do so (dishonest type) (see Lemma 1 and Corollary 1 in Appendix A). Importantly, communication costs also vary across messages. First, note that the literal meaning of x_1 is that the sender is uninformed, that of x_2 is that the state had a higher chance of being Red than $Blue$, while x_3 represents silence or making no claim about any state dimension. These messages can only influence the naive receiver's beliefs about the payoff relevant characteristic, and only x_2 changes the naive receiver's beliefs away from their prior and toward the belief that the state is

[6] Senders may incur an influence cost also from the size of the material loss that different messages can have on the receivers. In our setting, we hold this constant so the expected consequence for the receiver is influenced only by the inaccuracy of the induced beliefs.

Red (as suggested by the message). Consequently, the naive receiver's beliefs following each message are:

$$\begin{cases} \mu_{R^N}(\theta = Red | m \in \{x_1, x_3\}) = \dfrac{11}{20} \\ \mu_{R^N}(\theta = Red | m = x_2) > \dfrac{11}{20} \end{cases}$$

Thus, x_2 has a higher influence cost than x_1 and x_3 since it leads to more inaccurate beliefs in the naive receiver when $\theta = Blue$ and the sender could reveal this truthfully. The messages also differ in terms of the falsehood cost incurred by the sender when the sender has a choice (i.e., when the state is $Blue$ and the sender is informed). Specifically, x_2 and x_3 are both truthful, regardless of the sender's type, while x_1 is true only when the sender is uninformed, according to Definition 2. Therefore, x_1 has the highest falsehood cost. Direct lies incur a greater social image cost than evasions. When the sender lies directly ($m = Red$), the sophisticated receiver will correctly infer the message was deceptive. When the sender evades ($m \in \{x_1, x_2, x_3\}$), neither the sophisticated nor the naive receiver can infer whether the message was truthful even after the payoff realization. Hence, all evasive messages have a lower social image cost than the direct lie. Moreover, all evasive messages as well as the direct lie are equally deceptive when the sender knows that $\theta = Blue$, as the sender could have truthfully revealed this.

We note that when comparing x_1 (IGNORANCE) with x_2 (PARTIAL), the former has a higher falsehood cost but a lower influence cost. To enable a complete ranking of all messages, we assume that the falsehood cost is at least as high as the influence cost.[7] Summing over the different costs for each message, we obtain the following ranking of communication costs for the messages in our framework:

$$C(m = Red) > C(m = x_1) \geq C(m = x_2) \geq C(m = x_3) > C(m = Blue)$$

Given these characteristics, we show (see Appendix A) that the equilibria of this game have the following properties (leading to the following predictions):

1. Only truthful types send the truthful ($m = Blue$) message, to which the receiver responds with $a = Blue$.
2. If there are enough truthful types, both the direct and the evasive messages are equilibrium strategies, to which the receiver responds with the same action, $a = Red$. Hence, the expected payoff to the dishonest sender from both strategies is the same (and equal to h).
3. The lower the communication cost of a message, the more likely a sender is to choose it. Therefore, direct lying is the least likely to occur in equilibrium, followed by ignorance, partial truth and then silence.
4. The more likely receiver will learn if the sender deceived, the lower the deception rate.

[7] This assumption was guided by a pilot survey (described in detail in Appendix C).

3. Experimental Design and Procedures

3.1. Experimental Design

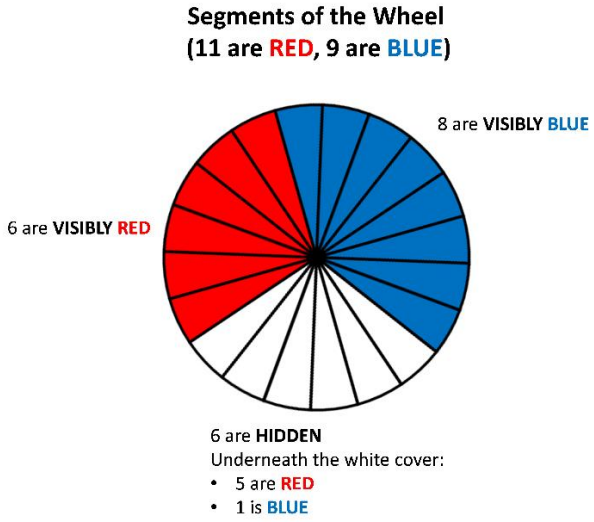
We conducted three experiments. Our empirical strategy mirrors the theoretical framework with all three experiments involving a one-shot interaction between an informed sender and an uninformed receiver.

3.1.1 Senders' behaviour

Experiments 1 (Sender-Hidden) and 2 (Sender-Open) investigated the effect of the communication space on senders' behaviour.

3.1.1.1. The Sender-Hidden experiment. Participants were allocated either the role of sender or receiver. The game structure was common knowledge. The state of the world was determined by using the visual setup depicted in Figure 1. Specifically, a wheel composed of 20 equal segments was spun, and one segment was randomly selected. The colour of this segment could be either Red or Blue, with Red being realized on 11 segments and Blue being realised on the remaining 9 segments.

Figure 1. The 20-segment wheel



As shown in Figure 1, the segment colour could be either visible or hidden. With $\frac{7}{10}$ probability a visible segment was selected so the sender was informed about its colour; and with $\frac{3}{10}$ probability a hidden segment was selected so the sender was uninformed. After the segment was selected, a costless message was sent to the receiver. The message was the only information the receiver obtained. The receiver then guessed whether the segment was Blue or Red. Subsequently, payoffs for both parties were realized, depending on the actual colour of the selected segment and the receiver's guess. The payoff structure is summarized in Table 2 (similar to Table 1, with h = 2 and l = 1). There was a potential

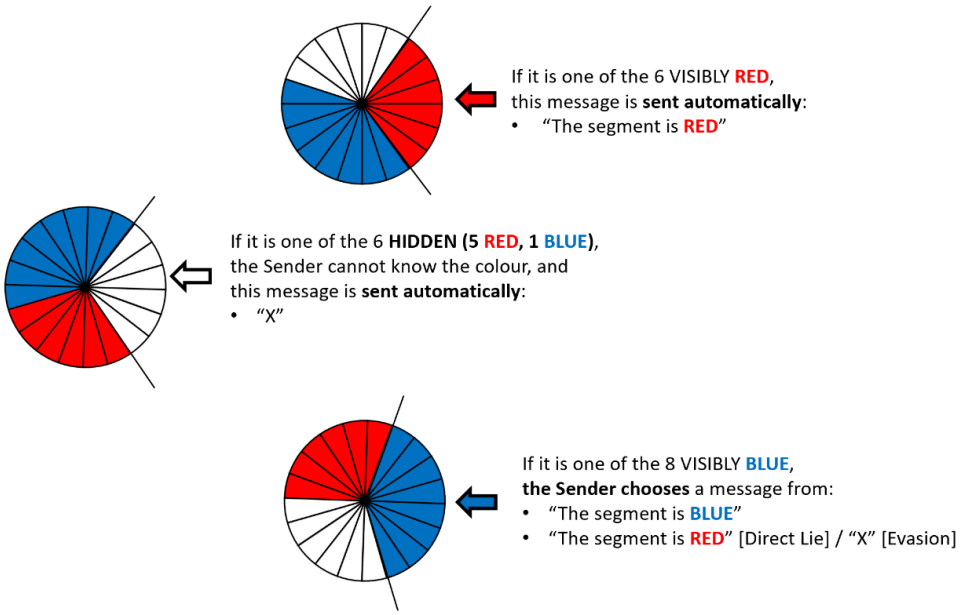
conflict of interest as the sender earned more if the receiver guessed Red, independently of the true state, whereas the receiver earned more if his guess correctly matched the state.

Table 2. Payoff structure of the experimental game (S, R)

		Receiver's guess	
		Red	Blue
Segment	Red	£2, £2	£1, £1
	Blue	£2, £1	£1, £2

To study the psychological cost of deception, we contrasted two decision environments, one comprising a single treatment where participants could lie directly (DIRECT), and one with three evasion treatments (IGNORANCE, PARTIAL, SILENCE). An overview of the structure is shown in Figure 2.

Figure 2. Summary of the decision environments



In both environments, the sender chose which message to send only when the segment was visibly Blue and she therefore had an incentive to deceive.[8] When players' interests were aligned, i.e., the segment was visibly Red, the automatic message "The segment is RED" was sent; if the randomly drawn segment was hidden, another automatic message was sent. This message was one variant of the set X introduced earlier in Section 2 depending on the treatment. The exact messages used in our game are given in Table 3.

[8] Empirical evidence shows the sender almost always (99.3% of the time) sends the truthful option when interests are aligned (Khalmetski et al., 2017).

Table 3. The set X of evasive messages used in the game

Treatment	Message
IGNORANCE	"I don't know the colour of the segment"
PARTIAL	"The segment **was** more likely to be RED than BLUE"
SILENCE	" " (Silence)

When there was a conflict of interest, i.e., the segment was visibly Blue, the DIRECT and evasion treatments diverged. In DIRECT, the sender could tell the truth with the message "The segment is BLUE" or lie directly with "The segment is RED." In the evasion treatments, the sender chose whether to tell the truth with the message "The segment is BLUE" or evade with one of the messages from X depending on the treatment.

The key to our design is that the receiver could not ex-ante distinguish between truth and deception. In DIRECT, when the message "The segment is RED" was received it could be because it was sent automatically when the segment was visibly Red, or because the sender lied directly. In the evasion treatments, when the X message was received it could be because it was sent automatically when the segment was hidden, or because the sender chose to evade. In all treatments, therefore, deception was ex-ante credible.

3.1.1.2. The Sender-Open experiment. An important feature of the Sender-Hidden experiment is that the receiver could infer if they were deceived only in DIRECT. A receiver who got the message "The segment is RED" and followed the recommendation could infer he was deceived since his payoff was £1 instead of the £2. However, evasion was ex-post non-verifiable, since the evasive message came with a positive probability of the segment being Blue, if it was sent automatically from an uninformed sender. As a result, the social image cost of being perceived as a deceiver was higher in DIRECT.

To pin down the role of social image concerns, Sender-Open controlled for the social image cost associated with different deceptive messages. In all treatments, before senders decided which message to send to the receiver, they were informed that, after the receiver made his guess, he would learn if the selected segment was visible or hidden, and therefore if the message was chosen by the sender or sent automatically. Thus, it was highly and equally salient that there would be full revelation of the sender's type. Apart from this, the two experiments were identical.

Note here that a significant part of the senders may had already abstained from deceiving for social image reasons in Sender-Hidden, as their deception was observable by the experimenter, leaving only little room for an effect of social image in Sender-Open. However, the scope of Sender-Open was not to test for the already well-established finding in the dishonesty literature about the existence of social image costs (e.g., Abeler et al., 2019; Bašić and Quercia, 2022; Gneezy et al., 2018; Khalmetski

and Sliwka, 2019; Fries et al., 2021), but to test for differences in deception rates between direct lies and evasion when social image concerns are held constant.

3.1.1.3. Senders' beliefs. Senders' beliefs about how receivers responded to messages are important for identifying the psychological cost of deceptive communications. Senders, for instance, might believe receivers were more likely to choose Red following an evasive message rather than a direct lie, which would then lead them to choose evasions more frequently. To examine whether any observed differences across treatments were driven by differences in these expectations, and not by differences in the psychological cost of communication, we elicited those expectations in an incentivized manner. Each sender estimated the percentage of receivers who guessed Red, after receiving the message that the segment is Blue and the percentage who guessed Red after receiving the alternative (potentially deceptive) message.

It is also well known that people like to adhere to what they believe others will do (e.g., Bicchieri and Xiao, 2009; Colzani et al., 2023; Gächter et al., 2017; Isler and Gächter, 2022; Kimbrough and Vostroknutov, 2016; Kölle and Quercia, 2021; te Velde and Louis, 2022) and, indeed, failing to conform will be an additional psychological cost either for evading or not. To investigate this possibility, senders estimated the percentage of other senders who chose the deceptive message when the segment was visibly Blue to examine whether they were more likely to deceive if they believed others were deceiving too.

In line with Abeler et al. (2019) senders were paid £0.10 per question if their estimates were correct within 3 percentage points. These beliefs were elicited after senders had chosen their message.

3.1.2 Receivers' behaviour

Experiment 3 (Receiver-Hidden) tested the effect of the communication space on receivers' behaviour, to examine the proportion of receivers guessing Red (hereafter called the persuasion rate) and the monetary implications of the different deceptive communications.

3.1.2.1. The Receiver-Hidden experiment. Receiver-Hidden used the design of Sender-Hidden. The only difference was that instead of senders' expectations, we elicited receivers' expectations regarding senders' behaviour as described next. As in Sender-Hidden, the receivers were not informed if the sender was deceiving or telling the truth.

3.1.2.2. Receivers' beliefs. Receivers' beliefs about the likelihood the sender chose the deceptive option are crucial to shed light on whether they believed all deceptive messages were equally informative. For each deceptive message, we elicited receivers' estimates of the percentage of senders who chose the deceptive option when the segment was visibly Blue. To test for adherence to norms, we also elicited estimates of the percentage of other receivers who guessed Red after receiving the deceptive message. As with senders' beliefs, receivers were paid £0.10 per question if their estimate

was correct within 3 percentage points, and their expectations were elicited after they had made their guess.

3.1.3. Discussion of design choices

The specific distribution of segments on the 20-segment wheel was chosen for two reasons. First, it ensured the probability that the deceptive message was sent by a non-deceitful sender was equal across treatments: in 6 out of 14 cases, the Red message was non-deceptive as it was sent by a sender who indeed observed a Red segment, and the evasive message was non-deceptive as it was sent by a sender who observed a hidden segment. Direct lying and evasion were therefore equally credible. This is important, since previous research has shown how increasing the probability of a statement being perceived as true makes the statement more credible, and as such significantly increases lying when the statement is not true (Abeler et al., 2019). Second, the distribution of segments ensured the expected benefit of evasion in equilibrium was not higher than the expected benefit of a direct lie: if senders chose evasion more often than direct lies, it was not because evasion was in expectation more profitable, but because it was less psychologically costly.

In all experimental treatments, we used the strategy method (Selten, 1967). Senders pre-defined which message they wanted to send to the receiver conditional on the segment being visibly Blue. Similarly, receivers guessed the segment's colour conditional on each message they could receive. For Sender-Hidden and Sender-Open, since we focused on senders, we used a matching protocol of ten senders for one receiver to maximize the power of our statistical analysis within our budget (see e.g., Erat and Gneezy, 2012 for a related partial matching protocol). Similarly, for Receiver-Hidden we used a matching protocol of ten receivers for one sender. To use the available resources efficiently, we first collected data for Sender-Hidden, to establish the existence of any difference in psychological costs across the different deceptive communications. We then collected data for Sender-Open and Receiver-Hidden in a sequential order.

To determine the required sample size in each experiment, we conducted a power analysis based on unequal sample sizes between DIRECT and each evasion treatment. This ensured adequate power in the unlikely possibility that the three versions of DIRECT — differing only in the message sent automatically when the sender is uninformed — would differ significantly. In such a case, we could not pool across the three versions of DIRECT and would have to separately compare each version with the corresponding evasion treatment. Our power analysis showed that with 80% power and 5% probability of a type I error, we would need 282 participants in each treatment, to detect a small-to-medium effect size with unequal sample sizes between each version of the DIRECT and the respective evasion treatment.[9] We thus set our target sample to 300 participants in each evasion treatment and 100 in each

[9] Power calculations were conducted using http://powerandsamplesize.com/Calculators/Compare-2-Proportions/2-Sample-Equality. We ran a pilot study to calibrate the incentives in DIRECT, where we found that the deception rate using a high bonus of £2 and a low one of £1 was 25%. We used this number as a guideline for the deception rate in DIRECT for the power analysis. In the actual experiment deception rates were higher.

DIRECT variation. The design, hypotheses and detailed analysis plan were pre-registered via the Open Science Framework and are available at https://osf.io/65hbc/.

As a pre-test, before running our experiments we conducted a pilot survey, where a separate group of participants (N = 201) considered a setting similar to our sender-receiver game. Participants studied a set of possible messages (truth telling, direct lying and various evasions including silence, partial truth and feigned ignorance) and then rated their deceptiveness in case of a conflict of interest on a scale from 1 (Not at all deceptive) to 7 (Very deceptive). Each participant rated all messages: first the truth-telling message, then then direct lie one, then the evasions in a randomized order either from the perspective of the sender, or the receiver.[10] In line with our hypotheses, telling a direct lie was perceived as more deceptive than evading; evasions followed in the order of feigned ignorance, partial truth, and silence; truth telling was the least deceptive (for all paired t-test $p < 0.001$, besides the comparison between silence and partial truth, where $p = 0.001$). Detailed design and results of the pilot survey are reported in Appendix C.

3.2 Experimental Procedures

All experiments were implemented online using samples drawn from Prolific (http://www.prolific.co) and programmed using Qualtrics (http://www.qualtrics.com/). The Humanities and Social Sciences Research Ethics Committee at the University of Warwick reviewed and approved the procedures (18/18-19 for Sender-Hidden and Sender-Open, and 18/18-19 AM01 for Receiver-Hidden). Participants took 12 minutes on average to complete the experiment as sender, and 11 minutes as receiver. Each person participated in only one experimental treatment. We restricted our sample to UK residents with at least 90% past approval rate on Prolific. Participants received a flat fee of £1 for taking part, plus an additional payment ranging from £1 to £3.30 (or £3.20 in Receiver-Hidden) depending on their decisions and the decisions of other participants. The experiments included comprehension questions concerning the instructions, which participants had to answer correctly before proceeding to the main task. We conducted all experiments in two waves: first, we simultaneously collected data from all senders randomly allocated in one of the experimental treatments, and second, we simultaneously collected data from all receivers randomly allocated in one of the experimental treatments. Payoffs to both parties were announced after all responses were received. Experimental instructions are in Appendix D.

4. Hypotheses

We describe the pre-registered hypotheses that are derived from the preceding theoretical framework. The first five hypotheses refer to senders, while the last one refers to receivers.

[10] Deceptiveness judgements are relatively insensitive to the role of the responder (we only find 2/13 differences significant at the 5%, and 1/13 significant at the 10%); therefore, we pool participants' responses irrespective of whether they evaluate a message from the perspective of the sender or the receiver. Results per respondent's type are available on request.

Hypothesis 1: In Sender-Hidden, the deception rate is lowest in DIRECT.

Hypothesis 2: In Sender-Hidden, the deception rate is higher in PARTIAL and SILENCE compared to IGNORANCE.

Hypothesis 3: In Sender-Hidden, the deception rate is higher in SILENCE than in PARTIAL.

We now turn to the effect of social image. There are two plausible hypotheses about its effect depending on the relative costs of the different deceptive communications. If social image costs have no effect, any observed differences in Sender-Hidden should remain in Sender-Open. Otherwise, if any effect observed in Sender-Hidden is completely attributable to differences in the social image cost between DIRECT and the evasion treatments, the deception rate should be indistinguishable across experimental treatments in Sender-Open.

Hypothesis 4a: In Sender-Open, the deception rate is lowest in DIRECT.

Hypothesis 4b: In Sender-Open, the deception rate in DIRECT is equal to the deception rate in any of the evasion treatments.

Lastly, we expect senders to be less likely to deceive in each treatment of Sender-Open, where the receiver is explicitly informed about the sender's potential deception compared to the respective treatment of Sender-Hidden.

Hypothesis 5: The deception rate in Sender-Open is lower than in Sender-Hidden.

Regarding receivers' behaviour, we expect they are more likely to choose Red after receiving the direct lie ("The segment is Red") compared to the alternative deceptive messages in the evasive treatments. This is because we predict there will be enough naive receivers that will take messages at face value and therefore always choose Red after the direct lie.

Hypothesis 6: In Receiver-Hidden, the persuasion rate in DIRECT is higher than that in IGNORANCE, PARTIAL or SILENCE, as well as than the average persuasion rate across the three evasion treatments.

5. Results

In this section we report the experimental results. All hypothesis tests are two tailed, as pre-registered. We first analyse the results focusing on senders (Sender-Hidden and Sender-Open), and then we turn to the receivers (Receiver-Hidden). We also conduct two analyses that are not in our pre-registration. First, we compare the DIRECT treatment and the three evasion treatments pooled. Second, we bring the Sender-Hidden and Receiver-Hidden data together to examine the welfare consequences of the different deceptive communications.

5.1. The Sender-Hidden experiment

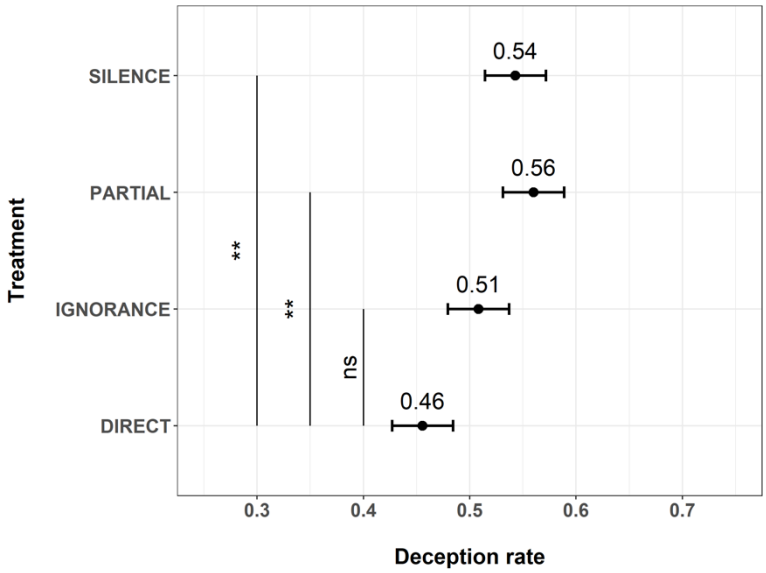
5.1.1. Sample characteristics

The sample consisted of 1,210 participants randomly distributed across the four treatments. Their average age was 36.3, 65% were female, and 86% completed higher education (college or above).[11]

5.1.2. Senders' message choice

Figure 3 presents the choice frequencies for the deceptive message across the four treatments.[12] This was significantly lower in DIRECT than in both PARTIAL ($\chi^2(1, 605) = 6.58$, $p = 0.010$; $d = 0.21$) and SILENCE ($\chi^2(1, 607) = 4.63$, $p = 0.031$; $d = 0.17$). DIRECT and IGNORANCE did not differ significantly ($\chi^2(1, 608) = 1.68$, $p = 0.195$; $d = 0.11$). Overall, however, DIRECT produced the lowest deception rate when pooling over all evasions, ($\chi^2(1, 1210) = 6.04$, $p = 0.014$; $d = 0.16$). Consistent with Hypothesis 1, this suggests the psychological cost of deception was higher via direct lying than evasion.

Figure 3. Deception rate across treatments in Sender-Hidden



*Note*s. The figure depicts the deception rate (x-axis) across treatments (y-axis). Standard errors are plotted as horizontal segments over each frequency (dot). Statistical differences across treatments are depicted with vertical lines accompanied by a statistical significance symbol: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$, ns $p > 0.10$.

[11] Tables B18-B20 in the Appendix depict summary statistics for the sample demographics across treatments for all experiments. There is no evidence that the demographics were unbalanced across treatments but in any case, we controlled for them in the regressions.

[12] Recall that DIRECT used three different versions for the automatic message coming from the uninformed sender (the versions used in the three evasion treatments). These messages were not part of the sender's message choice set in DIRECT, so we did not expect this to affect the sender's decision to deceive. Nevertheless, before analysing this treatment as one, we tested for any effect on the decision to lie coming from the type of automatic message associated with the uninformed sender. A Chi-square test comparing the deception rate across the three versions of DIRECT revealed no significant differences ($\chi^2(2, 305) = 2.41$, $p = 0.300$). For the rest of the analysis, in line with our pre-registration, we pooled across the three versions of DIRECT and treated them as a unitary set of observations.

We complement this analysis with a probit regression where the dependent variable is the decision to choose the deceptive option and the main independent variables are the experimental treatments. The regression results are presented in Table 4. Column (1) presents the main, pre-registered model comparing each evasion treatment with DIRECT where, we control for beliefs and demographics. In Appendix B, Table B27, we present the analysis without controlling for beliefs and show that the conclusions remain unchanged. Column (2) presents the comparison of DIRECT with all evasion treatments pooled.

Table 4. Probit analysis of choosing the deceptive option in Sender-Hidden

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.036	
	(0.045)	
PARTIAL	0.130***	
	(0.044)	
SILENCE	0.125***	
	(0.044)	
EVASIONS_Pooled		0.099***
		(0.037)
B(a=Red|m=non-Blue)	0.001	0.001
	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.001	-0.000
	(0.001)	(0.001)
B(others-deceive)	0.009***	0.009***
	(0.001)	(0.001)
Female	-0.078**	-0.078**
	(0.033)	(0.033)
Age	0.002	0.002
	(0.001)	(0.001)
Higher education	0.050	0.053
	(0.046)	(0.046)
Observations	1,193	1,193

Notes: Marginal effects from a probit regression in Sender-Hidden. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). IGNORANCE, PARTIAL and SILENCE are dummies for those treatments, DIRECT is the excluded category. B(·) are the sender's beliefs. "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Consistent with the findings just reported, senders were 13 percentage points more likely to choose the deceptive option in PARTIAL compared to DIRECT ($p = 0.003$) and 12.5 percentage points more likely to do so in SILENCE compared to DIRECT ($p = 0.005$). The 3.6 percentage points difference between IGNORANCE and DIRECT was not significant ($p = 0.430$), consistent with our view that IGNORANCE has a higher falsehood cost than the other evasions.

> **Result 1.** When evasion was non-verifiable, the deception rate in DIRECT was lower than in SILENCE or PARTIAL, while the rates did not differ between DIRECT and IGNORANCE.

The belief regarding whether other senders would deceive had a significant positive effect on the likelihood of choosing the deceptive option ($p < 0.001$) in line with a desire to conform to what others do. There was also a significant gender effect ($p = 0.017$), with females being less likely than males to choose the deceptive option, but this effect was not found in our other experiments.

We next compare how often the deceptive option was chosen across the three evasion treatments. Based on our pre-registered non-parametric analysis, we find no support for Hypotheses 2 and 3 (focusing on differences between the evasion treatments) as the proportion choosing the deceptive option did not significantly differ across any of these pairwise comparisons. Using the Chi-square test, the deception rate in IGNORANCE was not significantly different from PARTIAL ($\chi^2(1, 603) = 1.62$, $p = 0.203$) or SILENCE ($\chi^2(1, 605) = 0.74$, $p = 0.391$). Similarly, PARTIAL was statistically indistinguishable from SILENCE ($\chi^2(1, 602) = 0.18$, $p = 0.68$).

> **Result 2.** When evasion was non-verifiable, the proportion of senders choosing the deceptive option did not significantly differ across the three evasive treatments.

However, this does not necessarily mean that the three evasions were similarly psychologically costly since the differences between evasions with respect to the DIRECT treatment were fairly sizable after controlling for potential individual heterogeneity due to beliefs and demographics as can be seen in Table 4. Indeed, when comparing the coefficients in Column (1) of Table 4 for the evasion treatments we found that the coefficient of IGNORANCE was significantly smaller than that of both PARTIAL ($\chi^2(1, 1183) = 4.51$, $p = 0.034$) and SILENCE ($\chi^2(1, 1183) = 4.24$, $p = 0.039$). These differences become significant due to heterogeneity in beliefs about how likely others were to deceive – a variable that varied largely with the decision to send the deceptive message (see distribution of beliefs conditional on message choice in Figure B1 and corresponding statistical analysis in Tables B9-B11 in Appendix B). When we do not control for this variable in the probit regression, the differences between the coefficients of the evasion treatments are no longer significant despite directional similarities (see Table B30 and subsequent linear hypothesis tests in Appendix B). Overall, this suggests that the falsehood cost is potentially larger than the influence cost in our setting.

5.1.3. Senders' beliefs

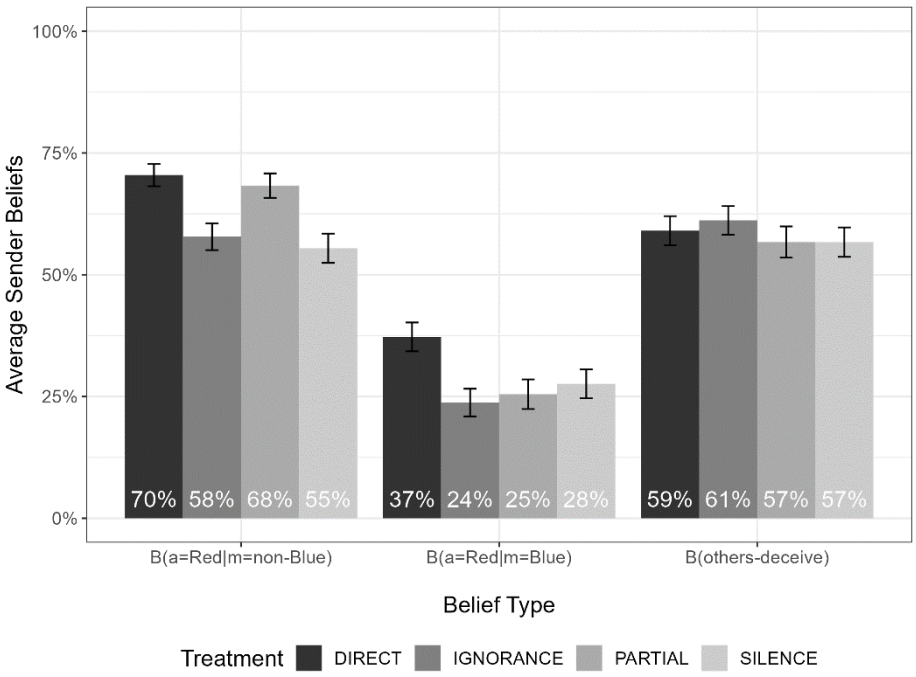
Figure 4 presents average sender beliefs across the four treatments, for all senders, irrespective of the sender's choice (deceptive or truthful) (see Appendix B for the analysis of belief distributions across treatments and decisions as well as the results of pairwise comparisons).

First, we find that messages differed in their judged effectiveness (B(a=Red|m=non-Blue), H(3) = 79.07, $p < 0.001$). In particular, senders believed that receivers were significantly less likely to choose Red after the IGNORANCE and SILENCE message compared to the DIRECT and PARTIAL message.

> **Result 3.** When evasion was non-verifiable, senders believed that receivers were more likely to act in senders' favour when the message was a direct lie or a partial truth than when keeping silent or feigning ignorance.

This supports our view that the influence cost of deception was lower in IGNORANCE and SILENCE than in PARTIAL or DIRECT. It also suggests that the higher likelihood to deceive in the evasion treatments was not due to a higher perceived expected benefit since senders believed receivers were *less* likely to choose their most preferred action in those treatments.

Figure 4. Average sender beliefs across treatments in Sender-Hidden



*Note*s. The figure depicts the mean reported sender belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar). B() indicates beliefs.

Surprisingly, we also find that treatments differed significantly in senders' belief about the likelihood the receiver chose Red after being honestly told the state was Blue (B(a=Red|m=Blue), H(3) = 56.05, $p < 0.001$). Although not central to our questions, we believe this could be due to several reasons including the fact that senders judged receivers as being more likely to reward truthful senders

in DIRECT, noise or order effects since we always elicited these beliefs after the ones regarding m=non-Blue, which were of primary interest.

However, this does not mean it was more advantageous to tell the truth in DIRECT than in the evasion treatments. In fact, in all treatments, the average judged likelihood of receivers choosing Red after the deceptive message was significantly higher than after the truthful Blue message (DIRECT: $t(304) = 16.62$, $p < 0.001$; IGNORANCE: $t(302) = 18.70$, $p < 0.001$; PARTIAL: $t(299) = 19.38$, $p < 0.001$; SILENCE: $t(301) = 13.57$, $p < 0.001$). According to their beliefs, senders should always make the deceptive choice to maximise their monetary earnings. Since this is not what we observed in the data, it appears that deception incurs psychological costs so that many senders are willing to earn less to avoid it.

Finally, we find no difference across treatments in the senders' beliefs about the likelihood that other senders would deceive (B(others-deceive), $H(3) = 5.47$, $p = 0.140$). Participants estimated that about 60% of people would deceive, which is an overestimate (it was closer to 50%) but not an extreme one.

5.2. The Sender-Open experiment

5.2.1. Sample characteristics

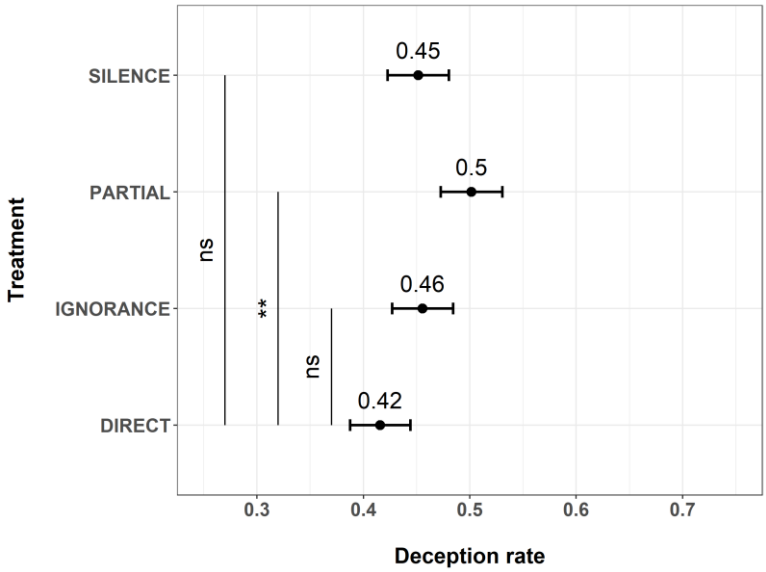
The sample consisted of 1,204 participants randomly distributed across the four treatments. Their average age was 36.6 years, 63% were female, and 88% completed higher education (college or above).

5.2.2. Senders' message choice

Figure 5 displays the deception rates.[13] As in Sender-Hidden, the lowest deception rate was in DIRECT. The pattern in the comparisons between DIRECT and the evasion treatments was also similar to Sender-Hidden with one exception: the deception rate in SILENCE was no longer significantly higher than that in DIRECT ($\chi^2(1, 602) = 0.78$, $p = 0.377$; $d = 0.07$). The deception rate in PARTIAL remained significantly higher than in DIRECT ($\chi^2(1, 600) = 4.45$, $p = 0.035$; $d = 0.17$), while IGNORANCE remained statistically indistinguishable from DIRECT ($\chi^2(1, 608) = 0.98$, $p = 0.321$; $d = 0.08$).

[13] As in Sender-Hidden, in DIRECT we used three different versions for the automatic message coming from the uninformed sender (the versions used in the three evasion treatments). Before analysing this treatment as one, we tested for any effect on the decision to lie coming from the specific automatic message associated with the uninformed sender. A Chi-square test comparing the deception rate across the three versions of DIRECT suggested no significant differences ($\chi^2(2, 303) = 1.87$, $p = 0.393$). We therefore pooled across the three versions of this treatment and analysed them as a unitary set of observations for our main hypothesis testing.

Figure 5. Deception rate across treatments in Sender-Open



*Note*s. The figure depicts the deception rate (x-axis) across treatments (y-axis). Standard errors are plotted as horizontal segments over each frequency (dot). Statistical differences across treatments are depicted with vertical lines accompanied by a statistical significance symbol: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$, ns $p > 0.10$.

We proceed with a probit analysis predicting choice of the deceptive option from experimental treatment and control variables. The results reported in Table 5 corroborate the main findings. After controlling for beliefs and demographic characteristics (Table 5, Column 1), senders were 10.8 percentage points more likely to choose the deceptive option in PARTIAL compared to DIRECT and this difference was statistically significant ($p = 0.020$). There were no significant effects of IGNORANCE and SILENCE. As in Sender-Hidden, beliefs about how likely others were to deceive had a significant positive effect on choice of the deceptive option ($p < 0.001$), suggesting a role for conformity. We also find a negative effect of beliefs about how likely receivers were to choose Red after receiving the message that the segment is Blue ($p = 0.039$).

> **Result 4.** When evasion was verifiable, the deception rate in DIRECT was significantly lower than in PARTIAL, but it did not significantly differ from the ones in IGNORANCE and SILENCE.

Result 4 is contrary to Hypothesis 4b, according to which we should not observe any difference between DIRECT and the evasion treatments once we control for differences in the social image costs associated with the different deceptive messages. Such an outcome would imply that the differences observed in Sender-Hidden (Result 1) were only due to differences in social image costs. Result 4 instead suggests that the psychological cost of deception was lower for partial truth than for direct lying even after controlling for social image concerns. We therefore find partial support for Hypothesis 4a.

Table 5. Probit analysis of choosing the deceptive option in Sender-Open

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	-0.007	
	(0.047)	
PARTIAL	0.108**	
	(0.047)	
SILENCE	0.017	
	(0.047)	
EVASIONS_Pooled		0.041
		(0.038)
B(a=Red|m=non-Blue)	0.000	0.000
	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.001**	-0.001**
	(0.001)	(0.001)
B(others-deceive)	0.012***	0.012***
	(0.001)	(0.001)
Female	0.013	0.015
	(0.033)	(0.033)
Age	0.001	0.001
	(0.001)	(0.001)
Higher education	-0.001	-0.001
	(0.050)	(0.050)
Observations	1,188	1,188

Notes: Marginal effects from a probit regression in Sender-Open. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). IGNORANCE, PARTIAL and SILENCE are dummies for those treatments, DIRECT is the excluded category. B(·) are the sender's beliefs. "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Next, we conduct pairwise comparisons of the evasion treatments to investigate whether they differ when evasion is verifiable. Using the pre-registered non-parametric analysis, they do not. Without controlling for beliefs or demographic characteristics, the deception rate in IGNORANCE was statistically indistinguishable from that in PARTIAL ($\chi^2(1, 602) = 1.27$, $p = 0.259$) and SILENCE ($\chi^2(1, 604) = 0.01$, $p = 0.917$), while deception rates in PARTIAL and SILENCE did not differ either ($\chi^2(1, 596) = 1.50$, $p = 0.220$), a result in line with the findings in Sender-Hidden (Result 2).

Result 5. When evasion was verifiable, the deception rate did not significantly differ across the three evasion treatments.

However, as in Sender-Hidden, this does not necessarily mean that the three evasions were similarly costly since the heterogeneity driven by beliefs and demographics may reduce these differences. Indeed, when controlling for this heterogeneity by comparing the coefficients in Column (1) of Table 5 for the evasion treatments we find that the coefficient of IGNORANCE was significantly smaller than that of PARTIAL ($\chi^2(1, 1178) = 6.28$, $p = 0.012$) but not different from that of SILENCE ($\chi^2(1, 1178) = 0.27$, $p = 0.601$). The coefficient of PARTIAL was also significantly larger than that of SILENCE ($\chi^2(1, 1178) = 3.90$, $p = 0.048$).[14] Since PARTIAL has a higher influence cost than IGNORANCE but a lower falsehood cost, as in Sender-Hidden, these findings suggest that the falsehood cost of IGNORANCE outweighs the influence cost of PARTIAL, or that the social image cost of IGNORANCE exceeds that of PARTIAL. The latter explanation holds also for the difference between PARTIAL and SILENCE, since these two evasions do not differ in terms of deception and falsehood costs, but PARTIAL has a higher influence cost than SILENCE. This means that the social image cost may vary across evasions.

Next, we test Hypothesis 5 by comparing the average deception rate in each treatment of Sender-Open, with that in the corresponding treatment of Sender-Hidden. The deception rates across all treatments are lower in Sender-Open than in Sender-Hidden, although only for SILENCE is the pairwise comparison statistically significant – 45% vs 54% ($\chi^2(1, 601) = 5.04$, $p = 0.025$; $d = 0.18$). This result is confirmed by a probit analysis controlling for sender's beliefs and demographics which suggests making evasion verifiable in SILENCE decreases the deception rate with 14.66 percentage points ($p = 0.001$).[15] An overall analysis shows that the deception rate significantly decreases with 8.5 percentage points in Sender-Open compared to Sender-Hidden ($p < 0.001$, see Table B13 in Appendix B for the full regression table).[16]

Result 6. The deception rate was lower when evasion was verifiable than when it was not.

5.2.3. Senders' beliefs

Figure 6 presents average sender beliefs across the four treatments in Sender-Open, for all senders, irrespective of the sender's choice (see Appendix B for the analysis of belief distributions

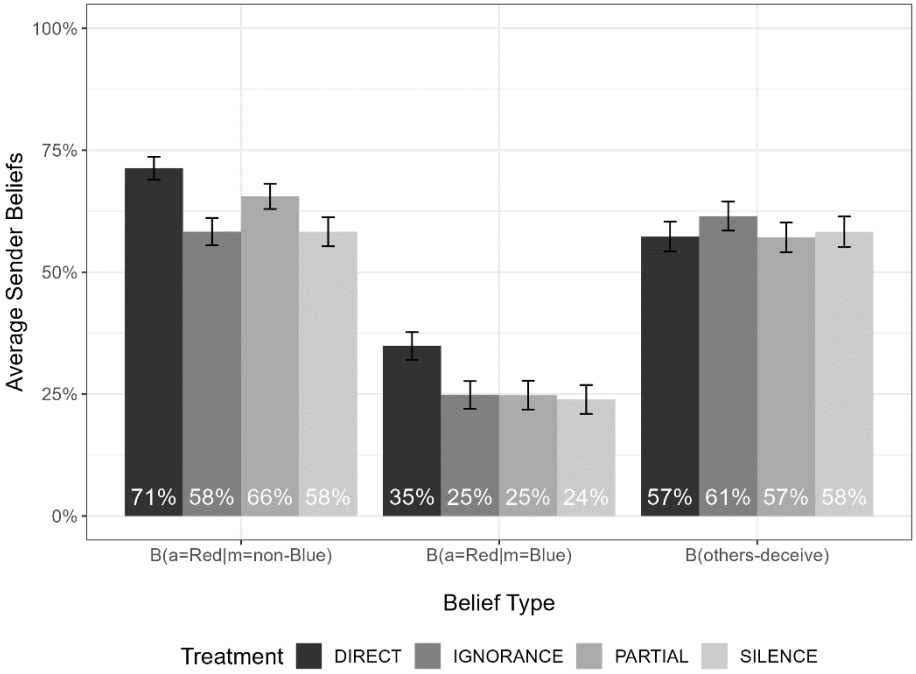
[14] Like in Sender-Hidden, the main source of heterogeneity is represented by senders' beliefs about how likely others are to deceive (B(others-deceive)) as the difference in coefficients becomes insignificant when we no longer control for this variable in the probit regression (see Table B31 and corresponding linear hypothesis analysis in Appendix B).

[15] Result 6 should be interpreted with caution as Sender-Hidden and Sender-Open were not conducted simultaneously. Sender-Hidden was run first, to investigate whether evasion is less psychologically costly than direct lying while social image costs are not equal. After finding support for our main hypothesis, we ran Sender-Open, 7 weeks after, to isolate the role of social image (since this only made sense if differences were observed in Sender-Hidden). Nevertheless, to enhance comparability, we held constant the day of the week and time of day data were collected. Moreover, the demographics do not differ significantly across experiments (Age: $H(2373) = 0.06$, $p = 0.813$; Female: $\chi^2(1, 2411) = 0.49$, $p = 0.483$; Higher education: $\chi^2(1, 2397) = 1.31$, $p = 0.253$).

[16] Note this overall analysis was not pre-registered.

across treatments and decisions as well as the results of pairwise comparisons). The belief distribution is very similar to that in Sender-Hidden.

Figure 6. Average sender beliefs across treatments in Sender-Open



*Note*s. The figure depicts the mean reported sender belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar).

As in that experiment, we find significant differences across treatments with respect to the average sender's beliefs about the likelihood the receivers chose Red after the deceptive ("non-Blue") message ($H(3) = 60.31$, $p < 0.001$). The direction of these differences is in line with that found in Sender-Hidden which further strengthens the robustness of Result 3.

> **Result 7.** When evasion was verifiable, senders believed that receivers were more likely to choose the action implied by the message when the message was a direct lie or a partial truth than when keeping silent or feigning ignorance.

Furthermore, the treatments differed again also in the sender's beliefs that the receiver would choose Red if they received the Blue message (B(a=Red|m=Blue), ($H(3) = 46.58$, $p < 0.001$), with senders judging receivers as being more likely to reward truthful senders (i.e., to choose a=Red despite being honestly told the state was Blue) in DIRECT.

Nevertheless, senders still believed that the deceptive option was more profitable than the truthful one also when evasion could be verified, since the average judged likelihood that the receiver will choose Red after the deceptive message was always significantly higher than after the non-deceptive one (DIRECT: $t(302) = 18.53$, $p < 0.001$; IGNORANCE: $t(304) = 17.81$, $p < 0.001$; PARTIAL: $t(296) = 18.89$, $p < 0.001$; SILENCE: $t(298) = 17.74$, $p < 0.001$).

Finally, we again find no difference across treatments in the sender's beliefs about the likelihood that other senders would deceive (B(others-deceive), H(3) = 4.80, $p = 0.187$), with senders estimating that about 60% of receivers would deceive, a modest overestimate.

5.3. The Receiver-Hidden experiment

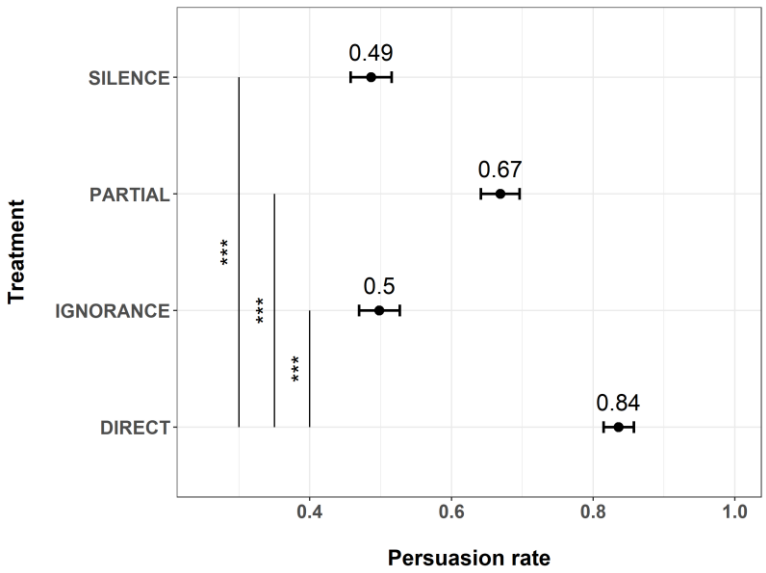
5.3.1. Sample characteristics

The sample consisted of 1,201 participants randomly distributed across the four treatments. Their average age was 40.4, 49% were female, and 87% completed higher education (college or above).

5.3.2. Receivers' guess

The proportion of receivers guessing Red, which is the senders' favourable option, is depicted in Figure 7. DIRECT had a much higher persuasion rate than all other treatments. In particular, DIRECT had a significantly higher persuasion rate compared to IGNORANCE ($\chi^2(1, 602) = 77.20$, $p < 0.001$; $d = 0.77$), PARTIAL ($\chi^2(1, 598) = 22.45$, $p < 0.001$; $d = 0.39$) and SILENCE ($\chi^2(1, 599) = 81.62$, $p < 0.001$; $d = 0.79$).

Figure 7. Persuasion rate across treatments in Receiver-Hidden



*Note*s. The figure depicts the persuasion rate (x-axis) across treatments (y-axis). Standard errors are plotted as horizontal segments over each frequency (dot). Statistical differences across treatments are depicted with vertical lines accompanied by a statistical significance symbol: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$, ns $p > 0.10$.

We confirm this first order analysis with a probit regression. The marginal effects from this analysis predicting the persuasion rate from experimental treatments and control variables are shown in Table 6. The results corroborate our main findings. Compared to DIRECT, receivers were 23.2 percentage points less likely to guess Red in IGNORANCE ($p < 0.001$), 12.7 percentage points less likely to do so in PARTIAL ($p = 0.010$), and 25.6 percentage points less likely to do so in SILENCE ($p < 0.001$). Beliefs about the behaviour of other receivers had a significant positive effect on the likelihood

of guessing Red ($p < 0.001$), while beliefs about the percentage of senders who chose to deceive had a significant negative effect ($p < 0.001$).

Result 8. The persuasion rate is higher in DIRECT than in all evasion treatments.

Table 6. Probit analysis of Persuasion Rate in Receiver-Hidden

	Dependent variable:	
	Guess RED	
	(1)	(2)
IGNORANCE	-0.230***	
	(0.049)	
PARTIAL	-0.125**	
	(0.049)	
SILENCE	-0.253***	
	(0.048)	
EVASIONS_Pooled		-0.182***
		(0.034)
B(a=Red|m=non-Blue)	0.003***	0.010***
	(0.001)	(0.001)
B(S-deceives)	-0.003***	-0.003***
	(0.001)	(0.001)
Female	0.005	0.003
	(0.031)	(0.031)
Age	-0.000	-0.000
	(0.001)	(0.001)
Higher education	-0.020	-0.027
	(0.046)	(0.046)
Observations	1,188	1,188

Notes: Marginal effects from a probit regression in Receiver-Hidden. The dependent variable is whether the receiver guessed RED (1 if yes, 0 if not). IGNORANCE, PARTIAL and SILENCE are dummies for those treatments, DIRECT is the excluded category. B(·) are the receiver's beliefs. Column (1) reports the regression without demographic controls, Column (2) with demographic controls, where "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Surprisingly, when we compare the persuasion rate in the three evasion treatments, we find that receivers were more likely to guess Red in PARTIAL than in IGNORANCE ($\chi^2(1, 602) = 18.01$, $p < 0.001$; $d = 0.35$) and SILENCE ($\chi^2(1, 599) = 20.38$, $p < 0.001$; $d = 0.37$), while IGNORANCE and SILENCE were statistically indistinguishable ($\chi^2(1, 603) = 0.08$, $p = 0.774$, $d = 0.02$). This suggests that the receivers were more likely to be deceived when the sender more strongly advises a guess of

Red. These results are confirmed when comparing the evasion treatments using the probit analysis shown in Table 6.[17]

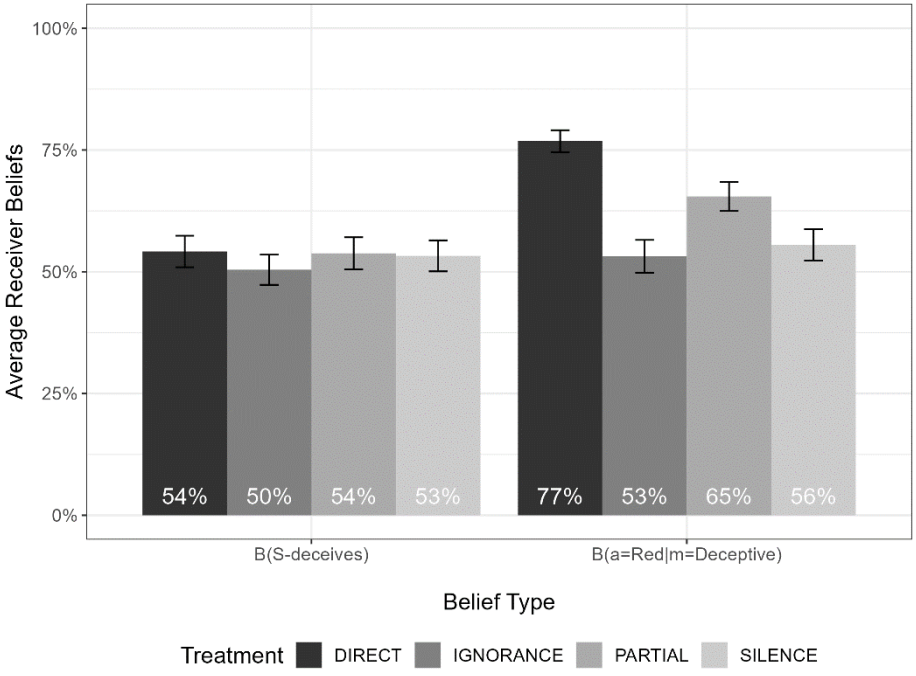
Result 9. The persuasion rate was significantly higher in PARTIAL compared to IGNORANCE and SILENCE.

Taken together, results 8 and 9 suggest that a significant proportion of receivers interpreted messages at face value and followed the senders' recommendation as the naive type of receiver would do in our theoretical analysis. This also implies that the higher influence costs senders may have associated with the DIRECT and PARTIAL messages were justified.

5.3.2. Receivers' beliefs

Recall that we elicited two types of beliefs from receivers: (1) about the likelihood that senders would choose the deceptive message when they observed the Blue segment (B(S-deceives)), and (2) about the likelihood that the other receivers would choose Red when receiving the potentially deceptive message which varied across treatments (B(a=Red|m=Deceptive)). The average receiver beliefs in Receiver-Hidden are presented in Figure 7.

Figure 7. Average receiver beliefs across treatments in Receiver-Hidden



[17] In particular, PARTIAL was statistically different both from IGNORANCE (without ($\chi^2(1, 1195) = 5.48$, $p = 0.019$) and with controlling for demographic characteristics ($\chi^2(1, 1179) = 5.33$, $p = 0.021$)), and from SILENCE (without ($\chi^2(1, 1195) = 8.41$, $p = 0.004$) and with controlling for demographic characteristics ($\chi^2(1, 1179) = 8.16$, $p = 0.004$)). IGNORANCE was statistically indistinguishable from SILENCE both without ($\chi^2(1, 1195) = 0.30$, $p = 0.584$) and with demographic controls ($\chi^2(1, 1179) = 8.28$, $p = 0.596$).

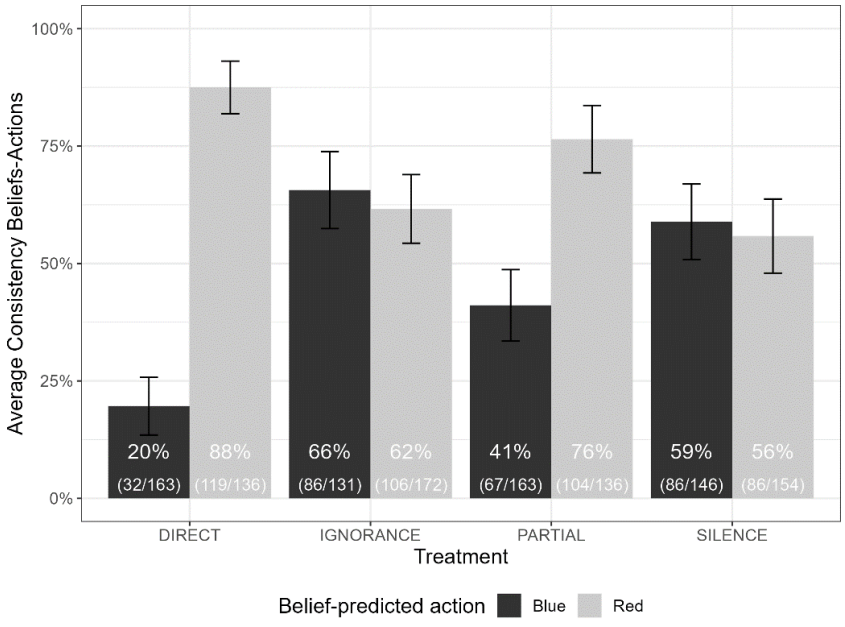
*Note*s. The figure depicts the mean reported receiver belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar).

First, we find no difference across treatments in the receivers' beliefs about the likelihood that the senders would deceive ($H(3) = 4.123$ $p = 0.248$).

Result 10. Receivers believed that senders were equally likely to deceive across all treatments.

According to their beliefs, receivers should guess Red in the same rate across treatments. This is not what we observe in our data which suggests receivers' guessing decisions are not (solely) driven by whether they think the message is truthful. Indeed, when we look at how often a belief indicating the receiver thinks the message is more likely truthful (B(S-deceives) ≤ 50%) was associated with that receiver guessing Red after a potentially deceptive message and vice-versa, we find significant differences across treatments. In particular, the consistency rate between belief-predicted guesses and actual guesses in DIRECT was 50.5%, lower than the rate in IGNORANCE (63.4%), PARTIAL (57.2%) and SILENCE (57.3%).[18] Given the simple structure of the game, the most likely driver of receivers' behaviour, other than beliefs, is the strength of the recommendation implied by the face value of the message. We find some evidence for this channel when disaggregating the consistency of beliefs and actions based on whether receivers' beliefs would guess Blue or Red guess (see Figure 8).

Figure 8. Average receiver beliefs across treatments in Receiver-Hidden



[18] The difference between DIRECT and IGNORANCE was significant, between DIRECT and SILENCE weakly significant and between DIRECT and PARTIAL not significant, whereas differences between the evasion treatments were not significant (see Appendix B, Table B17, for the results of the pairwise comparisons using Chi-squared tests). However, the differences between DIRECT-PARTIAL and DIRECT-SILENCE increased and became significant when considering beliefs strictly lower than 50% to be consistent with a guess of Red as that lowered the consistency rate in DIRECT to 45.2% without affecting much the rate in IGNORANCE (62.4%), PARTIAL (54.5%) or SILENCE (55.7%).

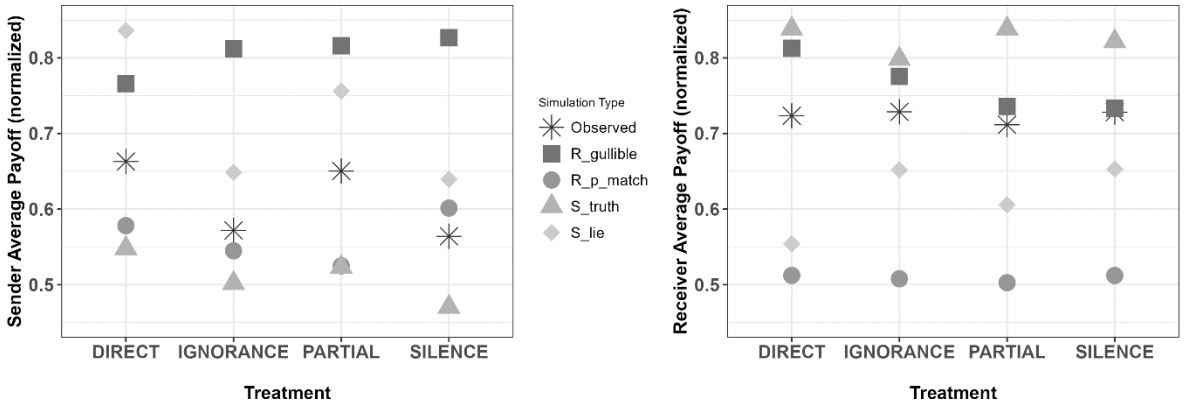
The striking finding from this disaggregation is that receivers' decision to guess Blue was much less predictable by their beliefs about senders' truthfulness rather than the decision to guess Red, in DIRECT (χ^2 (1, 299) = 137.42, $p < 0.01$) and PARTIAL (χ^2 (1, 299) = 41.23, $p < 0.01$), whereas no such difference was found in IGNORANCE (χ^2 (1, 303) = 0.18, $p = 0.67$) and SILENCE (χ^2 (1, 300) = 0.03, $p = 0.87$). This discrepancy was not due to participants simply reporting beliefs consistent with their actions, since that would result in a higher number of receivers for whom the belief-predicted action was Red rather Blue in DIRECT and PARTIAL. Yet, we see that the opposite is true: in both of these treatments the number of receivers for whom the belief-predicted action is Red (N = 136), is lower than that for whom it is Blue (N = 163). We interpret this as suggestive evidence for the effect that the senders' message has on the weight receivers assign to their beliefs when making a guess: the clearer the recommendation for a guess is in a message, the lower the weight the receiver puts on their belief and the higher the weight he puts on the face value of the message. When no such direct recommendation is explicit in the message, like in IGNORANCE and SILENCE, beliefs may be more salient for receiver's decision-making process, and in our data, they are predicting equally well both a Blue and a Red guess.

With respect to beliefs about how other receivers behave, receivers seem to be (correctly) projecting this type of naivety on the other participants in their role, as their estimates about how likely other receivers are to guess Red after the deceptive message differed across treatments (H(3) = 128.62, $p < 0.001$). In particular, receivers believed other receivers were more likely to guess Red in DIRECT and PARTIAL compared to IGNORANCE and SILENCE (see Appendix B for the results of the pairwise comparisons). This pattern was similar irrespective of whether the receiver guessed Red or Blue, though the levels were lower in the latter case (see Figure B3 in Appendix B). This is to be expected, since the receivers are predicting others will do what they themselves are doing (see e.g., Ross et al., 1977 for a seminar paper on the consensus effect, but also Dawes, 1989; Engelmann and Strobel, 2000; Vanberg, 2019; for related discussions on its rationality).

6. Welfare analysis

What are the welfare consequences of evasion? Although our experiments were not expressly designed to answer this question, our data can, nevertheless, reveal some new insights. We investigated this by simulating matches between senders from Sender-Hidden and receivers from Receiver-Hidden and calculate average *potential* payoffs under a number of theoretically relevant scenarios. To maximise accuracy, we paired each receiver with all senders (approximately 300 matches for each receiver in each treatment). Figure 9 presents the average simulated payoffs for senders and receivers across treatments.

Figure 9. Simulated average payoffs for senders (left panel) and receivers (right panel)



*Note*s. The figure depicts simulated average payoffs (y-axis) for each treatment (x-axis). Payoffs are normalized to 0 (equivalent to £1 in the experiment) and 1 (equivalent to £2 in the experiment).

The "Observed" type includes the expected payoffs given the observed decisions of senders and receivers in our experiments. We also included four other benchmarks where we simulated alternative decisions on the part of the sender or the receiver. The "R_p_match" type is based on the observed decisions of the senders but assumes receivers ignore the message and choose Red 11 out of 20 times. This is one of a family of possible ways the receiver could respond while ignoring the message, and corresponds to "probability matching," a well-established tendency for people to predict probabilistic outcomes by matching event probabilities (e.g., Koehler and James, 2010; Vulcan, 2000). Probability matching is by no means the only way receivers could ignore messages, but it is a useful baseline and more natural than assuming receivers are simply choosing randomly.

The "R_gullible" type uses the observed decisions of the senders but assumes receivers are benevolently trusting, i.e. they always guess Red unless the message is that the segment is Blue. We call this "gullible" because this is what receivers would be expected to do if the message were legitimate, meaning the segment were indeed Red, or hidden. The R_gullible type would choose Red even when receiving the evasive message because the probability of a Red segment is 5/6 given the evasive message when the segment is hidden.

The "S_truth" type assumes senders always tell the truth while receivers respond as observed. This reflects the *potential* cost of scepticism or disbelief. Finally, the "S_lie" type is similar to the "S_truth" type except that it assumes that senders always lie. It therefore reflects the potential cost of trusting behaviour on the receiver side.

First, we note that there is a significant amount of information sent and received since both players' simulated payoffs given the observed choices (stars in Figure 9) across all treatments are significantly higher than what they would be if receivers were ignoring messages (circles in Figure 9). Second, and not surprisingly, if senders were always truthful (triangles in Figure 9), receivers would benefit most, while senders would be significantly harmed. This analysis also suggests that, across all

treatments, welfare could be increased if receivers were less sceptical. This is particularly true in the DIRECT and IGNORANCE treatment.[19]

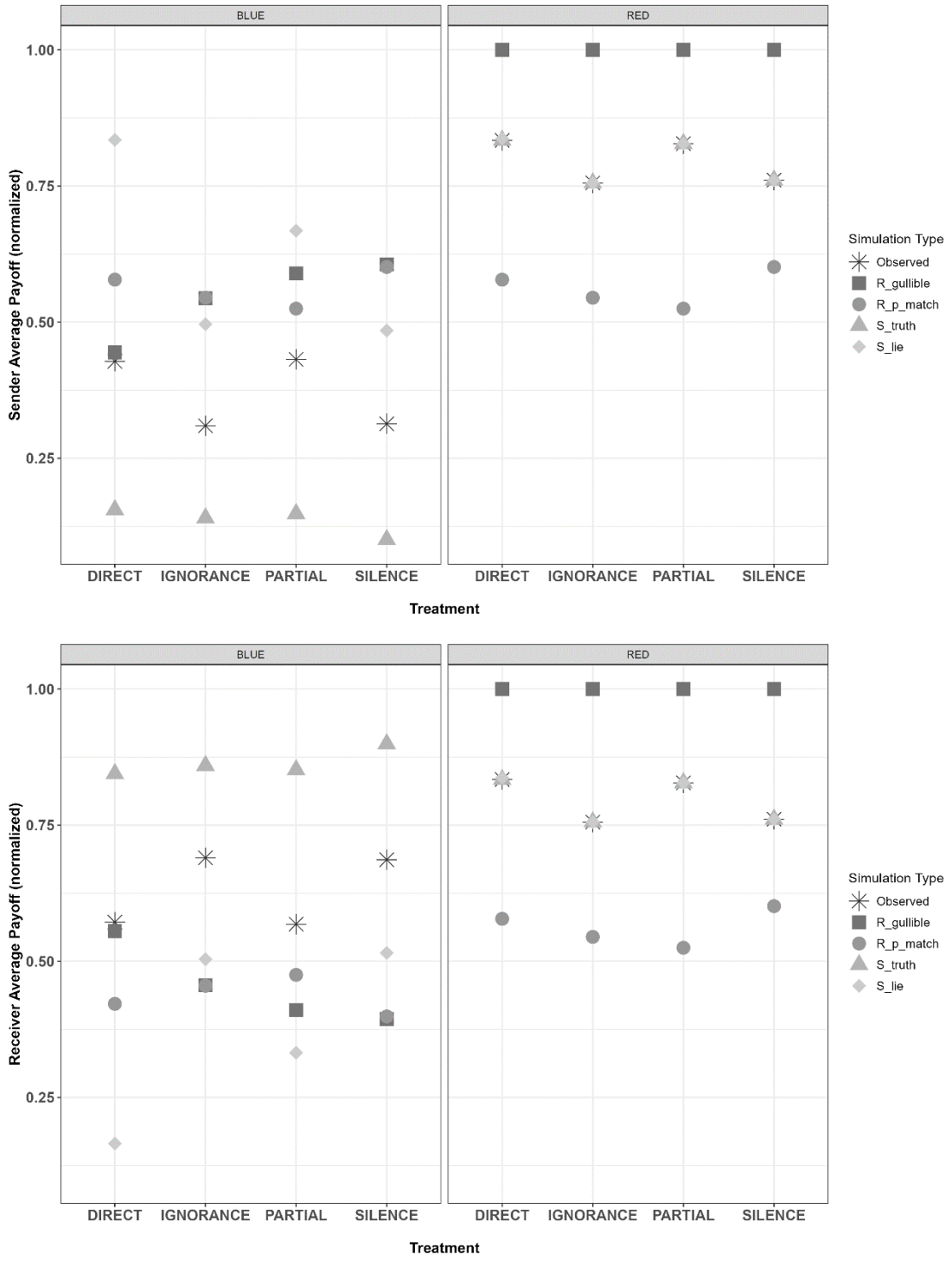
Focusing on the "Observed" type payoffs, we find that overall, receivers are not hurt by evasion, but senders do significantly worse in IGNORANCE and SILENCE compared to DIRECT and PARTIAL (see Table B24 in Appendix B). The implication is that the former two types of evasion are the most likely to decrease overall welfare. We note, though, that receivers' two types of guessing errors (guessing Red when the segment was Blue – a false negative – or Blue when the segment was Red – a false positive) were equally costly in our setting. In many other situations, this may not be the case. For example, it may be costlier to buy a house with underlying issues that remain hidden at the time of the contract than to forego a good deal. Similarly, it may be costlier to convict an innocent person than to absolve a guilty one.

To get some insights about what might happen in such instances and better understand the scale of the costs associated with the two types of errors, we disentangled the expected payoffs between cases where the drawn colour of the segment was Red, and hence interests were aligned, and cases where the colour was Blue and interests were mis-aligned (see Figure 10).

From this decomposition we learn first, that because receivers are less likely to guess Red in the IGNORANCE and SILENCE treatments. When interests are aligned both players do significantly worse in these treatments compared to DIRECT and PARTIAL (see Table B26 in Appendix B). But increasing this likelihood (or reducing scepticism) is not a panacea as this would significantly harm receivers in all evasion treatments when interests are misaligned. In these situations, the receiver was actually worse off in DIRECT and PARTIAL where he was more likely to incorrectly guess Red (see Table B25 in Appendix B). Taken together, this exploratory analysis suggests that evasion can be materially harmful for both receivers and senders. While the material consequences of evasion for senders may be compensated by the lower psychological costs, receivers are unlikely to benefit from any such psychological gains when being deceived through evasion. And, of course, senders receive even lower psychological costs if they choose not to deceive at all.

[19] All statistical test results are presented in Tables B22 and B23 in Appendix B.

Figure 10. Simulated average payoffs for senders (top panel) and receivers (lower panel) by state



*Note*s. The figure depicts simulated average payoffs (y-axis) for each treatment (x-axis) separate for cases when the segment's colour was BLUE and for when it was RED. Payoffs are normalized to 0 (equivalent to £1 in the experiment) and 1 (equivalent to £2 in the experiment).

7. Conclusion

Our paper seeks to establish whether evasion is less psychologically costly than direct lies. Although suggestive evidence in support of this proposition has previously been provided (e.g., Khalmetski et al., 2017; Serra-Garcia et al., 2011; Turmunkh et al., 2019), this paper is the first to confirm it by contrasting a wide range of environments, while isolating the psychological cost of each communication. We do so in the context of novel variation of a sender-receiver game, where an informed sender can benefit from deceiving an uninformed receiver.

We find that senders do not always choose to deceive, but they are more likely to do so when they can evade rather than lie directly. We show that even after eliminating the increased plausible deniability from evasion, some types of evasion are still chosen more frequently than direct lies. This suggests that the preference for evasion is not only due to social image concerns but is also driven by intrinsic considerations. By analysing multiple types of evasion, we identify different intrinsic channels that influence the preference for evasion including an aversion to take advantage of (or act on) an opportunity to deceive, an aversion to state something that is literally untrue as well as an aversion to influence a listener's beliefs further away from the truth. Further support for the relatively higher costs of direct lies comes from the analysis of senders' beliefs. In particular, senders believe the receiver will be more likely to choose the option best for the sender under a direct lie than under an evasion, and thus direct lies have a higher *perceived* appeal in terms of persuasiveness compared to evasion. Nevertheless, senders choose direct lies less frequently suggesting that the experienced communication costs of direct lies are often greater than the perceived benefits.

We then compare how persuasive these deceptive communications *actually* are. We show that, as senders correctly anticipate, receivers are much more likely to act in the senders' favour after a direct lie than after an evasion. One of the most striking findings of our work is that receivers are also much more likely to choose the sender's preferred action following an evasion when this evasion suggests a clear recommendation. Although receivers believe that senders are equally deceptive across all communications, they are too trusting and interpret the messages at face value: the clearer the recommendation, the lower the weight they put on their belief.

Our findings have important implications both for the prevalence and the deterrence of deception. First, our work implies that deception is likely more widespread than suggested by previous estimates based on direct lying only, as in the great majority of the dishonesty literature documented in the meta-analyses of Abeler et al. (2019) and Gerlach et al. (2019). Many people might refrain from direct lies yet engage in evasions due to their lower psychological cost. Second, relying on reputation-sensitive mechanisms like increased transparency and shaming penalties that is often recommended to reduce unethical behaviour (see e.g., Abeler et al., 2019; Bø et al., 2015) might be less effective when evasion is possible, both because these the psychological cost of deception will be lower when evasion is possible, and because individuals choosing evasion are less likely to be held accountable. Thus, enforcing deterrence policies that rely on reputation, might not be helpful, and could even backfire. This

is suggested by the work of Tergiman and Villeval (2021) who find that increasing reputation costs does not make managers lie less, but rather switch from detectable to deniable lies.

We argue that communication in settings with asymmetric information and conflict of interest should be explicit, rather than free form, ensuring that any deception must take the form of direct lying rather than evasion. For instance, in job interviews, where applicants might misrepresent their skills, employers should ask direct rather than open questions. A similar suggestion emerges from research on vague disclosure showing that less flexible disclosure protocols can increase information transmission (e.g., Deversi et al., 2018) and firms will use more flexible protocols to evade or hide information at a cost to the consumer. Consider, for example, how firms who possess unfavourable information about themselves remain strategically silent because consumers do not distinguish them from firms without information (e.g., Dye, 1985; Jung and Kwon, 1988; Sah and Read, 2020) or how managers who foster a reputation for being uninformed are treated with less scepticism by consumers (Einhorn and Ziv, 2008). Our findings confirm that a demand for statements that contain instrumental information is important to reduce such deceptive communications.

Our study is a first step towards a complete understanding of the distinction between lying and evasion, and by design we excluded some key factors that may make evasion even more likely than direct lying. Two of these relate to what can be called the "menu" of deception. To cleanly measure the associated psychological cost, we restricted participants to a single type of deceptive communication that was relevant for the environment we created. Yet, outside of the lab, people can simultaneously choose between a large variety of evasive moves. Different contexts will render different evasions more or less beneficial to the deceiver, partly (but not entirely) due to their being more or less credible and detectable.

Consider, for example, the manager we introduced earlier, who must choose between different ways of avoiding telling their employee the bad news. She will want to choose the best way to slip out of her obligations, and this will depend on the circumstances. If, for instance, it is feasible that the employee will never know when the decision was made, then partial truth is a good tactic, because the manager does not have to incur the falsehood cost, and yet at the same time will appear to answer the question. If on the other hand the employee might learn that the decision had already been made at the time of the conversation, something along the lines of "I don't know what decision the board has reached" might be a better choice, because even though it incurs a falsehood cost it has a greater chance of credible deniability. Silence, or changing the subject, can work if the employee is easily side tracked. In general terms, "I don't know" can be chosen when it is credible the speaker has not learned a fact, silence when multiple questions are asked and some can be left unanswered, and partial truth when this can masquerade as the whole truth. Other evasions will similarly be more appropriate in different

contexts. For instance, the very popular "I don't remember"[20] is best used when an event has occurred long ago. The differing psychological costs associated with each item on the wide variety of deceptions people have in more naturalistic situations is likely to encourage evasion.

Another menu effect might operate through the comparison between options. For instance, we might expect someone to be more likely to deceive if their choice is between lying, evasion and truth telling than if it is between evasion and truth telling, simply because evasion may seem positively virtuous if one of its alternatives is lying directly. Because we restricted people either to truth telling or a single deceptive message we could not capture either the effect of greater flexibility in evasion, or the effect of some evasions being relatively more virtuous than others. This menu effect would be in line with the self-concept maintenance theory of Mazar et al. (2008), who suggest that people face a trade-off between gains from deception and maintaining a positive self-image, and solve this by trying to keep a balance between the two, as illustrated by die-rolling experiments where people deceive but not to the maximum extent (e.g., Fischbacher and Föllmi-Heusi, 2013). As such, a narrative for deceiving via evasion while still maintaining a self-image of honesty might be easier to generate (Bénabou et al., 2018). We leave these possibilities open for future research.

[20] See for instance https://www.politico.com/story/2017/06/25/washington-defense-trump-russia-239914 for the role of "faulty" memories in US politics.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115-1153.

American Medical Association Journal of Ethics, Opinion 8.082 – Withholding Information from Patients, *Virtual Mentor*. 2012;14(7):555-556.

Bašić, Z., & Quercia, S. (2022). The influence of self and social image concerns on lying. *Games and Economic Behavior*, *133*, 162-169.

Bénabou, R., Falk, A., & Tirole, J. (2020). Narratives, imperatives, and moral persuasion. University of Bonn, mimeo. Unpublished.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652-1678.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, *102*(5), 841-877.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, *22*(2), 191-208.

Bickart, B., Morrin, M., & Ratneshwar, S. (2015). Does it pay to beat around the bush? The case of the obfuscating salesperson. *Journal of Consumer Psychology*, *25*(4), 596-608.

Blume, A., Lai, E. K., & Lim, W. (2020). Strategic information transmission: A survey of experiments and theoretical foundations. In *Handbook of Experimental Game Theory*. Edward Elgar Publishing.

Bø, E. E., Slemrod, J., & Thoresen, T. O. (2015). Taxes on the internet: Deterrence effects of public disclosure. *American Economic Journal: Economic Policy*, *7*(1), 36-62.

Bok, S. (1978). Lying: Moral choices in public and private life. New York: Pantheon

Braghieri, L. (2023). Biased Decoding and the Foundations of Communication. CESifo Working Paper n. 10432. Unpublished.

Bucciol, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, *32*(1), 73-78.

Cai, H., & Wang, J. T. Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, *56*(1), 7-36.

Carson, T. L. (2010). Lying and deception: Theory and practice. *Oxford University Press.*

Charness, G., Samek, A. & van de Ven, J. (2020). What is Considered Deception in Experimental Economics? A Survey. Unpublished.

Cohen, K., & Kupferschmidt, K. (2020). The "very, very bad look" of remdesivir, the first FDA-approved COVID-19 drug. *Science,* Oct 28. https://www.sciencemag.org/news/2020/10/very-very-bad-look-remdesivir-first-fda-approved-covid-19-drug

Cohen, S., & Zultan, R. (2021). The Deceiving Game, *Journal of the American Philosophical Association*, 1-21.

Colzani, P., Michailidou, G., & Santos-Pinto, L. (2023). Experimental evidence on the transmission of honesty and dishonesty: A stairway to heaven and a highway to hell. *Economics Letters*, 111257.

Connelly, C. E., Zweig, D., Webster, J., & Trougakos, J. P. (2012). Knowledge hiding in organizations. *Journal of Organizational Behavior, 33(1),* 64-88.

Corran, E. (2018). Lying and Perjury in Medieval Practical Thought: A Study in the History of Casuistry. *Oxford University Press*.

Crawford, V. (1998). A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, *78*(2), 286-298.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 1431-1451.

Danziger, E. (2010). On trying and lying: Cultural configurations of Grice's Maxim of Quality. *Intercultural Pragmatics*, 7(2), 199-219.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1-17.

Deversi, M., Ispano, A., & Schwardmann, P. (2018). Spin doctors: A model and an experimental investigation of vague disclosure. *CESifo Working Paper No. 7244*, Unpublished.

Dufwenberg, M., & Dufwenberg, M. A. (2018). Lies in disguise–A theoretical analysis of cheating. *Journal of Economic Theory*, *175*, 248-264.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research*, 123-145.

Einhorn, E., & Ziv, A. (2008). Intertemporal dynamics of corporate voluntary disclosures. *Journal of Accounting Research, 46*(3), 567-589.

Egan, M., Matvos, G., & Seru, A. (2019). The market for financial adviser misconduct. *Journal of Political Economy*, *127*(1), 233-295.

Ellingsen, T., & Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, *98*(3), 990-1008.

Engelmann, D., & Strobel, M. (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, *3*, 241-260.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733.

Fallis, D. (2018). Lying and omissions. In J. Meibauer (Ed.), Oxford handbook of lying (pp. 183–192.). Oxford, UK: Oxford University Press.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525-547.

Forsythe, R., Lundholm, R., & Rietz, T. (1999). Cheap talk, fraud, and adverse selection in financial markets: Some experimental evidence. *The Review of Financial Studies*, *12*(3), 481-518.

Fries, T., Gneezy, U., Kajackaite, A., & Parra, D. (2021). Observability and lying. *Journal of Economic Behavior & Organization*, *189*, 132-149.

Gächter, S., Gerhards, L., & Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review, 97*, 72-86.

Galasinski, D. (2000). The language of deception: A discourse analytical study. Sage Publications.

Gaspar, J. P., Methasani, R., & Schweitzer, M. (2019). Fifty shades of deception: Characteristics and consequences of lying in negotiations. *Academy of Management Perspectives, 33*(1), 62-81.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin*, *145*(1), 1.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384-394.

Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, *108*(2), 419-53.

Grice, H. P. (1975). Logic and conversation. In G. Harman & D. Davidson (Eds.), The Logic of Grammar. Dickinson.

Gurun, U. G., Stoffman, N., & Yonker, S. E. (2018). Trust busting: The effect of fraud on investor behavior. *The Review of Financial Studies, 31*(4), 1341-1376.

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior, 18*(1), 59-92.

Hey, J. D. (1998). Experimental economics and deception: A comment. *Journal of Economic Psychology, 19*(3), 397-401.

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics*, *12*, 180-192.

Isler, O., & Gächter, S. (2022). Conforming with peers in honesty and cooperation. *Journal of Economic Behavior & Organization*, *195*, 75-86.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532.

Johnson, E. J., Meier, S., & Toubia, O. (2019). What's the catch? Suspicion of bank motives and sluggish refinancing. *The Review of Financial Studies*, *32*(2), 467-495.

Jung, W. O., & Kwon, Y. K. (1988). Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 146-153.

Kang, C., Packard, G., & Wooten, D. B. (2020). Beyond Truth and Lies: When and Why Consumers Evade. Unpublished.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies, 76*(4), 1359-1395.

Khalmetski, K., Rockenbach, B., & Werner, P. (2017). Evasive lying in strategic communication. *Journal of Public Economics*, *156*, 59-72.

Khalmetski, K., & Sliwka, D. (2019). Disguising lies—Image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics*, *11*(4), 79-110.

Khalmetski, K., & Tirosh, G. (2012). Two types of lies under different communication regimes. Unpublished.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608-638.

Koehler, J. D., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667-676.

Kölle, F., & Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. *Journal of Economic Behavior & Organization*, *190*, 691-703.

Kuran, T. (1997). Private truths, public lies. *Harvard University Press.*

Leibbrandt, A., Maitra, P., & Neelim, A. (2017). *Large Stakes and Little Honesty? Experimental Evidence from a Developing Country* (No. 13-17). Monash University, Department of Economics. Unpublished.

Levine, E., Hart, J., Moore, K., Rubin, E., Yadav, K., & Halpern, S. (2018). The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission. *Journal of Personality and Social Psychology*, *114*(1), 29.

Mahon, J. E. (2015). The definition of lying and deception. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Fall 2015 Edition). Available at http://plato.stanford.edu/archives/fall2015/entries/lying-definition/

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633-644.

McDaniel, T., & Starmer, C. (1998). Experimental economics and deception: A comment. *Journal of Economic Psychology*, *19*(3), 403-409.

Miller, A.C. (1998). President's memory lapses raise eyebrows in capital. *Los Angeles Times*, Sept 28, 1998. https://www.latimes.com/archives/la-xpm-1998-sep-28-mn-27284-story.html.

Morse, J. M. (2010). "Cherry picking": Writing from thin data. *Qualitative health research*, 20(1), 3-3.

New York Times (Dec 1, 2022). Transcript of Sam Bankman-Fried's interview at the DealBook Summit. https://www.nytimes.com/2022/12/01/business/dealbook/sam-bankman-fried-dealbook-interview-transcript.html.

O'Connor, K. M., & Carnevale, P. J. (1997). A nasty but effective negotiation strategy: Misrepresentation of a common-value issue. *Personality and Social Psychology Bulletin*, *23*(5), 504-515.

PBS. (2004, July 7). A conversation with former president Bill Clinton. *PBS NewsHour*. Retrieved from https://www.pbs.org/newshour/show/a-conversation-with-former-president-bill-clinton

Pittarello, A., Rubaltelli, E., & Motro, D. (2016). Legitimate lies: The relationship between omission, commission, and cheating. *European Journal of Social Psychology*, *46*(4), 481-491.

Rogers, T., & Norton, M. I. (2011). The artful dodger: Answering the wrong question the right way. *Journal of Experimental Psychology: Applied*, *17*(2), 139.

Rogers, T., Zeckhauser, R., 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*(3), 456.

Ross, Lee; Greene, David; House, Pamela (1977). "The 'false consensus effect': An egocentric bias in social perception and attribution processes". *Journal of Experimental Social Psychology*. **13** (3): 279–301.

Sah, S., & Read, D. (2020). Mind the (information) gap: Strategic nondisclosure by marketers and interventions to increase consumer deliberation. *Journal of Experimental Psychology: Applied*, *26*(3), 432.

Sánchez-Pagés, S., & Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, *61*(1), 86-112.

Sánchez-Pagés, S., & Vorsatz, M. (2009). Enjoy the silence: an experiment on truth-telling. *Experimental Economics*, *12*(2), 220-241.

Schauer, F., & Zeckhauser, R. (2007). Paltering. Harvard University, John F. Kennedy School of Government.

Schweitzer, M. E., & Croson, R. (1999). Direct Questions on Lies and Omissions. T*he International Journal of Conflict Management*, *10*(3), 225-248.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). Tübingen: J.C.B. Mohr (Paul Siebeck).

Serra-Garcia, M., Van Damme, E., & Potters, J. (2011). Hiding an inconvenient truth: Lies and vagueness. *Games and Economic Behavior*, *73*(1), 244-261.

Sheremeta, R. M., & Shields, T. W. (2013). Do liars believe? Beliefs and other-regarding preferences in sender–receiver games. *Journal of Economic Behavior & Organization*, *94*, 268-277.

Slater, T. (1910). Lying. In The Catholic Encyclopedia. New York: Robert Appleton Company. Retrieved December 23, 2020 from New Advent: http://www.newadvent.org/cathen/09469a.htm

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, *128*(3), 907-947.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76-105.

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal*, *119*(534), 47-60.

Tergiman, C., & Villeval, M. C. (2021). The Way People Lie in Markets: Detectable vs. Deniable Lies. Unpublished.

te Velde, V. L., & Louis, W. (2022). Conformity to descriptive norms. *Journal of Economic Behavior & Organization*, 200, 204-222.

Turmunkh, U., van den Assem, M. J., & Van Dolder, D. (2019). Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Science*, *65*(10), 4795-4812.

Vanberg, C. (2019). A short note on the rationality of the false consensus effect. Discussion Paper Series No. 662, University of Heidelberg.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*(1), 101-118.

White, J. J. (1980). Machiavelli and the bar: Ethical limitations on lying in negotiation. *American Bar Foundation Research Journal, 5*(4), 926-938.

For online publication

Appendix A. Detailed Theoretical Framework

In this appendix we present a more detailed theoretical model based on our deception game and provide mathematical proofs for our predictions regarding how behavioural differences can be driven by differences in psychological costs associated with direct lying and evasive communications.

2.1. The deception game (with more structure)

We consider a game with two players: a sender (S, she) and a receiver (R, he). The sender has private information about the state. She can communicate with the receiver, but she cannot take actions that have a direct impact on the two players' payoffs. The receiver does not have private information about the state, but his actions determine the payoffs of both parties.

The sender's type (θ) is represented by a three-dimensional state: $\Theta = \Theta_1 \times \Theta_2 \times \Theta_3$, where $\theta = (\theta_1, \theta_2, \theta_3)$ is an element of Θ, and $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$, $\theta_3 \in \Theta_3$. The dimensions capture elements that have both direct (Θ_1) and indirect (Θ_2) payoff consequences as well as elements that are common knowledge (Θ_3). This three-dimensional state space is necessary to implement *credible* evasions (defined later), that have an external counterpart in natural language and are not simply different labels for direct lies. It also makes for a more realistic depiction of a sender's type which is often more complex than the unidimensional depiction in standard sender-receiver games. For example, when selling a house, the quality of the house will directly affect the buyer's payoff (hence, the quality of the house is an element of Θ_1). However, the seller's expertise about the house – how informed she is about the positive and negative aspects of the house will have indirect effects as the price the buyer ends up paying will depend on what the seller can say about the house given her expertise and what the buyer ends up believing about its quality (hence, the seller's expertise is an element of Θ_2). There are also characteristics of the selling environment that are common knowledge, such as public statistics about the crime rate in the neighbourhood (which would be elements of Θ_3 in our framework). Such common knowledge and/or payoff irrelevant state characteristics can be used to implement truthful evasions. For instance, the seller can point to low general crime rates when, in fact, the next-door neighbours are notorious criminals. We now describe the specific parameters we chose for each dimension.

Θ_1 represents the primary payoff relevant characteristics of the state and consists of two elements: $\{Red, Blue\}$. $\theta_1 = Red$ is more likely than $\theta_1 = Blue$. Specifically, $\Pr(\theta_1 = Red) = \frac{11}{20}$, and $\Pr(\theta_1 = Blue) = 1 - \Pr(Red) = \frac{9}{20}$. For generalizability, we use a neutral framing for the values of θ_1 but through the later associations with the payoffs the sender and receiver can get in each case, these values can be interpreted as "Good" or "Bad." Θ_2 and Θ_3 include state characteristics that while not (directly) payoff relevant are needed to capture the differences between deceptive communications. Θ_2 represents secondary payoff relevant characteristics of the state of the world, indicating whether the sender has private information about θ_1. In particular, Θ_2 defines the sender's information type as

follows: with probability $Pr_I = \frac{7}{10}$, the sender is an *informed type* who knows the value of θ_1, the payoff relevant dimension of the state; we will denote this with $\theta_2 = I$. With probability $1 - Pr_I$, the sender is an *uninformed type* who does not know the value of θ_1; we will denote this with $\theta_2 = U$. Conditional on the sender being informed ($\theta_2 = I$), the probability that the payoff relevant dimension is Red (Pr $(\theta_1 = Red|\theta_2 = I)$ is equal to $\frac{3}{7}$, while if the sender is uninformed ($\theta_2 = U$), the respective probability (Pr $(\theta_1 = Red|\theta_2 = U)$ is equal to $\frac{5}{6}$. This means that θ_1 is more likely to be Red if the sender is uninformed, but more likely to be $Blue$ if the sender is informed.[21] Because the sender can be either informed or uninformed, evasions that claim ignorance (e.g., "I don't know the value of θ_1") are credible. That is, there are types who are genuinely ignorant and who would want the receiver to know this. Finally, we define Θ_3 to include any other common knowledge or payoff irrelevant state characteristics. In our setting, these include the probability distributions of θ_1 and θ_2 (i.e., $\{\Pr(\theta_1 = Red), \Pr(\theta_2 = I), \Pr(\theta_1 = Red|\theta_2 = I)\}$.

Timing. The timing of the game follows. First, nature determines the sender's type: $\theta = (\theta_1, \theta_2, \theta_3)$. The value of θ_3 is common knowledge. Then, if the sender is informed ($\theta_2 = I$), she observes θ_1 and chooses a message *m*, either the truth or a deception, from a set $M(\theta)$ that depends on her type. Then the payoff is realised.

In our study, each individual sender is restricted to a single deception. This is however varied across senders, so we consider deceptions covering several dimensions. Specifically, we consider a message space that includes: statements about the primary payoff relevant dimension, θ_1, statements about the sender's information type, θ_2, statements about other common knowledge state characteristics, θ_3, as well as (empty) non-statements. This entails $\cup M(\theta) = \{\Theta_1, \Theta_2, \Theta_3, \emptyset\}$. We sometimes refer to the subset including all messages that are not about the primary payoff relevant dimension, $\{\emptyset, \theta_2, \theta_3\}$, as X, where $x \in X$ is an element of this set.

The messages that can be sent depend on the sender's type. In two states she does not have a choice. If she is uninformed ($\theta = (\theta_1, U, \theta_3)$), $m \in X$ is always sent (which element of X is sent is common knowledge); if she is informed and $\theta_1$is Red ($\theta = (Red, I, \theta_3)$), the truthful message $m = Red$ is always sent.[22] Only when both the sender is informed, and θ_1 is Blue ($\theta = (Blue, I, \theta_3)$) does she have a choice. This choice is between telling the truth or sending the deceptive message, $m \in M = \{Blue, non - Blue\}$, where $non - Blue \in \{Red, x\}$. Note that the message $\{Blue\}$ is perfectly informative, since it can only be sent when the sender knows θ_1 is Blue. The receiver knows the value of the $non - Blue$ message available to the sender.

---

[21] As we show in the experimental design section, these parameters are chosen such that the expected material benefit of an evasive message is not larger than that of a direct lie. This ensures a preference for the evasive message cannot be due to higher expected material benefits.

[22] We assume the sender has no incentive to send a different message (as is clear from the payoff table).

After receiving the message, the receiver first guesses whether θ_1 is Red or Blue and then the payoffs are realised. We use a to denote the receiver's guess ($a \in \{Red, Blue\}$) and μ for the receiver's beliefs about the probability distribution over the states of the world ($\theta \in \Theta$), given the message. That is, μ assigns to each message m a probability distribution over Θ.

Payoffs. The payoff to the sender depends only on the receiver's action while the receiver's payoff depends both on his action and on θ_1. Hence, after observing his payoff, the receiver can be certain about θ_1 (the colour), but not about θ_2 (whether the sender was informed). Table 1 summarizes payoffs (the sender's payoff is listed first in each cell), where $h > l$.

Table 1. Payoff matrix $(\boldsymbol{\pi^S}, \boldsymbol{\pi^R})$

	$a = Red$	$a = Blue$
$\theta_1 = Red$	$(h,\ h)$	$(l,\ l)$
$\theta_1 = Blue$	$(h,\ l)$	$(l,\ h)$

Given the payoff structure, the sender maximizes her expected payoff if the receiver always chooses $a = Red$ (since $h > l$), while the receiver when his action matches the realisation of the primary payoff relevant state dimension (i.e., if $a = \theta_1$).

Definitions.

Here, we re-iterate the definitions provided in the main text, with their mathematical counterparts given the additional structure we impose in this appendix.

Definition 1 (Literal meaning). The literal meaning of m is the a priori, common understanding that $m = m_{\theta_{i\in\{1,2,3\},j}}$ implies that $\theta_{i\in\{1,2,3\},j} = \theta_i$, where θ_i is the value of the dimension of $\theta \in \Theta$ the message refers to; $\theta_{i\in\{1,2,3\},j}$ implies that θ_i takes the value θ_j.

Definition 2 (Direct message). A message $m = m_{\theta_{i,j}}$ is direct if $i = 1$.

Definition 3 (Evasive message). A message $m = m_{\theta_{i,j}}$ is evasive if $i \neq 1$ and $\theta_2 = I$ and $M(\theta) = \{\theta_1, x\}_1$, where $\theta_1 \in \Theta_1, x \in X$.

Definition 4 (Truth). A message $m = m_{\theta_{i,j}}$ is true if $\theta_{i,j} = \theta_i$, $\forall i \in \{1,2,3\}$.

Definition 5.0 (Lie). A message $m = m_{\theta_{i,j}}$ is a lie if $\theta_{i,j} \neq \theta_i$, $\forall i \in \{1,2,3\}$.

Definition 5.1 (Direct Lie). Formally, a message $m = m_{\theta_{1,j}}$ is a direct lie if $\theta_{1,j} \neq \theta_1$.

Definition 5.2 (Evasive Lie). A message $m = m_{\theta_{i,j}}$ is an evasive lie if $\theta_{i,j} \neq \theta_i$, $\forall i \neq 1$.

Definition 6 (Deception). A message $m = m_{\theta_{i,j}}$ is deceptive if $\mu\left(\theta_i | m_{\theta_{i,j}}\right) - Pr(\theta_i) > 0$, $\forall i \in \{1,2,3\}$ and S has the option to send $m' = m_{\theta_{i,j'}}$ for which $\mu\left(\theta_i | m_{\theta_{i,j}}\right) - Pr(\theta_i) > \mu\left(\theta_i | m_{\theta_{i,j'}}\right) - Pr(\theta_i)$.

Preferences. We assume senders may incur psychological costs from the message they choose and its potential implications. We also assume that receivers are one of two types: sophisticated (R^S) or

naive (R^N) (similar to e.g., Kartik, 2009).[23] A sophisticated receiver chooses the action that maximizes his expected payoff given his beliefs about the state distribution which are updated following Bayes' rule upon observing the sender's message.

In contrast, a naive receiver does not use Bayes' rule to update his beliefs about the state distribution, but rather interprets the message literally. Specifically, if a message makes no statement about the payoff relevant state dimension, the naive receiver's posterior belief about the distribution of the payoff relevant dimension θ_1 remains equal to his prior (i.e., $\mu_{R^N}(\theta_1 = Red) = \Pr(Red) = \frac{11}{20}$). If the message makes a statement about the payoff relevant state dimension, the naive receiver's posterior belief moves away from the prior in the direction suggested by the message, more so depending on the precision of the message. That is, if $m = Red$, $\mu_{R^N}(\theta_1 = Red|m) = 1$; if $m = Blue$, $\mu_{R^N}(\theta_1 = Red|m) = 0$; if $m = x$ and the message implies a higher probability for one of the two possible values for θ_1, then $\mu_{R^N}(\theta_1 = Red|m = x) \neq \frac{11}{20}$. Note that $\mu_{R^N}(\theta_1 = Red|m \neq Red)$ will always be strictly lower than when $m = Red$. The naive receiver then chooses $a = Red$ if their posterior belief suggests that $\theta_1 = Red$ is at least equally likely to $\theta_1 = Blue$, i.e., $\mu_{R^N}(\theta_1 = Red|m) \geq \frac{1}{2}$.

Furthermore, naive receivers do not draw inferences about the sender's message (i.e., whether it is deceptive or truthful) from the payoff realization. That is, if the sender sent $m = Red$ when they knew the colour of the state (θ_1) is $Blue$, and the receiver chooses $a = Red$ (or $a = Blue$) and therefore gets a payoff of l (or h), the naive receiver does not go through the inference process of comparing the payoff they should have gotten if the message they received was truthful with what they actually got to conclude that the deceptive message must have been chosen by the sender. The sophisticated receiver, however, does go through this inference process. Therefore, the likelihood that a deceptive message (in particular, a direct lie) will be interpreted as such depends on the proportion of sophisticated receivers in the population. This proportion will influence the magnitude of the social image cost described below. Let η be the proportion of naive receivers in the population (and $1 - \eta$ that of sophisticated receivers).

The utility of the sender (U^S) and the receiver (U^R) is given by the following functions:

$$U^S(\theta, m, a) = \pi^S(a) - c_d(\theta, m) - c_l(\theta, m) - c_i(\theta, m, \mu) - c_s(\theta, m, p_{vf}) \quad (1)$$

$$U^R(\theta, a) = \pi^R(\theta, a) \quad (2)$$

where:

$c_d(\theta, m)$ is the *deception cost* from sending a deceptive message. This is incurred whenever the sender chooses the non-truthful message (i.e., when $m \neq Blue$).

$c_f(\theta, m)$ is the *falsehood cost* incurred when the message is false (i.e., a lie). We will say that given θ, $c_f(\theta, m) > 0$ if $m = m_{\theta_{i,j}}$ and $\theta_j \notin \Theta$; $c_f(\theta, m) = 0$ otherwise.

[23] Kartik (2009) introduces naïve receivers in an alternative but equivalent way by assuming that receivers are likely to take a naïve action with a certain probability, e.g., η.

$c_i(\theta, m, \mu)$ is the *influence cost*, which increases with the difference between the sender's belief about the receiver's belief about θ_1 and its realized probability (i.e. given θ, m and m', $c_i(\theta, m, \mu) > c_i(\theta, m', \mu)$ if $\mu(\theta_1 = j|m, \theta_1 = i) > \mu(\theta_1 = j|m', \theta_1 = i), \forall i \neq j$).

$c_s(\theta, m, p_{vf})$ is the *social image cost* incurred when the sender's message is not the truth and increases with the probability the receiver can infer the sender was deceptive (p_{vf}).

We refer to the sum of all communication costs as C. Moreover, when the message is perfectly informative about the sender's type (i.e., the receiver can infer it from the message with certainty) or the sender does not have a choice regarding which message to send, we assume $C = 0$. This happens when $m = Blue$ (a perfectly informative message that is only available to the informed sender when $\theta_1 = Blue$) or when a message is sent automatically (i.e., either when $\theta_1 = Red$ or $\theta_2 = U$). Let λ be the probability that C is sufficiently low that the sender will behave as a standard material payoff maximizer and will therefore deceive if it is beneficial to do so. Consequently, $1 - \lambda$ is the probability that the sender's message is perfectly informative about θ.

2.2. Analysis

Our equilibrium solution concept is Perfect Bayesian Equilibrium (PBE). A PBE consists of a set of strategies for the sender and the receiver, and a set of beliefs for the receiver. The strategies are (m^*, a^*), where m^* is the sender's (pure) message strategy and a^* is the receiver's (pure) action strategy. The receiver's beliefs are given by μ^*, which assigns to each m a probability distribution over Θ such that the equilibrium strategies and beliefs satisfy sequential rationality and consistency of beliefs. Sequential rationality is that at any information set, a player uses a best response strategy given their beliefs and holding the other player's strategy constant; consistency of beliefs is that each player's beliefs follow Bayes' rule (wherever appropriate) and is consistent with the strategy profile. Unless μ_{R^S} differs from μ_{R^N}, we will omit the subscript to refer to the receiver's beliefs.

Note that since the (Red, I, θ_3) and the (θ_1, U, θ_3) sender types are not active players and therefore their behaviour is constant, in describing the equilibria we can omit reiterating their strategies and refer to the $(Blue, I, \theta_3)$ sender type simply as the sender.

First, we describe the equilibrium actions when senders do not incur any communication costs and receivers are sophisticated.

Lemma 1. In any PBE of the game where players are sophisticated and only care about material payoffs, S will choose either $m = Red$ or $m = x$ (depending on which one is available as the $non - Blue$ message). R will best reply with $a^*(m = Blue) = Blue$, $a^*(m = Red) = Blue$, $a^*(m = x) = Red$, when $non - Blue = Red$, and $a^*(m = Blue) = Blue$, $a^*(m = Red) = Red$, $a^*(m = x) = Blue$, when $non - Blue = x$.

Proof: Suppose $m^* = Blue$. When $non - Blue = Red$, $\mu(\theta_1 = Red|m = Red) = 1$; $\mu(\theta_1 = Blue|m = Blue) = 1$; $\mu(\theta_1 = Red|m = x) = \frac{5}{6}$. Consequently, the receiver best replies by choosing $a^*(m = Red) = Red$, $a^*(m = Blue) = Blue$, $a^*(m = x) = Red$. Therefore, the sender's

expected payoff is equal to l. The sender in this case has a profitable deviation to $m = Red$, where her payoff would be equal to h. Similarly, when $non - Blue = x$. Does the receiver have a profitable deviation from the above strategy when $m^* = Red$? If this is the sender's strategy when $non - Blue = Red$, receiver's beliefs are: $\mu(\theta_1 = Red|m = Red) = \frac{6}{14}$; $\mu(\theta_1 = Blue|m = Blue) = 1$; $\mu(\theta_1 = Red|m = x) = \frac{5}{6}$ and he best replies with $a^*(m = Blue) = Blue$, $a^*(m = Red) = Blue$, $a^*(m = x) = Red$. The sender has no profitable deviation in this case since she is indifferent between $m = Blue$ and $m = Red$. The argument for when $non - Blue = x$ follows similarly.

Next, we identify the critical value for the communication cost that determines whether a sender is a truth-teller. In doing so, we also describe the equilibrium strategies when senders have no communication costs and receivers are sophisticated.

Lemma 2. If and only if $C > h - l$, S will choose $m^* = Blue$, i.e., tell the truth. If C = 0, S will choose the potentially deceiving message.

Proof: Suppose $m^* = Blue$. Then, $\mu(\theta_1 = Red|m = Red) = 1$; $\mu(\theta_1 = Blue|m = Blue) = 1$; $\mu(\theta_1 = Red|m = x) = \frac{5}{6}$. Consequently, the receiver best replies by choosing $a^*(m = Red) = Red$, $a^*(m = Blue) = Blue$, $a^*(m = x) = Red$. Therefore, the sender's expected payoff is equal to l. By deviating to $m = Red$, her payoff would be equal to $h - C$. This is not a profitable deviation when $C > h - l$. Hence, only if this condition is met, that $C > h - l$, it is optimal for the sender to truthfully reveal the state (i.e., to use $m^* = Blue$ as their equilibrium strategy). Since $h > l$, this condition cannot be met when senders do not incur communication costs (C=0).

The difference between h and l (i.e., $h - l$) is the difference between the high and low payoffs in the game (see Table 1). Lemma 1 establishes the threshold above which a sender with the opportunity to deceive would find it optimal to tell the truth. Given this, we can now redefine λ as the probability that U^S is such that $C < h - l$, i.e., that the sender's communication costs are low enough to behave as an expected payoff maximizer. This property helps us differentiate between two psychological types of senders: truth-telling senders – S^T- whose communication costs are high enough such that they always tell the truth, and dishonest senders, who will lie when it is profitable – S^L. A corollary of Lemma 1 is that S^L, the dishonest sender, would never send the truthful message $Blue$ in equilibrium.

Corollary 1. The message strategy $m_{S^L} = Blue$ cannot be part of a PBE of the game.

Proof: Suppose that $m_{S^L}{}^* = Blue$ is S^L's equilibrium strategy. Then, the receiver's beliefs about the conditional distribution of the payoff relevant state dimension are:

$$\begin{cases} \mu_{R^S}(\theta_1 = Red \,|m = Red) = 1; \\ \mu_{R^S}(\theta_1 = Red \,|m = Blue) = 0; \\ \mu_{R^S}(\theta_1 = Red \,|m = x) = \frac{5}{6}; \end{cases} \qquad \begin{cases} \mu_{R^N}(\theta_1 = Red \,|m = Red) = 1; \\ \mu_{R^N}(\theta_1 = Red \,|m = Blue) = 0; \\ \mu_{R^N}(\theta_1 = Red \,|m = x) = p; \end{cases}$$

Given these beliefs, the sophisticated receiver's best reply is: $a_{R^S}(m = Red) = Red$; $a_{R^S}(m = Blue) = Blue$; $a_{R^S}(m = x) = Red$. The naive receiver's best reply is: $a_{R^N}(m = Red) = Red$; $a_{R^N}(m = Blue) = Blue$; $a_{R^N}(m = x) = Red$, if $p \geq \frac{1}{2}$; $a_{R^N}(m = x) = Blue$, if $p < \frac{1}{2}$. Then, the deceptive sender's utility from each message is:

$$U^{S^L}(m = Blue) = l;\ U^{S^L}(m = Red) = h;\ U^{S^L}(m = x) = (1 - \eta)h + p\eta h + (1 - p)\eta l$$

Since $h > l$, the deceptive sender has a profitable deviation to $m_{S^L} = Red$, showing that $m^*_{S^L} = Blue$ cannot be part of a PBE of the game.

It follows that in equilibrium, the message $Blue$ is perfectly informative about the state as it will only be sent by a truth-telling sender. When will the dishonest sender choose the direct lie over the evasive message in equilibrium? Proposition 1 states that if at least one quarter of the senders are truth-tellers, then the receiver's optimal action is to choose the sender's preferred action (i.e., Red) after either the Red message or the evasive one (x). This makes it optimal for the dishonest sender to choose the direct lie (i.e., $m_{S^L} = Red$) in equilibrium.

Proposition 1 (Direct lying equilibrium). If $\lambda \leq \frac{3}{4}$, the strategy set $m^*_{S^L} = Red, m^*_{S^T} = Blue, a^*(m) = Red, \forall m \in \{Red, x\}\ and\ a^*(m = Blue) = Blue$, constitutes a PBE of the game.

Proof: The receiver's beliefs about the state given this message strategy of the deceptive sender are equal to:

$$\begin{cases} \mu_{R^S}(\theta_1 = Red \,|m = Red) = \frac{6}{6+8\lambda}; \\ \mu_{R^S}(\theta_1 = Red \,|m = Blue) = 0; \\ \mu_{R^S}(\theta_1 = Red \,|m = x) = \frac{5}{6}; \end{cases} \qquad \begin{cases} \mu_{R^N}(\theta_1 = Red \,|m = Red) = 1; \\ \mu_{R^N}(\theta_1 = Red \,|m = Blue) = 0; \\ \mu_{R^N}(\theta_1 = Red \,|m = x) = p; \end{cases}$$

Following these beliefs, the naive receiver's optimal action when $m = Blue$ is $a_{R^N} = Blue$, while when $m = Red$, the naive receiver's optimal action is $a = Red$; but, when $m = x$, the naive receiver chooses $a_{R^N} = Red$ if $p \geq \frac{1}{2}$ and $a = Blue\ otherwise$.

The sophisticated receiver best replies by choosing $a_{R^S}(m = x) = Red$ and $a_{R^S}(m = Blue) = Blue$. When $m = Red$, the sophisticated receiver would optimally choose $a_{R^S} = Red$ as long as $\frac{6}{6+8\lambda} \geq \frac{1}{2}$. This condition is equivalent to $\lambda \leq \frac{3}{4}$. The deceptive sender does not have a profitable deviation since the payoff they obtain by $m_{S^L} = Red$ is at least equal to what they would get by sending $m_{S^L} = x$ and greater than what they would get if they sent $m_{S^L} = Blue$.

Hence, as long as $\lambda \leq \frac{3}{4}$, $m^*_{S^L} = Red$ is an equilibrium strategy for which the deceptive sender's expected material payoff is equal to g.

Proposition 2 states that for the receiver to optimally choose Red after the evasive message (x), at least half of the senders need to be truth-tellers and the evasive message is such that the naive receivers believe that $\theta_1 = Red$ is at least as likely as $\theta_1 = Blue$. Given this, it is optimal for the dishonest sender to choose the evasive message (i.e., $m_{S^L} = x$) in equilibrium.

Proposition 2 (Evasive equilibrium). If $\lambda \leq \frac{1}{2}$ and $\mu_{R^N}(\theta_1 = Red | m = x) \geq \frac{1}{2}$, the strategy set $m^*_{S^L} = x, m^*_{S^T} = Blue,\ a^*(m) = Red, \forall m \in \{Red, x\}$ and $a^*(m = Blue) = Blue$, constitutes a PBE of the game.

Proof: R's beliefs about the state given this S^L message strategy are equal to:

$$\begin{cases} \mu_{R^S}(\theta_1 = Red\ | m = Red) = 1; \\ \mu_{R^S}(\theta_1 = Red\ | m = Blue) = 0; \\ \mu_{R^S}(\theta_1 = Red\ | m = x) = \frac{5}{6+8\lambda}; \end{cases} \qquad \begin{cases} \mu_{R^N}(\theta_1 = Red\ | m = Red) = 1; \\ \mu_{R^N}(\theta_1 = Red\ | m = Blue) = 0; \\ \mu_{R^N}(\theta_1 = Red\ | m = x) = p; \end{cases}$$

Following these beliefs, the naive receiver's optimal action when $m = Blue$ is $a_{R^N} = Blue$, while when $m = Red$, the naive receiver's optimal action is $a_{R^N} = Red$; when $m = x$, the naive receiver chooses $a_{R^N} = Red$ if $p \geq \frac{1}{2}$ and $a_{R^N} = Blue$ otherwise.

The sophisticated receiver best replies by choosing $a_{R^S}(m = Red) = Red$ and $a_{R^S}(m = Blue) = Blue$. When $m = x$, the sophisticated receiver would optimally choose $a_{R^S} = Red$ as long as $\frac{5}{6+8\lambda} \geq \frac{1}{2}$, i.e., $\lambda \leq \frac{1}{2}$, otherwise they would optimally choose $a_{R^S} = Blue$.

Suppose $\lambda \leq \frac{1}{2}$ and the sophisticated receiver optimally chooses $a_{R^S}(m = x) = Red$ and that $p \geq \frac{1}{2}$ and the naive receiver optimally chooses $a_{R^N}(m = x) = Red$ also. The deceptive sender does not have a profitable deviation since the payoff they obtain by $m_{S^L} = x$ is equal to what they would get by sending $m_{S^L} = Red$ and greater than what they would get if they sent $m_{S^L} = Blue$.

Hence, as long as $\lambda \leq \frac{1}{2}$ and $p \geq \frac{1}{2}$, $m^*_{S^L} = x$ is an equilibrium strategy for which the deceptive sender's expected material payoff is equal to g.

Note that if $m^*_{S^L} = x$ is an equilibrium strategy, $m^*_{S^L} = Red$ is also an equilibrium strategy since the constraint for the latter is stricter than for the former. Importantly, the expected payoff to the dishonest sender from both strategies is the same (and equal to h). This ensures that the expected material benefit of evasive deception is not larger than that of direct lying. This is summarized in the following remark.

Remark 1. If $\lambda \leq \frac{1}{2}$ and $\mu_{R^N}(\theta_1 = Red|m = x) \geq \frac{1}{2}$, S^L is equally well off by choosing $m = Red$ or $m = x$.

2.2.1. Introducing specific evasive messages

We now restrict the game to the specific messages used in our experiment about the state dimensions that are not primary payoff relevant. This is the following set *X*:

x_1(IGNORANCE) = "I don't know the colour of the state"

x_2 (PARTIAL) = "The state was more likely to be Red than Blue"

x_3 (SILENCE) = $\emptyset$

Note that the literal meaning of x_1 is that the sender is uninformed ($\theta_2 = U$), that of x_2 is that the primary payoff relevant dimension had a higher chance of being Red, rather than $Blue$ ($\Pr(\theta_1 = Red) > \Pr(\theta_1 = Blue)$), while x_3 represents silence or making no statement about any state dimension. These messages can only influence the naive receiver's beliefs about the payoff relevant state dimension, and only x_2 (PARTIAL) changes the naive receiver's beliefs away from their prior and toward the belief the state is Red (as suggested by the message).[24] Consequently, the naive receiver's beliefs following each message are:

$$\begin{cases} \mu_{R^N}(\theta_1 = Red|m \in \{x_1, x_3\}) = \dfrac{11}{20}; \\ \mu_{R^N}(\theta_1 = Red|m = x_2) > \dfrac{11}{20}. \end{cases}$$

Given the definition of the influence cost (c_i), x_2 has a higher influence cost than x_1 and x_3 since it leads to more inaccurate beliefs in the naive receiver when $\theta_1 = Blue$ and the sender could reveal this truthfully. The messages also differ in terms of the falsehood cost (c_f) incurred by the sender when the sender has a choice (i.e., when $\theta = (Blue, I, \theta_3)$). Specifically, x_2 and x_3 are both truthful, regardless of the sender's type, while x_1 is true only when $\theta_2 = U$, according to definition 2. Therefore, x_1 has the highest falsehood cost. Direct lies incur a greater social image cost than evasions. When the sender lies directly ($m = Red$), the sophisticated receiver will correctly infer the message was deceptive. When the sender evades ($m \in \{x_1, x_2, x_3\}$), neither the sophisticated nor the naive receiver can learn whether the message was truthful even after observing the payoff realization. Hence, $p_{vf}(m = Red) > p_{vf}(m \in \{x_1, x_2, x_3\})$, which means all the evasive messages have a lower social image cost (c_s) than the direct lie. Moreover, all evasive messages as well as the direct lie are equally deceptive when the sender knows that $\theta_1 = Blue$, as the sender could have truthfully revealed this. Lastly, the truthful message ($m = Blue$) has a communication cost equal to 0, the lowest of all messages.

[24] For our comparative analysis, it does not matter by how much the naïve receiver's belief is strengthened, or how many such receivers will be influenced in that manner. What matters is that there is a positive probability of that case happening.

Next, we combine this analysis and rank all possible messages available to the sender when $\theta = (Blue, I, \theta_3)$ based on their communication costs.

Deception cost:

$c_d(\theta, m = Red) = c_d(\theta, m = x_1) = c_d(\theta, m = x_2) = c_d(\theta, m = x_3) > c_d(\theta, m = Blue)$

Falsehood cost:

$c_f(\theta, m = Red) = c_f(\theta, m = x_1) > c_f(\theta, m = x_2) = c_f(\theta, m = x_3) = c_f(\theta, m = Blue)$

Influence cost:

$c_i(\theta, m = Red, a) > c_i(\theta, m = x_2, a) > c_i(\theta, m = x_1, a) = c_i(\theta, m = x_3, a) > c_i(\theta, m = Blue, a)$

Social image cost:

$$c_s(\theta, m = Red, p_{vf}) > c_s(\theta, m = x_1, p_{vf}) = c_s(\theta, m = x_2, p_{vf}) = c_s(\theta, m = x_3, p_{vf}) > c_s(\theta, m = Blue, p_{vf})$$

Summing across these inequalities we find that:

$$C(m = Red) > C(m = x_1) \geq C(m = x_2) \geq C(m = x_3) > C(m = Blue) \quad (3)$$

Recall that as long as $\lambda \leq \frac{1}{2}$, material payoff for the dishonest sender is the same for either message $m = Red$ or $m = x$ (Remark 1). Furthermore, equation (3) states that the communication costs associated with $m = Red$ are strictly higher than those associated with $m = x$, where $x \in \{x_1, x_2, x_3\}$. Therefore, since the material benefits from the four messages are equal in equilibrium, the likelihood that $m = x$ or $m = Red$ will be chosen in equilibrium instead of the truthful $m = Blue$, depends on the probability that the expected benefit of sending a deceptive message ($h - C(m \in \{Red, x\})$) is greater than the expected benefit of sending the truthful message ($l - C(m = Blue)$). That is, it depends on the probability that $h - C(m = Red) > l - C(m = Blue)$ and that $h - C(m = x) > l - C(m = Blue)$. Since the communication cost of being truthful is equal to 0 ($C(m = Blue) = 0$), these inequations can be rewritten as $C(m = Red) < \mathrm{h} - l$ and $C(m = x) < \mathrm{g} - l$. We assume that $C \sim U(0, n)$ and $0 \leq h - l \leq n$. Since $C(m = Red) > C(m = x)$, it follows that $\Pr(C(m = Red) < h - l) < \Pr(C(m = x) < h - l)$. A similar argument can be applied to comparing the likelihood that each evasive message will be chosen in equilibrium.

Prediction 1. If $x \in \{x_1, x_2, x_3\}$, senders are more likely to choose $m = x$ than $m = Red$ on the equilibrium path. Moreover, the lower is $C(m = x_i), i \in \{1,2,3\}$, the higher the likelihood that senders will choose $m = x_i$.

Prediction 1 essentially states that the lower the communication cost of a message, the more likely a sender is to choose it. Therefore, the direct lying message is the least likely to occur in equilibrium.

Next, we consider the case where, after the payoff realization, the receiver learns whether the sender is informed (i.e., learns the value of θ_2) and whether the sender had a non-singleton message choice (i.e., whether the value of θ_1 is $Blue$), and the sender knows this. In this case it is certain that choosing the direct lie will be interpreted as deceptive, since the inference process from the own payoff realization has been eliminated and even the naive receivers will understand this is the case. This holds also for

evasions where it is now clear for both the sophisticated and the naive receivers the sender made a deceptive choice. Hence, the social image cost the sender incurs when sending an evasive message is equal to that of a direct lie. Formally: $c_s(\theta, m = Red, p_{vf} = 1) = c_s(\theta, m = x_1, p_{vf} = 1) = c_s(\theta, m = x_2, p_{vf} = 1) = c_s(\theta, m = x_3, p_{vf} = 1) > c_s(\theta, m = Blue, p_{vf} = 1)$. Based on a similar argument as for Prediction 1, we formulate the following:

Prediction 2. The likelihood $m = x$ or $m = Red$ when $p_{vf} = 1$ is lower than when $p_{vf} < 1$.

Prediction 2 states that whenever the probability the receiver will find out whether the sender sent a deceptive message increases, the rate of deception will decrease.

Appendix B. Additional Analyses

Do average beliefs differ across treatments?

In the following tables we present the results of multiple comparison tests (Tukey HSD) for differences in mean beliefs. For each pairwise comparison, the tables include the size of the difference in average beliefs, the upper and lower bounds of the 95% confidence interval for this difference, and the corresponding p-value from the Tukey HSD test (which adjusts for multiple comparisons).

Sender-Hidden experiment

Table B1. Comparison of senders' beliefs about the likelihood of receivers guessing Red after the truthful message (Blue) in Sender-Hidden

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-13.5	-18.9	-8.0	0.00
PARTIAL – DIRECT	-11.8	-17.2	-6.3	0.00
SILENCE – DIRECT	-9.6	-15.1	-4.2	0.00
PARTIAL – IGNORANCE	1.7	-3.8	7.2	0.85
SILENCE – IGNORANCE	3.8	-1.6	9.3	0.27
SILENCE – PARTIAL	2.1	-3.4	7.6	0.75

Table B2. Comparison of senders' beliefs about the likelihood of receivers guessing Red after the deceptive message (non-Blue) in Sender-Hidden

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-12.7	-17.6	-7.8	0.00
PARTIAL – DIRECT	-2.2	-7.1	2.8	0.67
SILENCE – DIRECT	-15.0	-20.0	-10.1	0.00
PARTIAL – IGNORANCE	10.5	5.6	15.4	0.00
SILENCE – IGNORANCE	-2.4	-7.3	2.6	0.60
SILENCE – PARTIAL	-12.9	-17.8	-7.9	0.00

Table B3. Comparison of senders' beliefs about the likelihood of other senders choosing the deceptive message (non-Blue) in Sender-Hidden

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	2.1	-3.5	7.7	0.77

PARTIAL – DIRECT	-2.3	-8.0	3.3	0.71
SILENCE – DIRECT	-2.4	-8.0	3.3	0.70
PARTIAL – IGNORANCE	-4.4	-10.1	1.2	0.18
SILENCE – IGNORANCE	-4.5	-10.1	1.2	0.17
SILENCE – PARTIAL	-0.0	-5.7	5.6	1.00

Sender-Open experiment

Table B4. Comparison of senders' beliefs about the likelihood of receivers guessing Red after the truthful message (Blue) in Sender-Open

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-10.0	-15.4	-4.7	0.00
PARTIAL – DIRECT	-10.1	-15.5	-4.7	0.00
SILENCE – DIRECT	-11.0	-16.3	-5.6	0.00
PARTIAL – IGNORANCE	-0.0	-5.4	5.3	1.00
SILENCE – IGNORANCE	-0.9	-6.3	4.4	0.97
SILENCE – PARTIAL	-0.9	-6.3	4.5	0.97

Table B5. Comparison of senders' beliefs about the likelihood of receivers guessing Red after the deceptive message (non-Blue) in Sender-Open

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-13.0	-18.0	-8.0	0.00
PARTIAL – DIRECT	-5.8	-10.8	-0.8	0.02
SILENCE – DIRECT	-13.0	-18.0	-8.0	0.00
PARTIAL – IGNORANCE	7.2	2.2	12.2	0.00
SILENCE – IGNORANCE	0.0	-5.0	5.0	1.00
SILENCE – PARTIAL	-7.2	-12.3	-2.2	0.00

Table B6. Comparison of senders' beliefs about the likelihood of other senders choosing the deceptive message (non-Blue) in Sender-Open

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	4.2	-1.4	9.8	0.22
PARTIAL – DIRECT	-0.2	-5.8	5.5	1.00
SILENCE – DIRECT	1.0	-4.7	6.6	0.97

PARTIAL – IGNORANCE	-4.4	-10.0	1.3	0.20
SILENCE – IGNORANCE	-3.2	-8.9	2.4	0.46
SILENCE – PARTIAL	1.1	-4.6	6.8	0.96

Receiver-Hidden experiment

Table B7. Comparison of receivers' beliefs about the likelihood of senders choosing the deceptive message (non-Blue) in Receiver-Hidden

Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-3.8	-9.7	2.2	0.36
PARTIAL – DIRECT	-0.4	-6.3	5.6	1.00
SILENCE – DIRECT	-0.9	-6.8	5.0	0.98
PARTIAL – IGNORANCE	3.4	-2.6	9.3	0.46
SILENCE – IGNORANCE	2.9	-3.1	8.8	0.60
SILENCE – PARTIAL	-0.5	-6.5	5.4	1.00

Table B8. Comparison of receivers' beliefs about the likelihood of other receivers guessing Red after the deceptive message (non-Blue) in Receiver-Hidden

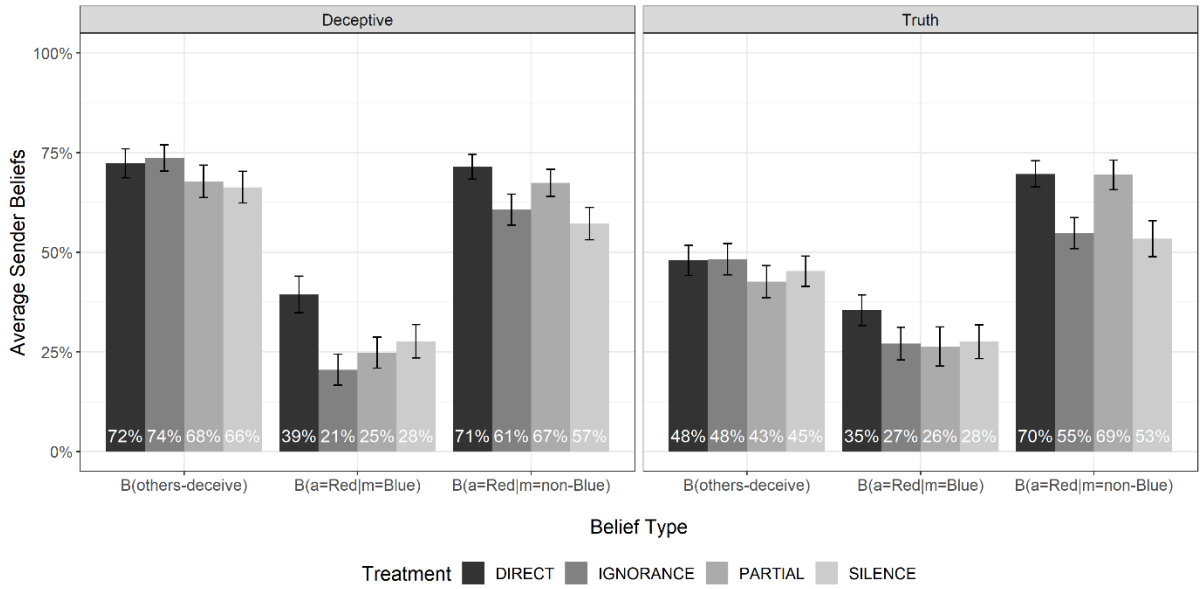
Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	-23.7	-29.2	-18.1	0.00
PARTIAL – DIRECT	-11.4	-16.9	-5.8	0.00
SILENCE – DIRECT	-21.3	-26.8	-15.8	0.00
PARTIAL – IGNORANCE	12.3	6.8	17.8	0.00
SILENCE – IGNORANCE	2.4	-3.2	7.9	0.69
SILENCE – PARTIAL	-9.9	-15.5	-4.4	0.00

Is the distribution of beliefs across treatments affected by the decision?

Sender-Hidden experiment

The following figure presents the distribution of sender beliefs across treatments and choice of message.

Figure B1. Average sender beliefs across treatments and message in Sender-Hidden



*Note*s. The figure depicts the mean reported sender belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar).

The following tables present the results of ANOVA tests for differences in mean senders' beliefs across treatments and choice of message (deceptive vs. truth).

Table B9. ANOVA results of senders' beliefs about the likelihood of receivers guessing Red after the truthful message (Blue) across treatments and decision in Sender-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	33150	11050	16.21	0.00
Decision	1	310	310	0.45	0.50
Treatment x Decision	3	4290	1430	2.10	0.10
Residuals	1202	819593	682	NA	NA

Table B10. ANOVA results of senders' beliefs about the likelihood of receivers guessing Red after the deceptive message (non-Blue) across treatments and decision in Sender-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	50897	16966	30.70	0.00
Decision	1	1714	1714	3.10	0.08
Treatment x Decision	3	2517	839	1.52	0.21
Residuals	1202	664350	553	NA	NA

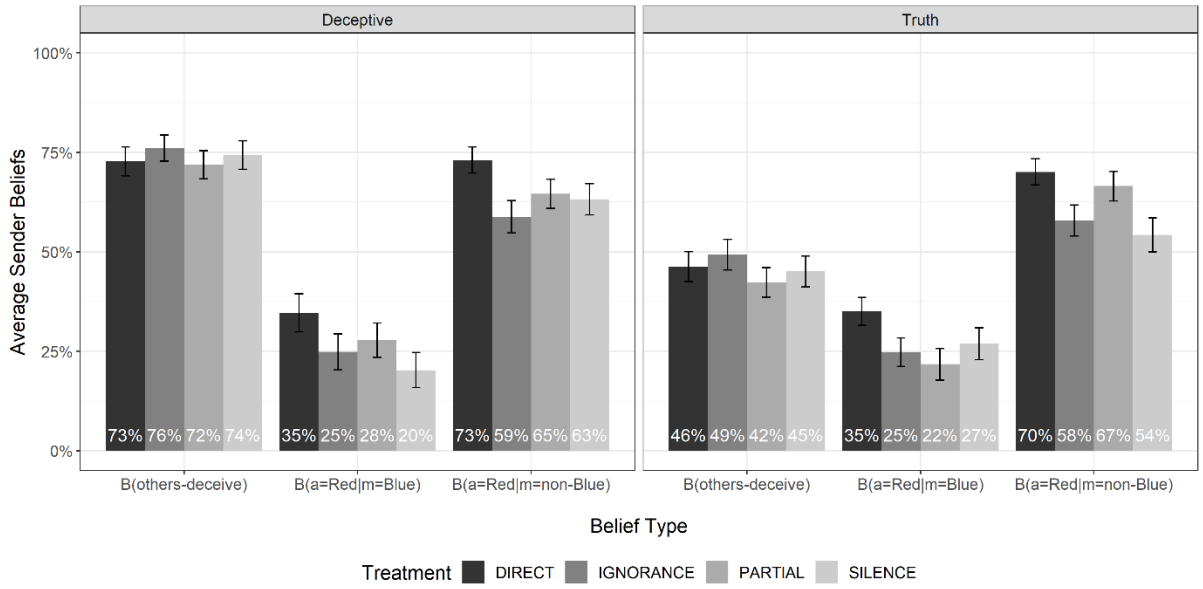
Table B11. ANOVA results of senders' beliefs about the likelihood of other senders choosing the deceptive message (non-Blue) across treatments and decision in Sender-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	4173	1391	2.40	0.07
Decision	1	172769	172769	297.72	0.00
Treatment x Decision	3	902	301	0.52	0.67
Residuals	1202	697530	580	NA	NA

Sender-Open experiment

The following figure presents the distribution of sender beliefs across treatments and choice of message.

Figure B2. Average sender beliefs across treatments and message in Sender-Open



*Note*s. The figure depicts the mean reported sender belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar).

Table B12. ANOVA results of senders' beliefs about the likelihood of receivers guessing Red after the truthful message (Blue) across treatments and decision in Sender-Open

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	24524	8175	12.56	0.00
Decision	1	12	12	0.02	0.89
Treatment x Decision	3	5987	1996	3.06	0.03
Residuals	1196	778740	651	NA	NA

Table B13. ANOVA results of senders' beliefs about the likelihood of receivers guessing Red after the deceptive message (non-Blue) across treatments and decision in Sender-Open

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	35993	11998	21.35	0.00
Decision	1	2207	2207	3.93	0.05
Treatment x Decision	3	4724	1575	2.80	0.04
Residuals	1196	672080	562	NA	NA

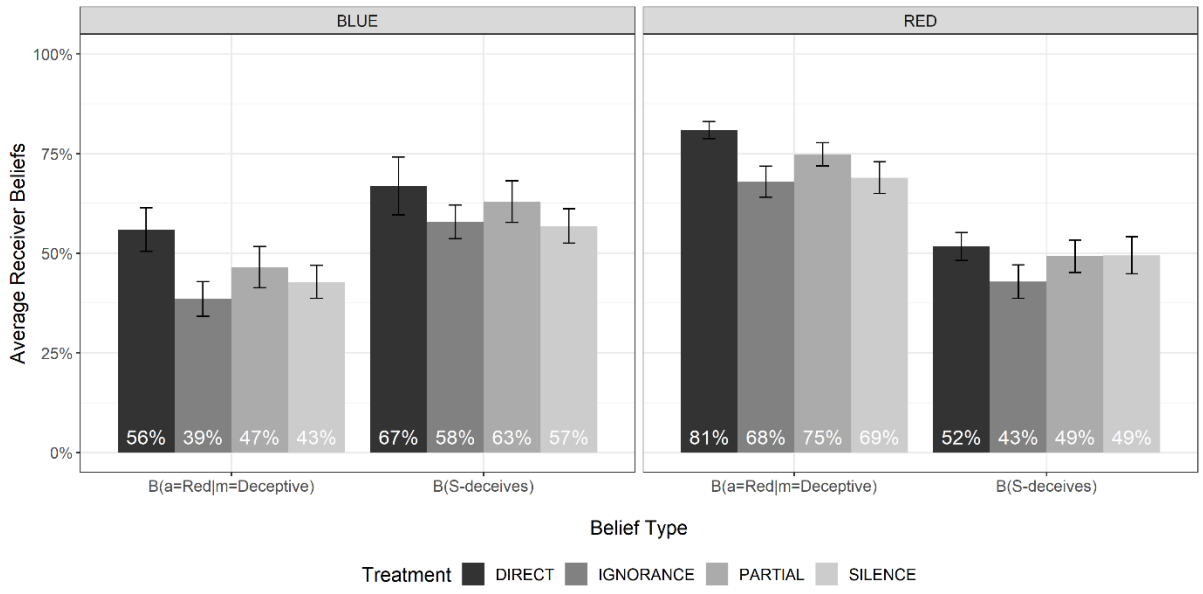
Table B14. ANOVA results of senders' beliefs about the likelihood of other senders choosing the deceptive message (non-Blue) across treatments and decision in Sender-Open

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	3735	1245	2.32	0.07
Decision	1	233493	233494	435.16	0.00
Treatment x Decision	3	587	196	0.36	0.78
Residuals	1196	641734	537	NA	NA

Receiver-Hidden experiment

The following figure presents the distribution of receiver beliefs across treatments and guess after the potentially deceptive message.

Figure B3. Average receiver beliefs across treatments and guess after the potentially deceptive message in Receiver-Hidden



*Note*s. The figure depicts the mean reported receiver belief (y-axis) for each elicited belief and treatment (y-axis). Standard errors are plotted as vertical segments over each mean belief (bar).

The following tables present the results of ANOVA tests for differences in mean receivers' beliefs across treatments and guess (RED vs. BLUE) after the potentially deceptive message.

Table B15. ANOVA results of receivers' beliefs about the likelihood of senders choosing the deceptive message (non-Blue) across treatments and guess in Receiver-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	9890	3297	4.30	0.01
Decision	1	39692	39692	51.81	0.00
Treatment x Decision	3	2834	945	1.23	0.30
Residuals	1193	914031	766	NA	NA

Table B16. ANOVA results of receivers' beliefs about the likelihood of other receivers guessing Red after the deceptive message (non-Blue) across treatments and decision in Receiver-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	27874	9291	17.40	0.00
Decision	1	179358	179358	335.88	0.00
Treatment x Decision	3	722	241	0.45	0.72
Residuals	1193	637048	534	NA	NA

Table B15. ANOVA results of receivers' beliefs about the likelihood of senders choosing the deceptive message (non-Blue) across treatments and guess in Receiver-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	2627	3297	4.30	0.01
Decision	1	40028	39692	51.81	0.00
Treatment x Decision	3	2834	945	1.23	0.30
Residuals	1193	914031	766	NA	NA

Table B16. ANOVA results of receivers' beliefs about the likelihood of other receivers guessing Red after the deceptive message (non-Blue) across treatments and decision in Receiver-Hidden

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	104834	34945	65.441	0.00
Decision	1	194661	194661	364.541	0.00
Treatment x Decision	3	722	241	0.45	0.72
Residuals	1193	637048	534	NA	NA

The following table presents the results of Chi-squared tests for differences across treatments in consistency rates between receiver's guess and belief that the sender is deceptive in Receiver-Hidden.

Table B17. Comparison of receivers' guess-belief consistency rate in Receiver-Hidden

Treatments Compared	Mean Difference	Chi-squared test	p-value
DIRECT – IGNORANCE	-12.9	χ^2 (1, 602) = 10.16	0.00
DIRECT – PARTIAL	-0.07	χ^2 (1, 598) = 2.69	0.10
DIRECT – SILENCE	-0.07	χ^2 (1, 599) = 2.81	0.09
PARTIAL – IGNORANCE	-0.06	χ^2 (1, 602) = 2.40	0.12
SILENCE – IGNORANCE	-0.06	χ^2 (1, 603) = 2.29	0.13
SILENCE – PARTIAL	0.00	χ^2 (1, 599) = 0.00	0.97

Are deception rates in the evasion treatments lower in Sender-Open compared to Sender-Hidden?

Table B18. Analysis of deception rates across Sender-Open and Sender-Hidden

	Dependent variable:				
	Choice of deceptive option				
	(Overall)	(DIRECT)	(IGNORANCE)	(PARTIAL)	(SILENCE)
Sender-Open	-0.064***	-0.045	-0.067	-0.076*	-0.147***
	(0.018)	(0.044)	(0.045)	(0.046)	(0.045)
B(a=Red|m=non-Blue)	0.000	-0.000	0.002*	-0.001	0.002*
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.001*	-0.001	-0.002**	0.000	-0.001
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
B(others-deceive)	0.009***	0.011***	0.012***	0.011***	0.010***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.029	-0.088*	-0.049	-0.002	0.012
	(0.019)	(0.047)	(0.047)	(0.047)	(0.047)
Age	0.001	0.001	-0.001	0.004**	0.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Higher education	0.020	-0.060	-0.096	0.145**	0.085
	(0.027)	(0.072)	(0.066)	(0.067)	(0.064)
Treatment FE	Yes				
Observations	2,381	602	599	587	593

Notes: This table reports marginal effects from logit (Overall column) probit (DIRECT, IGNORANCE, PARTIAL and SILENCE columns) regressions for each treatment. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). Sender-Open is a dummy for the Sender-Open experiment. B(·) are the sender's beliefs. "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Sample characteristics and randomization check

Table B19. Sample characteristics and randomization check in Sender-Hidden

Treatment (N)	Age	Female	Higher education
DIRECT (305)	35.8 (0.63)	0.70 (0.03)	0.88 (0.02)
IGNORANCE (303)	36.0 (0.65)	0.64 (0.03)	0.84 (0.02)
PARTIAL (300)	37.3 (0.68)	0.62 (0.03)	0.88 (0.02)
SILENCE (302)	36.2 (0.70)	0.63 (0.03)	0.85 (0.02)
	H(3) = 3.31,	χ^2 (3, 1209) = 5.61,	χ^2 (3, 1203) = 2.95,
	p = 0.346	p = 0.132	p = 0.399

*Note*s. Means and standard errors (in parenthesis) in each treatment of the Sender-Hidden experiment. The last row displays p-values for the null hypothesis of perfect randomization (Chi-square test in case of binary variables and Kruskal-Wallis test in case of interval variables). "Age" is in years, "Female," and "Higher education" are dummy variables indicating female participants, and higher education (college or above).

Table B20. Sample characteristics and randomization check in Sender-Open

Treatment (N)	Age	Female	Higher education
DIRECT (303)	35.7 (0.72)	0.66 (0.03)	0.91 (0.02)
IGNORANCE (305)	37.1 (0.72)	0.60 (0.03)	0.87 (0.02)
PARTIAL (297)	36.9 (0.74)	0.64 (0.03)	0.86 (0.02)
SILENCE (299)	36.7 (0.74)	0.65 (0.03)	0.87 (0.02)
	H(3) = 2.69,	χ^2(3, 1202) = 2.45,	χ^2(3, 1194) = 3.98,
	p = 0.443	p = 0.484	p = 0.264

*Note*s. This table reports means and standard errors (in parenthesis) in each treatment of the Sender-Open experiment. The last row displays p-values for the null hypothesis of perfect randomization (Chi-square test in case of binary variables and Kruskal-Wallis test in case of interval variables). "Age" is in years, "Female," and "Higher education" are dummy variables indicating female participants, and higher education (college or above).

Table B21. Sample characteristics and randomization check in the Receiver-Hidden experiment

Treatment (N)	Age	Female	Higher education
DIRECT (299)	39.8 (0.77)	0.50 (0.03)	0.86 (0.02)
IGNORANCE (303)	40.6 (0.78)	0.49 (0.03)	0.87 (0.02)
PARTIAL (299)	40.7 (0.70)	0.48 (0.03)	0.85 (0.02)
SILENCE (300)	40.7 (0.77)	0.49 (0.03)	0.89 (0.02)
	H(3) = 1.34,	χ^2 (3, 1200) = 0.09,	χ^2 (3, 1192) = 2.14,
	p = 0.719	p = 0.993	p = 0.545

*Note*s. This table reports means and standard errors (in parenthesis) in each treatment of the Receiver-Hidden experiment. The last row displays p-values for the null hypothesis of perfect randomization (Chi-square test in case of binary variables and Kruskal-Wallis test in case of interval variables). "Age" is in years, "Female," and "Higher education" are dummy variables indicating female participants, and higher education (college or above).

Statistical tests for the comparison of simulated average payoffs

Table B22. Comparison of senders' simulated average payoffs in each treatment (t-test)

Simulation Type Compared	DIRECT	IGNORANCE	PARTIAL	SILENCE
Observed – R_p_match	t(298)=6.17, $p < 0.01$	t(302)=1.76, $p = 0.08$	t(298)=7.96, $p < 0.01$	t(299)=-2.41, $p = 0.02$
Observed – R_gullible	t(298)=-7.54, $p < 0.01$	t(302)=-15.81, $p < 0.01$	t(298)=-10.51, $p < 0.01$	t(299)=-16.96, $p < 0.01$
Observed – S_truth	t(298)=26.29, $p < 0.01$	t(302)=12.90, $p < 0.01$	t(298)=18.65, $p < 0.01$	t(299)=13.92, $p < 0.01$
Observed – S_lie	t(298)=-13.97, $p < 0.01$	t(302)=-12.75, $p < 0.01$	t(298)=-18.65, $p < 0.01$	t(299)=-13.82, $p < 0.01$

Table B23. Comparison of receivers' simulated average payoffs in each treatment (t-test)

Simulation Type Compared	DIRECT	IGNORANCE	PARTIAL	SILENCE
Observed – R_p_match	t(298)=27.28, $p < 0.01$	t(302)=40.76, $p < 0.01$	t(298)=35.28, $p < 0.01$	t(299)=46.70, $p < 0.01$
Observed – R_gullible	t(298)=-11.93, $p < 0.01$	t(302)=-9.00, $p < 0.01$	t(298)=-4.23, $p < 0.01$	t(299)=-1.10, $p = 0.27$
Observed – S_truth	t(298)=-26.29, $p < 0.01$	t(302)=-12.90, $p < 0.01$	t(298)=-18.65, $p < 0.01$	t(299)=-13.92, $p < 0.01$
Observed – S_lie	t(298)=-26.41, $p < 0.01$	t(302)=12.75, $p < 0.01$	t(298)=18.65, $p < 0.01$	t(299)=13.82, $p < 0.01$

Table B24. Comparison across treatments of players' simulated average payoffs given observed behaviour (t-test)

Treatments Compared	Sender	Receiver
DIRECT – IGNORANCE	t(594.01)=4.43, $p < 0.01$	t(527.65)=-0.56, $p = 0.57$
DIRECT – PARTIAL	t(585.12)=0.60, $p = 0.55$	t(551.48)=1.26, $p = 0.21$
DIRECT – SILENCE	t(589.05)=4.78, $p < 0.01$	t(476.21)=-0.54, $p = 0.59$
PARTIAL – IGNORANCE	t(599.21)=3.58, $p < 0.01$	t(594.6)=-2.23, $p = 0.03$
SILENCE – IGNORANCE	t(600.85)=0.36, $p = 0.73$	t(585.22)=-0.07, $p = 0.95$
SILENCE – PARTIAL	t(596.75)=-3.91, $p < 0.01$	t(560.35)=2.33, $p = 0.02$

Table B25. Comparison across treatments of players' simulated average payoffs given observed behaviour, conditional on the segment being Blue (t-test)

Treatments Compared	Sender	Receiver
DIRECT – IGNORANCE	t(558.10)=5.16, $p < 0.01$	t(558.10)=-5.16, $p < 0.01$
DIRECT – PARTIAL	t(547.19)=-0.25, $p = 0.80$	t(547.19)=0.25, $p = 0.80$
DIRECT – SILENCE	t(543.15)=4.85, $p < 0.01$	t(543.15)=-4.85, $p < 0.01$
PARTIAL – IGNORANCE	t(599.58)=-4.74, $p < 0.01$	t(599.58)=4.74, $p < 0.01$
SILENCE – IGNORANCE	t(599.82)=0.16, $p = 0.87$	t(599.82)=-0.16, $p = 0.87$
SILENCE – PARTIAL	t(596.80)=-4.49, $p < 0.01$	t(596.80)=4.49, $p < 0.01$

Table B26. Comparison across treatments of players' simulated average payoffs given observed behaviour, conditional on the segment being Red (t-test)

Treatments Compared	Sender	Receiver
DIRECT – IGNORANCE	t(591.71)=3.56, $p < 0.01$	t(591.71)=3.56, $p < 0.01$
DIRECT – PARTIAL	t(593.35)=-0.25, $p = 0.80$	t(593.35)=-0.25, $p = 0.80$
DIRECT – SILENCE	t(582.21)=3.38, $p < 0.01$	t(582.21)=3.38, $p < 0.01$
PARTIAL – IGNORANCE	t(598.39)=3.43, $p < 0.01$	t(598.39)=3.43, $p < 0.01$
SILENCE – IGNORANCE	t(599.95)=0.27, $p = 0.78$	t(599.95)=0.27, $p = 0.78$
SILENCE – PARTIAL	t(591.82)=3.23, $p < 0.01$	t(591.82)=3.23, $p < 0.01$

Probit analysis without controlling for any of the elicited beliefs

Table B27. Sender-Hidden experiment

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.046	
	(0.041)	
PARTIAL	0.089**	
	(0.041)	
SILENCE	0.078*	
	(0.041)	
EVASIONS_Pooled		0.071**
		(0.033)
Female	-0.054*	-0.054*
	(0.030)	(0.030)
Age	0.001	0.001
	(0.001)	(0.001)
Higher education	0.092**	0.094**
	(0.042)	(0.042)
Observations	1,193	1,193
Note:	*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.	

Table B28. Sender-Open experiment

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.048	
	(0.041)	
PARTIAL	0.093**	
	(0.041)	
SILENCE	0.040	
	(0.041)	
EVASIONS_Pooled		0.059*
		(0.033)
Female	0.016	0.016
	(0.030)	(0.030)
Age	-0.001	-0.001
	(0.001)	(0.001)
Higher education	0.010	0.010
	(0.045)	(0.045)
Observations	1,188	1,188
Note:	*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.	

Table B29. Receiver-Hidden experiment

	Dependent variable:	
	Guess RED	
	(1)	(2)
IGNORANCE	-0.372***	
	(0.041)	
PARTIAL	-0.208***	
	(0.045)	
SILENCE	-0.384***	
	(0.040)	
EVASIONS_Pooled		-0.284***
		(0.027)
Female	0.009	0.008
	(0.029)	(0.029)
Age	0.001	0.001
	(0.001)	(0.001)
Higher education	-0.019	-0.026
	(0.043)	(0.042)
Observations	1,188	1,188
Note:	*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.	

Probit analysis without controlling for the belief that others deceive or for gender

Table B30. Sender-Hidden experiment

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.055	0.041
	(0.042)	(0.045)
PARTIAL	0.087**	0.136***
	(0.041)	(0.043)
SILENCE	0.091**	0.130**
	(0.042)	(0.044)
B(a=Red|m=non-Blue)	0.001*	0.001
	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.000	-0.001
	(0.001)	(0.001)
B(others-deceive)		0.009***
		(0.001)
Female	-0.056*	
	(0.030)	
Age	0.001	0.002
	(0.001)	(0.001)
Higher education	0.090**	0.051
	(0.043)	(0.046)
Observations	1,193	1,193

Notes: Marginal effects from a probit regression in Sender-Hidden. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). IGNORANCE, PARTIAL and SILENCE are dummies for those treatments, DIRECT is the excluded category. B(·) are the sender's beliefs. "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Linear hypothesis tests for the comparison of coefficients of evasion treatments in Column (1):

- IGNORANCE vs PARTIAL: $\chi^2(1, 1184) = 0.60$, $p = 0.439$
- IGNORANCE vs SILENCE: $\chi^2(1, 1184) = 0.76$, $p = 0.382$
- PARTIAL vs SILENCE: $\chi^2(1, 1184) = 0.01$, $p = 0.932$

Linear hypothesis tests for the comparison of coefficients of evasion treatments in Column (2):

- IGNORANCE vs PARTIAL: $\chi^2(1, 1185) = 4.62$, $p = 0.032$
- IGNORANCE vs SILENCE: $\chi^2(1, 1185) = 4.21$, $p = 0.040$
- PARTIAL vs SILENCE: $\chi^2(1, 1185) = 0.02$, $p = 0.888$

Table B31. Sender-Open experiment

	Dependent variable:	
	Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.063	-0.009
	(0.042)	(0.047)
PARTIAL	0.098**	0.108**
	(0.042)	(0.047)
SILENCE	0.054*	0.017
	(0.042)	(0.047)
B(a=Red|m=non-Blue)	0.001*	0.000
	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.000	-0.001**
	(0.001)	(0.001)
B(others-deceive)		0.012***
		(0.001)
Female	0.016	
	(0.030)	
Age	-0.001	0.001
	(0.001)	(0.001)
Higher education	0.009	-0.002
	(0.045)	(0.050)
Observations	1,188	1,188

Notes: Marginal effects from a probit regression in Sender-Open. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). IGNORANCE, PARTIAL and SILENCE are dummies for those treatments, DIRECT is the excluded category. B(·) are the sender's beliefs. "Female" is a dummy variable indicating female participants, "Age" is in years and "Higher education" is a dummy variable indicating participants having completed higher education (college or above). Standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Linear hypothesis tests for the comparison of coefficients of evasion treatments in Column (1):

- IGNORANCE vs PARTIAL: $\chi^2(1, 1179) = 0.75$, $p = 0.387$
- IGNORANCE vs SILENCE: $\chi^2(1, 1179) = 0.05$, $p = 0.826$
- PARTIAL vs SILENCE: $\chi^2(1, 1179) = 1.16$, $p = 0.281$

Linear hypothesis tests for the comparison of coefficients of evasion treatments in Column (2):

- IGNORANCE vs PARTIAL: $\chi^2(1, 1180) = 6.54$, $p = 0.011$
- IGNORANCE vs SILENCE: $\chi^2(1, 1180) = 0.33$, $p = 0.564$
- PARTIAL vs SILENCE: $\chi^2(1, 1180) = 3.88$, $p = 0.049$

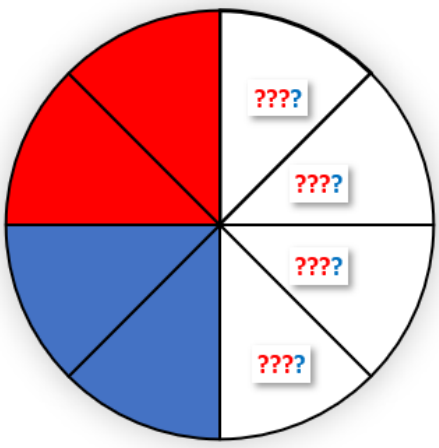
Appendix C. Preliminary Survey

C1. The Survey

Survey Procedures. The survey was conducted online prior to the experiments using Prolific (http://www.prolific.ac) and programmed using Qualtrics (http://www.qualtrics.com/). A total of 201 participants (69% female, mean age 33.5) completed the survey for a flat fee of £1 upon completion. The survey included comprehension questions, which participants had to answer correctly before proceeding to the evaluation. Experimental instructions are available at the end of this section.

Survey Design. Survey participants read about a hypothetical situation involving two parties, Person A (sender) and Person B (receiver). In particular, participants read a description of a setting resembling the setup of the actual experimental game, where an 8-segment wheel is spun, and one segment is randomly selected. The colour of the segment can be either Red or Blue, with Red being realized with probability 62.5%, and Blue with the remaining 37.5%. Half of the segments are visible, and half are hidden. Similar to the Hidden and Open evasion experiments, if the segment is visible, the sender sends a costless message to the receiver informing him about the colour of the segment; if the segment is hidden an automatic message is sent to the receiver. The receiver then makes a choice about the colour of the segment. The sender receives a bonus if the receiver guesses Red, while the receiver gets a bonus if he chooses correctly. To better visualize the different types of senders and their associated probabilities, participants were presented with the image of the wheel that would be spun which is depicted in Figure C1.

Figure C1. The 8-segment wheel



Participants rated the deceptiveness of the sent message, if the segment is visibly Blue i.e., there is a conflict of interest between the two parties. They were explained that the sender can choose between sending the truth (“The segment is BLUE”), sending a direct lie (“The segment is RED”) or sending an evasive message (message “X,”) that is the same as the automatic message in case the selected segment is hidden. Note here that in contrast to the experimental game, in the scenario used in the survey, the sender can choose between telling the truth, telling a direct lie or evade. The evasive messages available

to the sender include silence, partial truth and feigning ignorance. We further used eight more evasive messages as fillers, to ensure that participants take the survey seriously and rate the messages in a consistent manner. So, in total there are eleven possible versions of message X, although only one of these will be available to each pair of players. The versions of message X available to the sender are as follows.

x_1 = "I do not know the colour of the segment"

x_2 = "A hidden colour segment was chosen"

x_3 = "The segment is either RED or BLUE"

x_4 = "The segment **was** more likely to be RED than BLUE"

x_5 = "The segment **is** more likely to be RED than BLUE"

x_6 = "There are more RED segments"

x_7 = "There are both visible and hidden colour segments"

x_8 = "The current year is 2018"

x_9 = "Today is Friday"

x_{10} = "Today is Tuesday"

x_{11} = " " (Keep silent: a blank message containing no information)

Participants rated first the deceptiveness of the true and the direct lie message. Subsequently, they were reminded of these two ratings and judged the deceptiveness of each of the available evasive messages in a randomized order. Half of the participants judged the available messages from the perspective of the sender, and the remaining half from the perspective of the receiver. In line with the Hidden Evasion experiment, the receiver never finds out whether the message he received comes from an uninformed or a deceptive sender.,

Results

The evaluation ratings of all messages are depicted in Figure C2. Several interesting patterns can be observed eyeballing Figure C2. First, as expected, telling the truth is the least deceptive message, while telling a direct lie is the most deceptive one. Second, all the eleven evasive messages are significantly different from truth-telling (all paired t-test $p < 0.001$, see Table C1, column 2) and direct lying (all paired t-test $p < 0.001$, see Table C1, column 3). Third, we observe a large heterogeneity across participants' judgments on the evasive messages, suggesting that different messages entail different degree of deceptiveness, despite the fact their plain interpretation suggests simply the sender is uninformed. In particular, when it comes to the three evasive messages of interest that we used in the experimental games, feigned ignorance is judged harsher than partial truth followed by silence (see Table C2).

Figure C2. Deceptiveness ratings

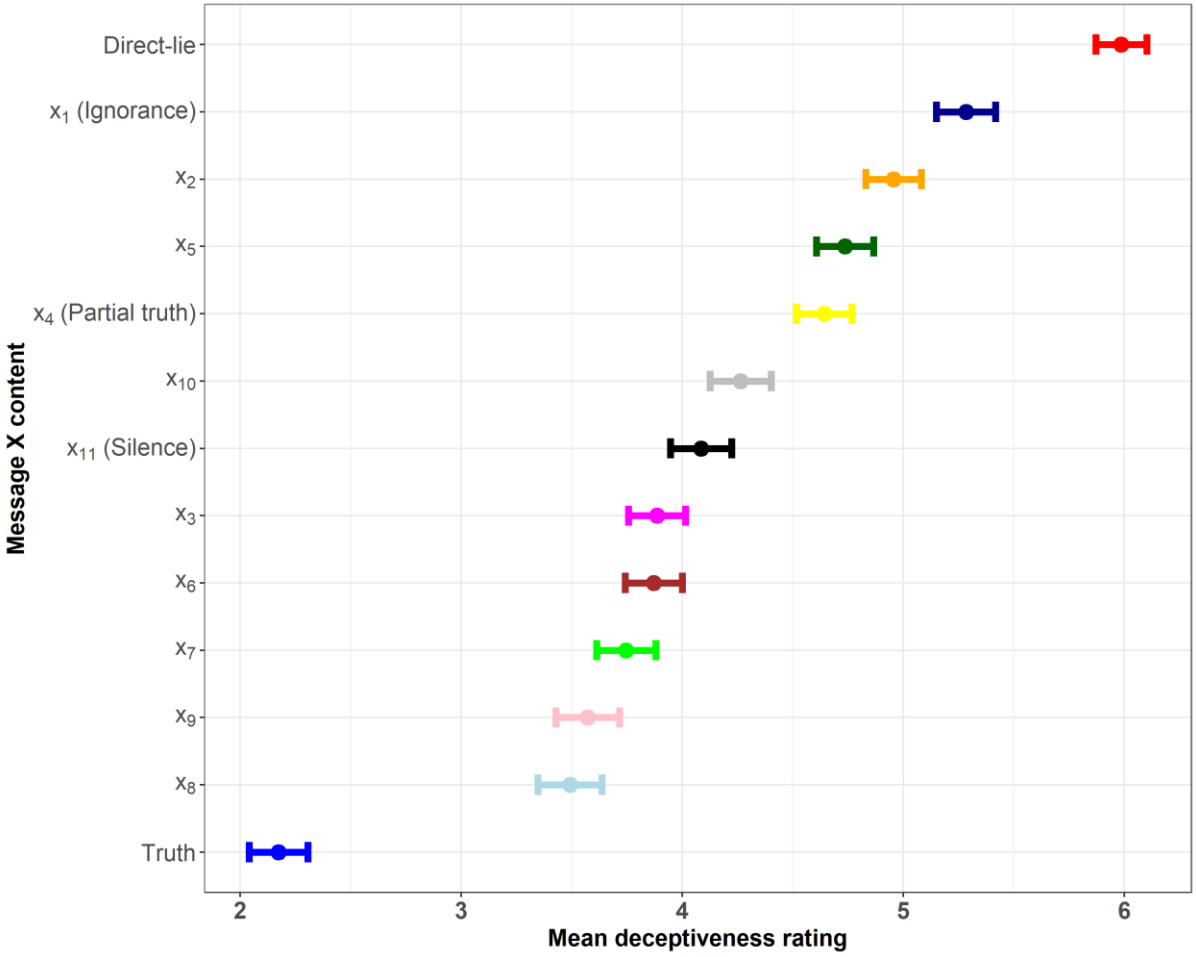


Table C1. T-test results for the comparison of all evasive messages with truth-telling and direct lying

Evasive message	Comparison with truth	Comparison with direct lie
"I do not know the colour of the segment"	$t(200) = -14.80$, $p < 0.001$	$t(200) = 4.68$, $p < 0.001$
"A hidden colour segment was chosen"	$t(200) = -13.62$, $p < 0.001$	$t(200) = 7.11$, $p < 0.001$
"The segment is either RED or BLUE"	$t(200) = -10.12$, $p < 0.001$	$t(200) = 11.93$, $p < 0.001$
"The segment was more likely to be RED than BLUE"	$t(200) = -13.62$, $p < 0.001$	$t(200) = 8.47$, $p < 0.001$
"The segment is more likely to be RED than BLUE"	$t(200) = -13.16$, $p < 0.001$	$t(200) = 7.98$, $p < 0.001$
"There are more RED segments"	$t(200) = -10.15$, $p < 0.001$	$t(200) = 11.70$, $p < 0.001$
"There are both visible and hidden colour segments"	$t(200) = -8.97$, $p < 0.001$	$t(200) = 12.00$, $p < 0.001$
"The current year is 2018"	$t(200) = -6.96$, $p < 0.001$	$t(200) = 12.96$, $p < 0.001$
"Today is Friday"	$t(200) = -7.61$, $p < 0.001$	$t(200) = 12.95$, $p < 0.001$
"Today is Tuesday"	$t(200) = -11.13$, $p < 0.001$	$t(200) = 9.46$, $p < 0.001$
"" (Keep silent: a blank message containing no information)	$t(200) = -10.58$, $p < 0.001$	$t(200) = 10.26$, $p < 0.001$

Table C2. T-test results for the comparison of feigned ignorance, partial truth and silence

Comparison	
PARTIAL – IGNORANCE	$t(200) = -3.44$, $p < 0.001$
SILENCE – IGNORANCE	$t(200) = -6.78$, $p < 0.001$
SILENCE – PARTIAL	$t(200) = -3.24$, $p = 0.001$

C2. Experimental Instructions for the Survey

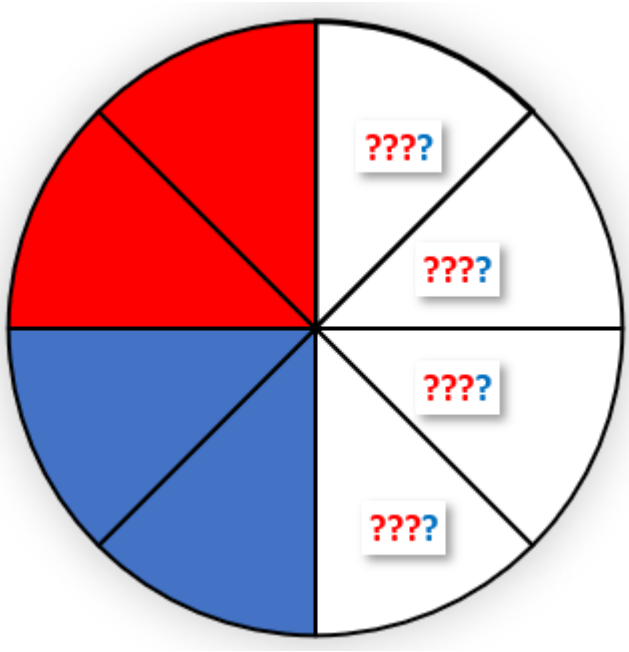
Below are the instructions for the survey. We provide the instructions from the perspective of Person A (the sender), and we use brackets ({}) to indicate the changes from the perspective of Person B (the receiver).

Welcome to this study about decision-making. You will read about a hypothetical situation involving two people, **Person A** and **Person B**, interacting in an experiment. Person A and Person B do not know one another and will never see each other.

Please read the description of the situation carefully. You will be asked questions that depend on your understanding of the situation.

----------(page break) ----------

At the beginning of the experiment a spinner like the one below with eight equal sized segments is spun and one random segment is selected.



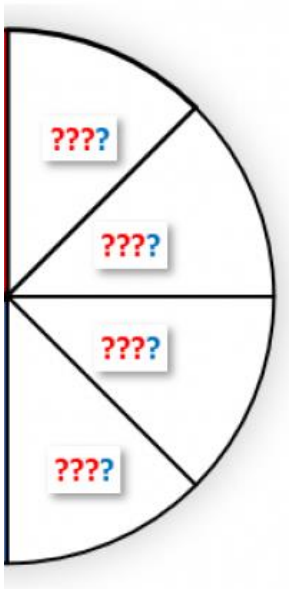
Four segments on the spinner are visibly coloured, two **RED** and two **BLUE**. The other four are hidden colour segments that have a white flap concealing the colour. Under the flap, three of these hidden colour segments are **RED** and one is **BLUE**. There is no way to know what colour is the segment if it is hidden. In total, the spinner has 5 **RED** and 3 **BLUE** segments.

----------(page break) ----------

Only Person A sees which segment is selected. Person B doesn't -- **neither during, nor after the experiment.**

Once the segment is selected, Person A must send a message to Person B.

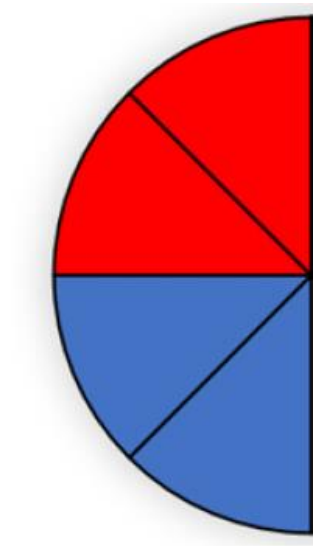
----------(page break) ----------



If a hidden colour segment is selected, Message X will automatically be sent from Person A to Person B without Person A having an option to choose this message.

The possible contents of Message X will be described later.

----------(page break) ----------



If a visibly coloured segment is selected, Person A will then choose a message to send to Person B. Person A's options are:

1. A message reporting the true colour of the segment, i.e. **RED** if the segment is **RED** and **BLUE** if the segment is **BLUE**.
2. A message reporting the opposite colour of the segment, i.e. **RED** if the segment is **BLUE** and **BLUE** if the segment is **RED**.

3. The same Message X that would be automatically sent if a hidden colour segment were selected.

----------(page break) ----------

The message from Person A will be shown to Person B, and then Person B will choose either **RED** or **BLUE**. The message will be the only information Person B has regarding the outcome of the spin.

Importantly, the colour chosen by Person B will determine the payments in the experiment.

Person A's bonus: If Person B chooses **RED** then Person A earns a bonus.

Person B's bonus: If Person B chooses the same colour as the segment then Person B earns a bonus.

Person A's bonus and Person B's bonus are equal.

Person B will **never** know if the selected segment was visible or hidden.

----------(page break) ----------

The following table presents all possible payment situations.

	Person B chooses RED	Person B chooses BLUE
RED segment	**Both** Person A and Person B get a bonus	**Neither** Person A **nor** Person B get a bonus
BLUE segment	**Only** Person A gets a bonus	**Only** Person B gets a bonus

----------(page break) ----------

There are eleven possible versions of Message X, although only one of these is available to each pair of players.

Below are the eleven possible versions of Message X:

- "I do not know the colour of the segment"
- "A hidden colour segment was chosen"
- "The segment is either **RED** or **BLUE**"
- "The segment **was** more likely to be **RED** than **BLUE**"
- "The segment **is** more likely to be **RED** than **BLUE**"

- "There are more **RED** segments"
- "There are both visible and hidden colour segments"
- "The current year is 2018"
- "Today is Friday"
- "Today is Tuesday"
- " " (Keep silent: a blank message containing no information)

----------(page break) ----------

Before you continue, however, please click below to indicate that you are not a robot.

----------(page break) ----------

The next questions ask you about the situation we have just described.

Which message will be sent to Person B when a hidden colour segment is selected?

- It depends on which message Person A will choose to send
- Message X is automatically sent

If a hidden colour segment is selected and Person B chooses **RED**, will Person A earn a bonus?

- Yes
- No
- It depends on the colour of the segment

If a visible colour segment is selected and Person B chooses **BLUE**, will Person A earn a bonus?

- Yes
- No
- It depends on the colour of the segment

If a hidden colour segment is selected and Person B chooses **BLUE**, will Person B earn a bonus?

- Yes
- No
- It depends on the colour of the segment

If the selected segment is **BLUE** and Person B chooses **RED**, who will earn a bonus?

- Person A
- Person B

- Both Person A and Person B
- Neither Person A nor Person B
- It depends on the message Person A sent

If Person B receives Message X, what can Person B infer?

- A visible colour segment was selected
- Either a hidden or a visible colour segment was selected

----------(page break) ----------

Imagine you are **Person A {Person B}** and that the segment is visible and **BLUE**.

Given these circumstances, please rate how deceptive it is for you to send each of the following messages to Person B.

{Given these circumstances, please rate how deceptive it would be for Person A to send you each of the following messages.}

Use a scale from 1 to 7 where 1 stands for “Not at all deceptive” and 7 stands for “Very deceptive."

----------(page break) ----------

Given that you are Person A, and the segment is visible and **BLUE**, how deceptive it is for you to send the following message to Person B?

{Given that you are Person B, and the segment is visible and **BLUE**, how deceptive would it be for Person A to send you the following message?}

“The segment is **RED**”

1	2	3	4	5	6	7
(Not at all deceptive)						(Very deceptive)

----------(page break) ----------

Given that you are Person A, and the segment is visible and **BLUE**, how deceptive it is for you to send the following message to Person B?

{Given that you are Person B, and the segment is visible and **BLUE**, how deceptive would it be for Person A to send you the following message?}

“The segment is **BLUE**”

1	2	3	4	5	6	7
(Not at all deceptive)						(Very deceptive)

----------(page break) ----------

Remember that you gave the message "The segment is **BLUE** " a rating of … and the message "The segment is **RED**" a rating of … .

Given that you are Person A, and the segment is visible and **BLUE**, how deceptive it is for you to send the following message to Person B?

{Given that you are Person B, and the segment is visible and **BLUE**, how deceptive would it be for Person A to send you the following message?}

"Placeholder for different versions of Message X presented in a randomised order"

1	2	3	4	5	6	7
(Not at all deceptive)						(Very deceptive)

----------(page break) ----------

Can you explain the reasoning behind your choices in the task? Specifically, how did you decide what rating to give to each message?

----------(page break) ----------

Thank you!

You're almost done, there are just another few questions for you to answer.

Q1. What is your gender?

- Female
- Male
- Other (please describe if you wish)
- I would prefer not to answer

Q2. What is your age?

- Please write your age in years ___
- I would prefer not to answer

Q3. What is your marital status?

- Single, never married
- Married or domestic partnership
- Divorced
- Widowed
- Separated
- I would prefer not to answer

Appendix D. Experimental Instructions

Below are the instructions for DIRECT from the sender's perspective for both experiments. We provide the instructions for the Hidden Evasion experiment, and we use brackets ({}) to indicate the changes in the Open Evasion experiment. The instructions for all evasion treatments were based on these treatments, with the corresponding modifications according to the treatment. Full set of instructions for the evasive treatments can be obtained from the authors.

Welcome and thank you for participating in this study. Every participant will receive £1 upon completion, and will earn an extra bonus.

All your decisions will be **anonymous** and no identifying information will be shared with other participants, **during** or **after** the study.

Please read the instructions carefully. During the study you will be asked a few questions to ensure that the instructions have been properly explained.

----------(page break) ----------

Instructions

Participants in this study take on one of two roles: **Sender** and **Receiver**. You will be randomly assigned one of these roles, but you don't know which one yet.

You will be randomly paired with another participant (another Prolific Academic worker) who will take the other role. If you are the Sender, the other participant will be the Receiver, and vice versa.

You will keep the same role for the entire study.

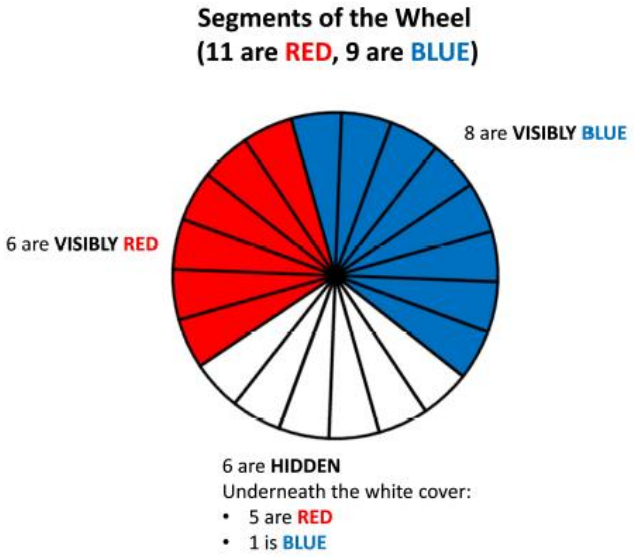
What follows is a description for **both** roles. You will learn your role after you have studied this description.

----------(page break) ----------

General description of the study

The following 20-segment wheel will be spun once to randomly select one segment. Each segment is **equally** likely to be selected.

As detailed below, there are "visibly" **RED** segments, "visibly" **BLUE** segments, and "hidden" segments that have a white cover but are either **RED** or **BLUE** underneath.



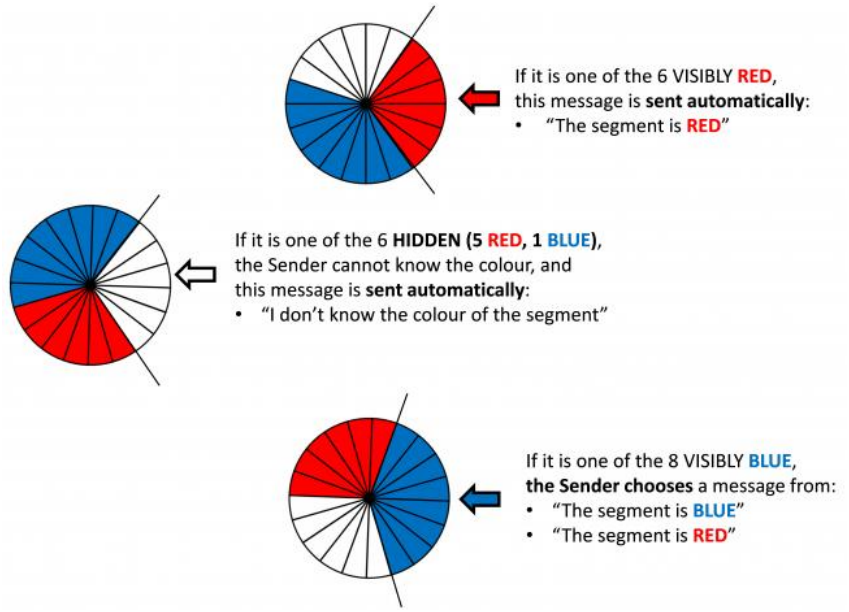
The Sender will observe the spin and its outcome, but the Receiver will not. Note that if a hidden colour segment is selected, the Sender cannot know whether the segment is **RED** or **BLUE**.

After the spin the Receiver will receive a message and then guess whether the segment is **RED** or **BLUE**. The Receiver earns more money if that guess is correct. The Sender earns more money if the Receiver guesses **RED**, no matter which colour the segment is.

More details about how the message is chosen and the exact earnings are shown next.

----------(page break) ----------

Before guessing the colour, the Receiver receives a message, which depends on the randomly selected segment as shown below:



Note that the Sender chooses a message only when the visibly **BLUE** segment is selected. In summary:

If the message is	Then the selected segment is
"I don't know the colour of the segment"	Hidden
"The segment is **BLUE**"	Visibly **BLUE**
"The segment is **RED**"	Either visibly **RED** or visibly **BLUE**

The Receiver will **never** directly observe which segment is selected -- neither **during**, nor **after** the study. The message is the **only** information the Receiver will have before guessing the colour.

The Receiver will **never** be told if the selected segment was **visible** or **hidden**, or if the message was **chosen by the Sender** or **sent automatically**.

{**Open Evasion experiment:**
The Receiver will **never** directly observe which segment is selected **during** the study. The message is the **only** information the Receiver will have before guessing the colour.

At the end of the study, **but only after guessing the colour and receiving the payment**, the Receiver will be told **more** about the Sender's decision making. Specifically, the Receiver will learn if the selected segment was **visible** or **hidden**, and if the message was **chosen by the Sender** or **sent automatically**.
}

----------(page break) ----------

Earnings: The Sender earns a £2 bonus only if the Receiver guesses **RED**; otherwise the Sender earns £1. The Receiver earns a £2 bonus only if their guess matches the actual colour; otherwise the Receiver earns £1.

The four possibilities are summarized below:

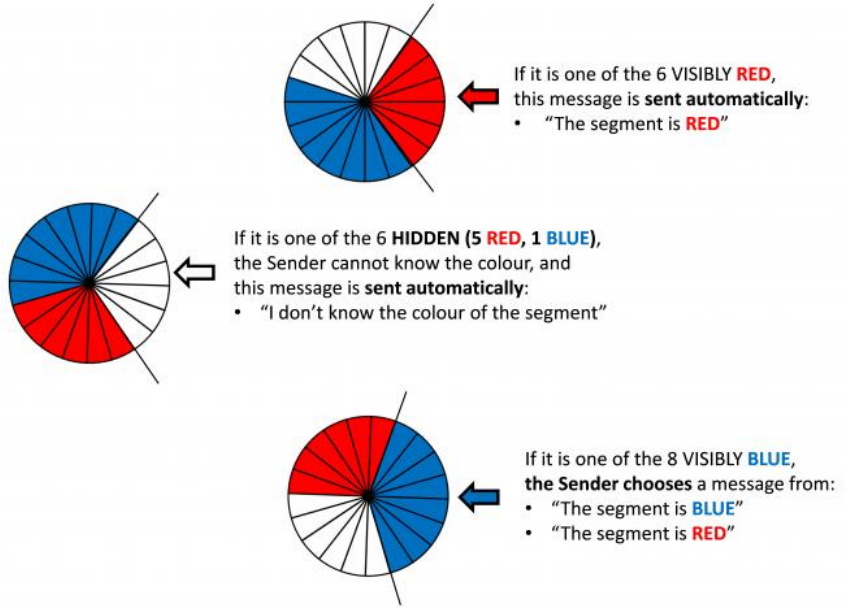
	Sender's bonus	Receiver's bonus
The segment is **RED** and the Receiver guesses **RED**	£2	£2
The segment is **RED** and the Receiver guesses **BLUE**	£1	£1
The segment is **BLUE** and the Receiver guesses **RED**	£2	£1
The segment is **BLUE** and the Receiver guesses **BLUE**	£1	£2

----------(page break) ----------

Summary

Step 1. A segment of the wheel is randomly selected. There are 11 **RED** and 9 **BLUE** in total.

Step 2. The Sender observes the segment and a message is sent to the Receiver as shown below:



Step 3. The Receiver guesses the segment's colour.

Step 4. The Sender earns a £2 bonus only if the Receiver guesses **RED**, and £1 otherwise. The Receiver earns a £2 bonus only if their guess matches the actual colour of the segment, and £1 otherwise.

Remember: The message is the **only** information the Receiver will have before guessing the colour of the segment. The Receiver will **never** be told if the selected segment was **visible** or **hidden**, or if the message was **chosen by the Sender** or **sent automatically**.

{**Open Evasion experiment**:

Remember: The Receiver will **never** directly observe which segment is selected **during** the study. The message is the **only** information the Receiver will have before guessing the colour. At the end of the study, **but only after guessing the colour and receiving the payment**, the Receiver will be told **more** about the Sender's decision making. Specifically, the Receiver will learn if the selected segment was **visible** or **hidden**, and if the message was **chosen by the Sender** or **sent automatically**.
}

----------(page break) ----------

This is the end of the instructions. Next you will be asked a few questions about these instructions.
Please review them before continuing.

----------(page break) ----------

Before you continue, however, please click below to indicate that you are not a robot.

----------(page break) ----------

We will now ask you some questions to ensure that the instructions are clear. You will be able to proceed with the study once you have answered all questions correctly.

Question 1. Which message will be sent to the **Receiver** when a hidden colour segment is selected?

- It depends on which message the Sender will choose
- "I don't know the colour of the segment"
- "The segment is **RED**"

----------(page break) ----------

Question 2. If a hidden colour segment is selected and the Receiver guesses **RED**, will the **Sender** earn the high (£2) bonus?

- Yes, irrespective of the actual colour of the segment
- No, irrespective of the actual colour of the segment
- It depends on the actual colour of the segment

----------(page break) ----------

Question 3. If a visible colour segment is selected and the Receiver guesses **BLUE**, will the **Sender** earn the high (£2) bonus?

- Yes, irrespective of the actual colour of the segment
- No, irrespective of the actual colour of the segment
- It depends on the actual colour of the segment

----------(page break) ----------

Question 4. If a hidden colour segment is selected and the Receiver guesses **BLUE**, will the **Receiver** earn the high (£2) bonus?

- Yes, irrespective of the actual colour of the segment
- No, irrespective of the actual colour of the segment
- Only if the actual colour of the segment is **BLUE**

----------(page break) ----------

Question 5. If the Receiver receives the message "The segment is **RED**," what is the selected segment?

- It can only be visibly **RED**
- It can only be visibly **BLUE**
- Either visibly **RED** or visibly **BLUE**

----------(page break) ----------

{Open Evasion experiment:

Question 6. Will the Receiver learn whether the selected segment was visible or hidden and if the message they received was chosen by the Sender or sent automatically?

- No, the Receiver will never learn
- Yes, but only after guessing the colour and receiving the payment
- Yes, before guessing the colour and receiving the payment

}

----------(page break) ----------

You have answered all questions correctly and can now proceed with the study. Press the button below to continue to the next page where you will observe your randomly assigned role for this study.

----------(page break) ----------

Your Role

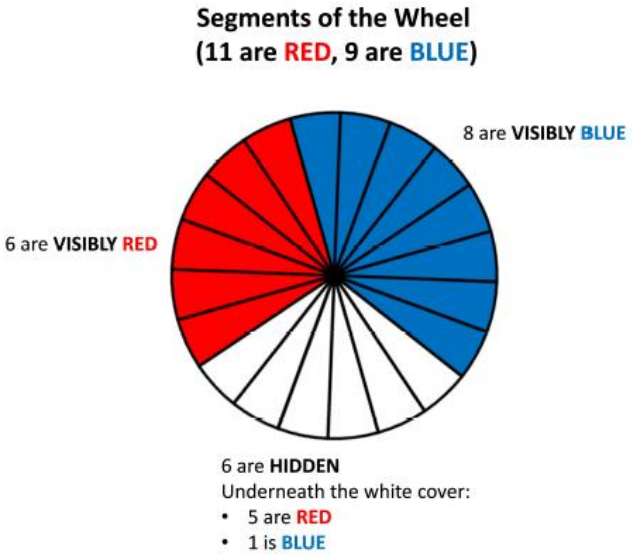
Your role in this study is that of: **Sender.**

Next, you will be asked to make your decision as a Sender.

----------(page break) ----------

Decision Screen

Recall the spinner that determines the colour of the segment:



We will now determine in advance the message that will be sent to the Receiver for each possible segment. If the segment is visibly **RED**, the message "The segment is **RED**" will be sent automatically. If the segment is hidden, the message "I don't know the colour of the segment" will be sent automatically. But if the segment is visibly **BLUE** you choose which message to send. Your choice will be implemented once the segment is selected, **only if** the segment is visibly **BLUE**.

Please choose the message you would like to send to the Receiver if the segment is visibly **BLUE** by selecting one of the following options:

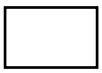
- "The segment is **BLUE**"
- "The segment is **RED**"

----------(page break) ----------

Before the spin, we will ask you what percent of participants you believe made certain decisions.

You will earn a bonus of £0.10 for each question you answer accurately (within 3 percentage points of the correct answer).

Here is the first question:
Please type a number from 0 to 100 to estimate the percent of **Receivers** in this study who guess **RED** after receiving the message "The segment is **RED**."



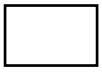
----------(page break) ----------

Before the spin, we will ask you what percent of participants you believe made certain decisions.

You will earn a bonus of £0.10 for each question you answer accurately (within 3 percentage points of the correct answer).

Here is the second question:

Please type a number from 0 to 100 to estimate the percent of **Receivers** in this study who guess **RED** after receiving the message “The segment is **BLUE**."



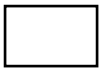
----------(page break) ----------

Before the spin, we will ask you what percent of participants you believe made certain decisions.

You will earn a bonus of £0.10 for each question you answer accurately (within 3 percentage points of the correct answer).

Here is the third question:

Please type a number from 0 to 100 to estimate the percent of **Senders** in this study (including you) who chose to send the message “The segment is **RED**," while the actual segment was visibly **BLUE**.



----------(page break) ----------

We will now select one of the 20 segments to determine which message will be sent to the Receiver.

Press the button below to spin the spinner.

----------(page break) ----------

The randomly selected segment is ___.

Therefore, the message ___ will be sent.

We will next send the message to the Receiver who will then have to guess whether the segment is **RED** or **BLUE**.

We will inform you of your bonus payments within 21 days.

----------(page break) ----------

Thank you! You're almost done, there are just another few questions for you to answer.

In a sentence or two, please describe the reasoning underlying your choice of which message to send if the segment was visible and **BLUE**.

----------(page break) ----------

What is your gender?

- Male
- Female
- Other (Please describe if you wish)
- I would prefer not to answer

What is your age?

- Please write your age in years
- I would prefer not to answer

What is the highest level of education you have completed?

- Less than secondary school
- Secondary school
- College or 6th form
- Undergraduate University degree
- Masters degree
- Doctoral or professional degree (JD, MD, PhD)
- Other (Please specify)
- I would prefer not to answer

----------(page break) ----------

You will be informed about your total earnings within 21 days. Please provide your Prolific ID number.