Alempaki, Despoina; Burdea, Valeria; Read, Daniel

Working Paper

Deceptive Communication

CESifo Working Paper, No. 9286

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Alempaki, Despoina; Burdea, Valeria; Read, Daniel (2021) : Deceptive Communication, CESifo Working Paper, No. 9286, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/245467

CESifo WORKING PAPERS

9286
2021

September 2021

Deceptive Communication

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

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 predictions in an experimental sender-receiver setting. We find senders prefer to deceive through evasion rather than direct lying. This is because they do not want to deceive others, and they do not want to be seen as deceptive. The specific language of evasion does not matter. The results suggest deception should be tested in more naturalistic contexts with richer language.

JEL-Codes: C910, D820, D830, 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, 2021
We are grateful to Daniele Nosenzo, Collin Raymond, and Daniel Seidmann for very valuable feedback. We also thank participants at the ESA 2020 Global meeting, the Winter School on Credence Goods, Incentives and Behavior 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 and the Birmingham-Warwick Strategic Information Network for useful comments and suggestions. Financial support from the ESRC's Network for Integrated Behavioural Science programme (ES/K002201/1 and ES/P008976/1) is gratefully acknowledged. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

Economic experiments have consistently demonstrated that people are averse to lying, i.e., making statements they believe to be false (see Mahon, 2015 for a formal definition), even when lying would benefit them materially and they cannot be detected or punished (e.g., Bucciol and Piovesan, 2011; Fischbacher and Föllmi-Heusi, 2013; Gerlach et al., 2019; Gneezy, 2005; Mazar et al., 2008). This fundamental lying aversion has been attributed to a psychological cost of lying mainly driven by two factors: a preference for being honest, which produces intrinsic costs of lying, and a preference for being *seen* as honest, which produces social image costs (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, and Bénabou and Tirole, 2006; Bernheim, 1994; Ellingsen and Johannesson, 2008 on social image).

Lying is not the only way to attempt deception. A famous non-lying attempt occurred when Bill Clinton denied his relationship with Monica Lewinsky (discussed in Carson, 2010; Rogers et al., 2017) by insisting that "there *is* no relationship" between them, pointedly using the present tense. This was not a lie since it did not contradict what Clinton believed – the relationship was over. But Clinton undoubtedly hoped to deceive at least some people. Importantly, Clinton apparently believed his carefully chosen words were more acceptable than a direct lie. Years later, he admitted this to journalist Jim Lehrer, observing his statement "was an intentional dodge because … I respect you. I didn't want to lie to you."[1]

Clinton's confession suggests it is easier to deceive through some methods than others, even if all methods have intentions and consequences in common. Another way to say this is that some deceptive communications are less psychologically costly than others. Indeed, rather than a simple dichotomy between truth and non-truth, there is a continuum in terms of psychological costs varying from telling "the truth, the whole truth and nothing but the truth," through various states in which some of the truth is withheld or bent a little or even a lot – we will use the term *evasion* to describe these intermediate states – through to direct lying or the assertion of falsehoods regarding instrumental information. Clinton's evasion was certainly not on the "truth" end of this continuum. To use the terms developed by Grice (1975), his statement violated the maxims of relation and quantity (as it did not relevantly and informatively answer the question of whether there was an improper relationship), yet it was also not an outright lie because it did not violate the maxim of quality (he believed his answer was true at the time).[2]

[1] PBS 1998, the interview is available at https://www.youtube.com/watch?v=XBzHnZiSv7U.

[2] See Danziger (2010), for an in-depth discussion of false utterances based on Grice's maxim of quality.

Despite the appeal, from the perspective of a communicator, of evasions over direct lies, there has been limited empirical research into their psychological costs. We fill the gap by providing a systematic analysis of the psychological cost of attempts to deceive by means of evasion. We use both theory and experiments to examine whether and when people find evasion easier than direct lying – even when the material consequences are held constant.

While the extremes of the truth-lie continuum are easily characterised, evasions are like Tolstoy's unhappy families in that they can take many forms, and like Wittgenstein's games in that while evasions are bound together by a family resemblance, there may be no definition that captures all evasions and excludes everything else.[3] We leave the search for precise definitions to those more qualified than us and focus our attention on three clearly identifiable and commonplace evasions: feigning ignorance, telling partial truths and remaining silent.

Feigning ignorance is a favourite among politicians. For example, when President Trump was asked whether he had been tested for COVID-19 on the day of the first presidential debate with Joe Biden in 2020, he replied by saying "I don't know. I don't even remember."[4] Feigning ignorance is not only used by politicians but is often employed in fiduciary relationships and by ordinary people in their everyday transactions. A seller of a house, for instance, might claim, when asked by potential buyers, that she does not know of any concerns, even though she knows there is a troubling report about subsidence lying on her desk.

The second class of evasion, illustrated by Clinton's testimony, is conducted by telling partial truths, in which only some of the truth is given and not the most relevant part. The same house seller asserts a partial truth when she says the furnace has been operating without trouble for twenty years (when, in fact, it is now in its twenty first year and is starting to act up).

The third class of evasion is remaining silent. The house seller might be selling because the neighbours are noisy and threatening, yet does not mention this to potential sellers, who are always invited for viewings when the neighbours are away or in a drunken sleep. Silence

[3] Evidence for the plurality of these intermediate states comes from major fact-checking organizations like Snopes (www.snopes.com), and Politifact (https://www.politifact.com/truth-o-meter/) or journals like the Washington Post (https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker/) that use facts to determine the truthfulness of factual claims in news articles, political speeches, social media posts etc. None of these sources uses a binary Truth-Lie scale to classify the statements of interest. For instance, PolitiFact's Truth-o-Meter uses a six-scale rating ("True," "Mostly True," "Half True," "Mostly False," "False," and "Pants on Fire"), and interestingly so, 54% of their ratings lie in the intermediate range between "Mostly True," and "Mostly False". Similar examples besides politics abound. For example, Gillespie (2008) suggests ten signs to detect greenwashing when it comes to sustainability claims by companies, ranging from "fluffy language: words or terms with no clear meaning" all the way to "irrelevant claims: emphasizing one tiny green attribute when everything else is un-green" and "outright lying: using totally fabricated claims or data" (Gillespie, 2008).

[4] The interview can be retrieved at: https://www.nbcnews.com/video/trump-claims-he-doesn-t-remember-if-he-was-tested-for-covid-before-the-debate-93934149786.

abounds in yet more momentous situations. One recent example occurred in negotiations between Gilead Sciences, maker of the potential COVID-19 treatment Remdesivir, and the EU. During negotiations Gilead did not disclose that not yet public clinical trials showed the drug was ineffective (Cohen and Kupferschmidt, 2020). They concluded negotiations one week before the clinical trial outcomes were made available to the EU.

There is a long tradition of viewing some or all evasions as more morally acceptable than lying. For instance, parents routinely enjoin their children to choose silence if they "don't have anything nice to say." Religious teachers often condemn lies while permitting or even condoning evasion (e.g., Corran, 2018).[5] Under persecution, for instance, most major world religions have found ways to allow their members to conceal their faith or even to act "as if" one's beliefs were different than they were (Kuran, 1995). Professional bodies also forbid direct lies but permit evasion. For example, the American Medical Association allows caregivers to withhold information for the comfort of the patient, at least for some period (see e.g., American Medical Association, Opinion 8.082; Bok, 1978). In business dealings, lying by telling a demonstrable falsehood is (typically) unacceptable, but almost any other deception is fair game. James J. White (1980) famously summed up the attitude: "On one hand the negotiator must be fair and truthful; on the other he must mislead his opponent." White's observation captures a widely held view that there are times when you are entitled to act in ways that lead others to believe something other than the truth, but these ways should not involve lying. Relatedly, academic researchers find it more defensible to selectively report studies that "worked," compared to falsifying data (John et al., 2012), or to omit information rather than explicitly lie to subjects in experiments (e.g., Charness et al., 2020; Hertwig and Ortmann, 2008; Hey, 1998; Krawczyk, 2019; McDaniel and Starmer, 1998).[6]

We investigate whether people prefer to evade rather than to lie directly, and if so to test for differences in the associated psychological costs. We begin our analysis with a theoretical framework based on an asymmetric information setting between an informed sender and an uninformed receiver where the sender has a material incentive to send a deceptive

[5] For instance, this is how the 13th century scholar St. Raymond of Penyafort's analysed the famous case of whether you should lie to murderers seeking to kill a man concealed in your house (cited in Slater, 1910): "The owner of the house where the man lies concealed, on being asked whether he is there, should as far as possible say nothing. If silence would be equivalent to betrayal of the secret, then he should turn the question aside by asking another — How should I know? — or something of that sort."

[6] McDaniel and Starmer (1998) for instance use the term "economy with the truth" yet call it "perfectly legitimate," while Hay (1998) argues "there is a world of difference between not telling subjects things and telling them the wrong things. *The latter is deception, the former is not*."

message to the receiver. This framework captures essential differences between forms of deception by differentiating between four possible psychological costs to the message sender: 1) a *deception* cost, incurred when the sender acts on the intention to create or maintain a false belief on the part of the receiver; 2) a *falsehood* cost, incurred when uttering a statement the sender believes to be false (note that deception and falsehood combined are what is conventionally called "lying"); 3) an *influence* cost, which increases in the likelihood that the receiver will take the wrong action based on the sender's message; 4) a *social image* cost, associated with the receiver's inferences about the sender's honesty.

The sum of these costs will generally be larger when the message is a direct lie than when it is an evasion. Compared to direct lies, evasions often have a lower falsehood cost (as do partial truth or silence where no untrue statement is uttered), and a lower influence cost (as do all three evasions we consider since they do not definitely guide the receiver towards the incorrect course of action). Evasions also often have a lower social image cost than direct lies because they permit plausible deniability: receivers can more easily verify direct lies are fraudulent, but not evasions. The main prediction derived from our theoretical framework is therefore that people will be more likely to evade than to lie directly. A second prediction is that people will be more likely to deceive via silence, followed by partial truth and then feigned ignorance. Moreover, since evasions often incur a lower social image cost because the receiver cannot find out they were deceived, the likelihood of choosing evasion compared to a direct lie should drop when the sender knows the receiver will find out.

We test these predictions using two pre-registered experiments having over 2,400 participants. Our experimental setup is based on cheap-talk sender-receiver games which have an extensive theoretical (e.g., Crawford and Sobel, 1982; Sobel, 2020), and empirical (e.g., Blume et al., 2020; Crawford, 1998; Gneezy, 2005) history. In our game there are two possible states of the world, *Red* and *Blue*. The sender receives a private signal that either specifies the state (e.g., "it is Red"), or leaves it uncertain (e.g., "it is either Red or Blue"). The sender then sends a cheap-talk message to the receiver that can be either truthful or deceptive. After receiving this message, the receiver chooses an action which determines the payoff for both players. There is a conflict of interest: the sender always wants the receiver to choose Red (and hence to believe the state is Red), whereas the receiver wants to choose the correct colour (and hence to believe the truth, whether Blue or Red). When the state is Blue, the sender faces a conflict of interest and will have an incentive to send a potentially deceptive message, such as a lie or evasion (hereafter just called a deceptive message). Our parameters are selected so that

the monetary incentive for senders to deceive via a direct lie does not exceed the monetary incentive to deceive via evasion.

We compare senders' behaviour across four treatments. In the direct lie treatment (DIRECT), 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).[7] The four treatments not only allow us to measure any differences in the psychological costs between a direct lie and evasion, but they also enable us to examine whether the language of evasion matters. As we explain later, the evasion treatments allow for plausible deniability, because the sender might really be uninformed and innocently send the evasive message, so the receiver cannot infer from their payoff whether the sender evaded or told the truth. Direct lies are not deniable because the receiver can infer they were lied to. The feedback participants get in Experiment 1, the *Hidden Evasion* experiment, maintains this property. Therefore, the social image cost is lower for an evasion than a direct lie. In Experiment 2, the *Open Evasion* experiment, we isolate the role of the social image cost by making it common knowledge in all treatments that the receivers will learn explicitly, after their guess, whether the sender disguised the truth by evading or by telling a direct lie.

We provide four main findings. First, in the Hidden Evasion experiment, DIRECT has a lower deception rate than all the evasion treatments, significantly so than PARTIAL and SILENCE. Second, social image costs play a considerable role, as the difference between DIRECT and the evasion treatments is reduced in the Open Evasion experiment, where the DIRECT versus SILENCE comparison ceases to be significant. However, social image costs are not the only driver of the difference between direct lies and evasion, since in the Open Evasion experiment, there remains significantly less deception in DIRECT than PARTIAL. This suggests that some evasions are associated with significantly lower intrinsic costs. Third, the language of evasion does not matter. In both experiments, we find statistically indistinguishable deception rates across the three types of evasion.

The fourth finding relates to another channel through which deception rates might differ -- senders' expectation about the potential benefits from a deceptive message. It might be, for instance, that senders evade more frequently than lying directly because they believe an evasive message is more likely to be interpreted in their favour by receivers. An incentivized elicitation

[7] Throughout the paper, for ease of exposition, we use uppercase letters to refer to the treatments' names, and lowercase letters to refer to types of deception in general.

of senders' beliefs about receivers' actions suggests this is not the case. If anything, senders believe that receivers are more likely to take the action favourable to the sender after a direct lie rather than an evasion. Taken together, our results suggest evasion is less psychologically costly than lying directly, and this is due to lower image as well as intrinsic costs.

This work has important policy implications because it can help us design interventions to reduce deception by considering the variety of channels through which behaviour is affected. For instance, while interventions that rely on reputation systems could be effective in reducing direct lies, where the deception can be relatively easily verified, they might be less effective for evasion, where the deceiver is less likely to be caught and held accountable. At the same time, our paper offers important insights into the pervasiveness of deception, and on how it might be even more widespread and costly than current best estimates. The major challenge in any attempt to measure the economic cost of deception is that it is usually estimated via cases of detected fraud, and as such, we can never know with certainty whether the cases we observe represent the whole iceberg. The fact that evasion is by its nature less detectable and certainly more deniable than direct lies, adds an extra layer of complexity in the already difficult task of calculating the societal cost of deceptive communications. As such, deception may in fact be more of a social and economic problem than previous literature has suggested (e.g., Egan et al., 2019; Gurun et al., 2018; Johnson et al., 2019).

Earlier studies have supported the view that people might refrain from telling direct lies when evasion is possible.[8] Serra-Garcia et al. (2011) show that sometimes senders use vague messages instead of precise but untruthful ones to disguise the truth in the context of a public good game. Similarly, senders frequently stay silent (e.g., Leibbrandt et al., 2017; Sánchez-Pagés and Vorsatz, 2009) or declare ignorance by pretending not to know (e.g., Khalmetski et al., 2017; Khalmetski and Tirosh, 2012) instead of telling a blatant lie in cheap-talk games. Also related is the study by Turmunkh et al. (2019), who analyse data from a TV game where players make non-binding pre-play statements about their willingness to cooperate in a prisoners' dilemma and find that many players who defect use malleable statements (evasions) to disguise their intentions rather than direct lies claiming to cooperate.

[8] Relevant to this work are also studies from various other disciplines: law (Schauer and Zeckhauser, 2007) psychology (e.g., Rogers and Norton, 2011; Rogers et al., 2017), marketing (e.g., Bickart et al., 2015; Kang et al., 2020) and philosophy (e.g., Carson, 2010; Cohen and Zultan, 2021). Closely related is also the important distinction made by scholars 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).

A common feature in earlier experimental studies is that senders have three options: tell the truth, lie, or evade. Our objective here is not to study whether people would opt for direct lies when evasion is possible. Instead, we want to understand whether the preference for evasion over direct lying is due to differences in psychological costs characterizing each communication in isolation, and not due to differences in perceived relative benefits which might arise when given all options side by side.[9] Conversely, in our experiment the sender has only two options; they can either tell the truth or deceive, with some being able to deceive by direct lying and others by evasion. Our study is (to the best of our knowledge) the first to provide a clean and direct test of whether evasion is less psychologically costly than outright lying. In addition, we are the first to systematically contrast the four different types of deception in a unified framework, to isolate the role of the social image cost in making evasion more attractive as a mean of deception, and to compare senders' beliefs about receivers' scepticism toward different forms of deceptive communication.

The remainder of the paper is organized as follows. Section 1 develops the theoretical framework. Section 2 describes the experimental design and procedures. Hypotheses are given in Section 3, experimental results in Section 4, and a discussion in Section 5.

1. Theoretical Framework

In this section we formally describe and analyse the setting of interest. We sketch a simple model to show how behavioural differences can be driven by differences in psychological costs associated with direct lying and evasive communications. The parameters used in the model have been chosen so they can be directly applied to our experiments, and in the following description we stick closely to our experimental design though the framework has more general application. That being said, there are situations not captured by our framework where the interactions between communication costs might lead to different conclusions.

1.1. The game

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.

[9] In Khalmetski et al. (2017), for example, 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.

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 $\Pr(\theta_1 = Blue) = 1 - \Pr(Red) = \frac{9}{20}$. 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.[10] 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. 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, about the sender's information type, θ_2, 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 will 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.[11] 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, $m \in M = \{Blue, Red, x\}$. Note that the message $\{Blue\}$ is perfectly informative, since it can only be sent when the sender knows θ_1 is Blue.

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

[10] 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.

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

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 $g > l$.

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

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

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

Definitions. Before describing players' utilities, it is useful to 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 the literal meaning of that message is that fact. For example, the literal meaning of "the state is Blue" is that the state is, indeed, Blue.

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*.*

Next, we distinguish between direct and evasive messages. A direct message states the value of the primary payoff relevant dimension 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 direct recommendation about the action the receiver should take.

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

An evasive message makes a statement about any dimension of the state that is not of primary payoff relevance when the sender has information about the dimension 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) since it refers to the past and not the present, and is probabilistic.

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* $\theta_1 \in \Theta_1, x \in X$.

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

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

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.

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

We further distinguish between direct and evasive lies. In line with Khalmetski et al. (2017), a lie is direct if it concerns a primary payoff relevant dimension. In the examples above, "the state is Blue" is a direct lie since colour is the primary dimension, and it is in fact Red.

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

A lie is evasive if it is about any other dimension of the sender's type that is not primary payoff relevant. 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. The implication of this will become clear later, when discussing the different psychological costs associated with different messages.

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

We follow Sobel (2020) and distinguish between lies and deceptions, using the latter to capture the interpretation of the messages by the receiver. Moreover, deception is defined relative to other available messages. 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 in question will lead the receiver further away from an accurate belief about the payoff relevant dimension. For instance, saying "I don't know the colour" is deceptive when one knows the colour 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). *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* $\mu\left(\theta_i|m_{\theta_{i,j}}\right) - Pr(\theta_i) > \mu\left(\theta_i|m_{\theta_{i,j'}}\right) - Pr(\theta_i)$.

In other words, messages are deceptive when they induce more inaccurate beliefs than would another available message. As in Sobel (2020), a belief $\mu\left(\cdot\,|m_{\theta_{i,j}}\right)$ is inaccurate if, given θ_i, $\mu\left(\theta_i|m_{\theta_{i,j}}\right) \in [0,1)$, that is, whenever the receiver believes that, given a message, the state

dimension is not 100% likely to take its true value. The farther from 1 this belief is, the more inaccurate it is.

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 naïve (R^N) (similar to e.g., Kartik, 2009).[12] 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 naïve 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 naïve 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 naïve 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 naïve 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, naïve 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 g), the naïve 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 naïve receivers in the population (and $1 - \eta$ that of sophisticated receivers).

[12] 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.

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.

$c_i(\theta, m, \mu)$ 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.

1.2. 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. All proofs are delegated to the Appendix.

Lemma 1. *If and only if $C > g - l$, S will choose $m^* = Blue$, i.e., tell the truth.*

The difference between g and l (i.e., $g - l$) 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 $C < g - l$, 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.*

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 (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.*

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 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.*

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 naïve 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.

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 g). 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* $m = x$.

1.2.1. 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 naïve receiver's beliefs about the payoff relevant state dimension, and only x_2 (PARTIAL) changes the naïve receiver's beliefs away from their prior and toward the belief the state is Red (as suggested by the message). Consequently, the naïve 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 naïve 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. Furthermore, neither the sophisticated nor the naïve receiver can learn whether the message was truthful after

observing the payoff realization when the sender chooses to evade ($m \in \{x_1, x_2, x_3\}$). However, when the sender lies directly ($m = Red$), the sophisticated receiver will correctly infer the message was deceptive. 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 is informed that $\theta_1 = Blue$, as the sender could have truthfully revealed this is the case. Lastly, the truthful message ($m = Blue$) has a communication cost equal to 0 and hence the lowest of all messages.

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:

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

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 ($g - C(m \in \{Red, x\})$) is greater than the expected benefit of sending the truthful message ($l - C(m = Blue)$). That is, it depends on the probability that $g - C(m = Red) > l - C(m = Blue)$ and that $g - C(m = x) > l - C(m = Blue)$. Since the communication cost of

being truthful is equal to 0 ($C(m = Blue) = 0$), these inequations can be rewritten as $C(m = Red) < \mathrm{g} - l$ and $C(m = x) < \mathrm{g} - l$. We assume that $C \sim U(0, n)$ and $0 \leq g - l \leq n$. Since $C(m = Red) > C(m = x)$, it follows that $\Pr(C(m = Red) < g - l) < \Pr(C(m = x) < g - l)$. 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 naïve receivers will understand this is the case. This holds also for evasions where it is now clear for both the sophisticated and the naïve 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 that the receiver will find out whether the sender sent a deceptive message increases, the rate of deception will decrease.

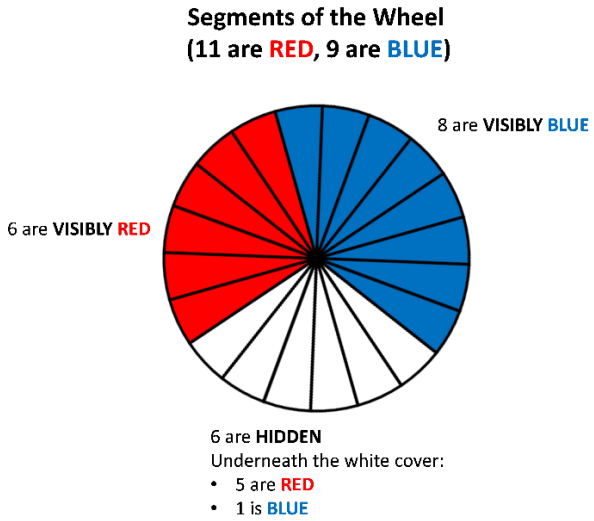
2. Experimental Design and Procedures

2.1. Experimental Design

To test our predictions, we conduct two experiments, the *Hidden Evasion* and *Open Evasion* experiments. Our empirical strategy mirrors the theoretical framework with both experiments involving a one-shot interaction between an informed sender and an uninformed receiver.

2.1.1. The Hidden Evasion experiment. Participants are allocated either the role of sender or receiver. The structure of the game is common knowledge. To determine the state of the world we use a setup as depicted in Figure 1. A wheel composed of 20 equal segments is spun, and one segment is randomly selected. The colour of this segment can be either Red or Blue, with Red being realized with probability 55%, and Blue with probability 45%.

Figure 1. The 20-segment wheel



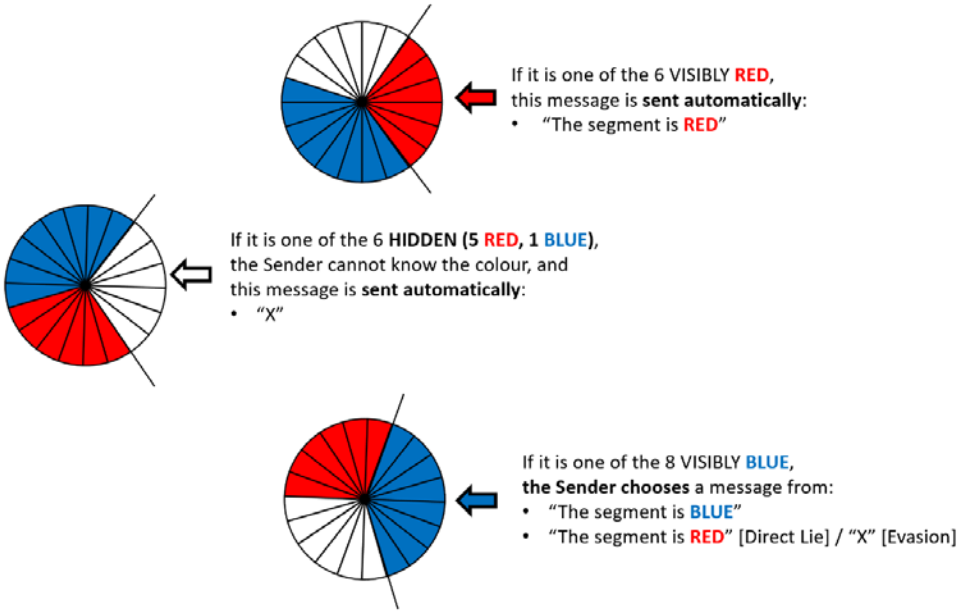
As shown in Figure 1, the segment colour can be either visible or hidden. With 70% probability a visible segment is selected so the sender is informed about its colour; with 30% probability a hidden segment is selected and the sender is uninformed. After the segment is selected, a costless message is sent to the receiver. The message is the only information the receiver obtains. After receiving the message, the receiver guesses whether the segment is Blue or Red. Subsequently, payoffs for both parties are realized, depending on the actual colour of the selected segment and the receiver's guess. Table 2 gives the payoffs for both players. There is a conflict of interest: the sender earns more if the receiver guesses Red, independently of the true state, whereas the receiver earns more, if his guess is correct, i.e., it matches the state.

Table 2. Payoff structure

		Receiver's guess	
		Red	**Blue**
State	**Red**	$\pi_S = £2; \pi_R = £2$	$\pi_S = £1; \pi_R = £1$
	Blue	$\pi_S = £2; \pi_R = £1$	$\pi_S = £1; \pi_R = £2$

To study the psychological cost of deception, we contrast two decision environments, one comprising a single treatment where participants can lie directly (DIRECT), and one with three evasion treatments. The structure of these decision environments is shown in Figure 2.

Figure 2. Summary of how messages are determined



In both direct lying and evasion environments, the sender chooses which message will be sent to the receiver only when the segment is visibly Blue, which is when the sender has a material incentive to deceive since she will benefit if the receiver believes the segment is Red.[13] In all treatments, when the segment is visibly Red, the automatic message "The segment is RED" is sent; if the randomly drawn segment is hidden, another automatic message is sent. We will call that message "X" and specify it in detail later, but for now think of message "X" as a placeholder for the different types of evasion we are investigating.

When there is a conflict of interest, meaning the segment is visibly Blue, the DIRECT and evasion treatments diverge. In DIRECT, the sender can either tell the truth or lie directly, whereas in the evasion treatments she can either tell the truth or evade. More specifically, in DIRECT, the sender can tell the truth with the message "The segment is BLUE" or lie directly with "The segment is RED." In the evasion treatments, the sender chooses whether to tell the truth with the message "The segment is BLUE" or evade with message "X."

The key to our design is that the receiver cannot ex-ante distinguish between truth and lies (in DIRECT) or truth and evasion (in the evasion treatments). In DIRECT, when the message "The segment is RED" is received it can be because it is sent automatically when the segment is visibly Red, or because the sender lied. In the evasion treatments, when the message

[13] We allowed the sender to choose the message only when they have an incentive to disguise the truth for two reasons. First, to attach natural meanings to messages, which is necessary for a literal interpretation of what constitutes a lie and second, to restrict the equilibrium strategies. Empirical evidence from a similar setting, shows the sender almost always (99.3% of the time) sends the truthful option when interests are aligned (Khalmetski et al., 2017).

"X" is received it could be because it is sent automatically when the segment is hidden, or because the sender chose to evade. In all treatments, therefore, deception is ex-ante credible.

Message "X" can take the form of the three evasive messages earlier introduced as *X*. We label the treatments: IGNORANCE, PARTIAL, and SILENCE. Table 3 depicts the exact content of the message for each evasion treatment. For instance, when the selected segment is visible and BLUE, a sender in IGNORANCE can choose between "The segment is BLUE" and "I don't know the colour of the segment." In the same treatment, if the sender is uninformed, the receiver automatically receives the message "I don't know the colour of the segment." The receiver cannot know whether the message "I don't know the colour of the segment" is sent automatically or chosen by the informed sender.

Table 3. Variations of message "X" per evasion treatment

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

By conducting pairwise comparisons between DIRECT and the three evasion treatments, we measure the psychological cost of evasion relative to the cost of direct lies. By comparing the deception rates across the three evasion treatments, we examine to what degree the language of evasion matters.

2.1.2. The Open Evasion experiment. An important feature of the Hidden Evasion experiment is that the receiver can infer if they were deceived only in DIRECT. If the receiver receives the message "The segment is RED" and follows the recommendation, he can infer he was deceived since his payoff will be £1 instead of the £2 he would receive if the segment was Red. However, evasion is ex-post non-verifiable, since the evasive message comes 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 recognised as a deceiver is highest in DIRECT.

To pin down the role of image concerns, the Open Evasion experiment controls for differences in the social image cost associated with different deceptive messages. In all treatments, including DIRECT, before senders decide which message to send to the receiver, they are informed that, after the receiver's guess, he will be told if the selected segment was visible or hidden, and if the message was chosen by the sender or sent automatically. Thus, in all treatments, it is now salient that there will be full revelation of the sender's type. Apart from this difference, the two experiments are identical.

2.1.3. Senders' beliefs. Senders' beliefs about how receivers interpret the messages they receive are important for identifying the psychological cost of deceptive communications. Senders, for instance, might believe receivers are more likely to choose Red following an evasive message rather than a direct lie, and therefore choose evasions more frequently. To examine whether any observed differences across treatments are driven by differences in the sender's expectations about the receiver's beliefs, and not by differences in the psychological cost of communication, we elicit those expectations. Each sender estimates the percentage of receivers who guess Red, after receiving the message that the segment is Blue (B(a=Red|m=Blue)) and the percentage who guess Red after receiving the alternative message which might be deceptive (B(a=Red|m=non-Blue)). The senders also estimate the percentage of other senders who choose the deceptive message (B(others-deceive)). Senders are paid £0.10 per question if their estimate is correct within 3 percentage points (in line with Abeler et al., 2019). Senders' beliefs are elicited after they have chosen their message.

2.1.4. Discussion of design choices. The specific distribution of segments on the 20-segment wheel was chosen for two reasons. First, it ensures the probability that the deceptive message is sent by a non-deceitful sender is equal across treatments: in 6 out of 14 cases, the Red message is non-deceptive as it is sent by a sender who indeed observed a Red segment, and the evasive message is non-deceptive as it is sent by a sender who observed a hidden segment. Direct lying and evasion are therefore equally credible. This is important, since previous research (Abeler et al., 2019) 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. Second, the distribution of segments ensures the expected benefit of evasion is not higher than the expected benefit of a direct lie: if people evade more often than directly lying, it is not because evasion is more profitable, but because it is less psychologically costly.

In all experimental treatments, we use the strategy method (Selten, 1967). Senders pre-define which message they want to send to the receiver conditional on the segment being visibly Blue. Similarly, receivers guess the segment's colour conditional on each message they may receive. Since we only analyse sender behaviour, we use a matching protocol of ten senders for each 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).

To determine the required sample size in each game, 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 senders 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.[14] We thus set our target sample to 300 senders in each evasion treatment and 100 in each DIRECT variation (to be matched with 30 and 10 receivers respectively). The design, hypotheses and 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 like our sender-receiver game. Participants studied a list of possible messages (truth telling, direct lying and various evasive statements including silence, partial truth and feigned ignorance) and then rated their deceptiveness in case of a conflict of interest, i.e., the randomly chosen segment was visibly Blue, 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.[15] In line with our hypotheses, telling a direct lie was perceived 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.

2.2 Experimental Procedures

We conducted the Hidden Evasion experiment in September 2019, and the Open Evasion experiment in November 2019. Both were implemented online using Prolific (http://www.prolific.ac) and programmed using Qualtrics (http://www.qualtrics.com/). The Humanities and Social Sciences Research Ethics Committee at the University of Warwick

[14] All 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.

[15] 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.

reviewed and approved the procedures (18/18-19). A total of 2,414 participants (average age = 36.5 years old; 64% female) completed the experiments in the role of sender, taking 12 minutes on average. Each participated in only one experimental treatment. We restricted our sample to UK residents that had 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 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 both 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 available in Appendix D.

3. Hypotheses

We describe the preregistered hypotheses that closely follow the preceding theoretical analysis. The first three hypotheses are restricted to the Hidden Evasion experiment and are based on Prediction 1 of our theoretical analysis according to which, the deceptive message is most psychologically costly in DIRECT, followed by IGNORANCE, PARTIAL and SILENCE.

Hypothesis 1: In the Hidden Evasion experiment, the proportion of senders choosing the deceptive option is lowest in DIRECT.

Hypothesis 2: In the Hidden Evasion experiment, the proportion of senders choosing to deceive is higher in PARTIAL and SILENCE compared to IGNORANCE.

Hypothesis 3: In the Hidden Evasion experiment, the proportion of senders choosing to deceive in SILENCE is higher than in PARTIAL.

We now turn to the effect of social image. There are two plausible hypotheses about the effect of social image depending on the relative costs of the different deceptive communications. If social image costs have no effect, any observed differences in the Hidden Evasion experiment should remain in the Open Evasion experiment. Otherwise, if any effect observed in the Hidden Evasion experiment 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 the Open Evasion experiment.

Hypothesis 4a: In the Open Evasion experiment, the proportion of senders choosing the deceptive option is lowest in DIRECT.

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

Lastly, in line with Prediction 2, we expect deception rates to be lower in each treatment of the Open Evasion experiment, where the receiver is explicitly informed about the sender's potential deception compared to the respective treatment of the Hidden Evasion experiment.

Hypothesis 5: The deception rate in the Open Evasion experiment is lower than in the Hidden Evasion experiment.

4. Results

In presenting the results of each game, we start with the treatment comparisons. We then explicitly discuss the role of beliefs. All hypothesis tests are two tailed, as pre-registered.

4.1. Hidden Evasion experiment

4.1.1. Sample characteristics

Our sample consists of 1,210 participants (65% female), randomly assigned to the four treatments. Their average age was 36.3 years old, with 87% having completed higher education (college or above). Table 4 depicts summary statistics for the sample demographics across treatments. The last row displays test statistics for the null hypothesis of perfect randomization. There is no evidence that the demographics are unbalanced across treatments but in any case, we control for these variables in our analysis.

Table 4. Sample characteristics and randomization check in Hidden Evasion

Treatment (N)	Age	Female	Higher education
DIRECT (305)	35.8 (0.63)	0.702 (0.03)	0.890 (0.02)
IGNORANCE (300)	37.3 (0.65)	0.623 (0.03)	0.886 (0.02)
PARTIAL (303)	36.0 (0.68)	0.637 (0.03)	0.852 (0.02)
SILENCE (302)	36.2 (0.70)	0.629 (0.03)	0.856 (0.02)
	H(3) = 3.31, p = 0.346	χ^2 (3, 1204) = 5.54, p = 0.136	χ^2 (3, 1191) = 3.06, p = 0.382

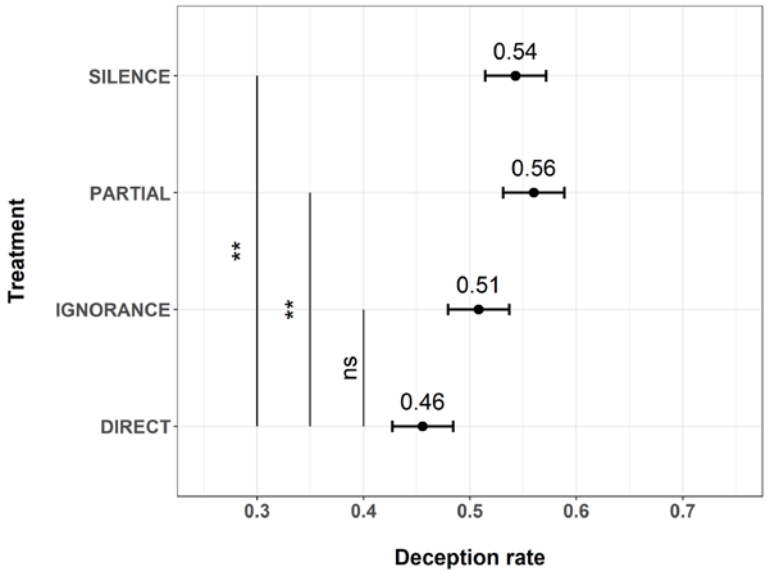
*Note*s. This table reports means and standard errors (in parenthesis) in each treatment of the Hidden Evasion experiment. The last row displays p-values for the null hypothesis of perfect randomization (Chi-square test in case of binary variables and Kruskal-Wallis test in case of interval variables).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

4.1.2. Senders' message choice

The average proportion of senders choosing the deceptive option (hereafter called deception rate) is close to 50%, similar to that typically observed in sender-receiver games (see meta-analytical estimates of 51% by Gerlach et al., 2019). Figure 3 presents the frequency with which

senders choose the deceptive message over the truth-telling one across the four treatments.[16] DIRECT leads to the lowest deception rate (and highest truth-telling). This is significantly lower than the frequency of deception in PARTIAL ($\chi^2(1, 605) = 6.167$, $p = 0.013$; $d = 0.21$) and SILENCE ($\chi^2(1, 607) = 4.284$, $p = 0.038$; $d = 0.17$) treatments, and not significantly lower than the rate in IGNORANCE ($\chi^2(1, 608) = 1.475$, $p = 0.225$; $d = 0.11$).

Figure 3. Deception rate across treatments in Hidden Evasion



*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 $p > 0.10$.

The Hidden Evasion treatments permit us to test Hypothesis 1, that deception will be less frequent via direct lying than via evasion. As Figure 3 shows, this hypothesis is supported only when comparing DIRECT with SILENCE and PARTIAL. This suggests the psychological cost of deception via direct lying is indeed higher than that of deception via silence or partial truth but we cannot confidently draw this conclusion for feigning ignorance.

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. We further control for senders' beliefs about receivers -

[16] Recall that in DIRECT we used three different versions for the automated 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 test for any effect on the decision to lie coming from the type of automated message associated with the uninformed sender. A Chi-square test comparing the deception rate across the three versions of DIRECT reveals no significant differences ($\chi^2(2, 305) = 2.405$, $p = 0.300$). For the rest of the analysis, in line with our pre-registration, we pool across the three versions of DIRECT and treat them as a unitary set of observations.

B(a=Red|m=Blue), B(a=Red|m=non-Blue) - and about other senders - B(others-deceive) (Table 5, column 1).

Table 5. Probit analysis of choosing the deceptive option in Hidden Evasion

	Dependent variable: Choice of deceptive option	
	(1)	(2)
IGNORANCE	0.044	0.044
	(0.045)	(0.045)
PARTIAL	0.148***	0.141***
	(0.043)	(0.044)
SILENCE	0.136***	0.132***
	(0.044)	(0.044)
B(a=Red|m=non-Blue)	0.001	0.001
	(0.001)	(0.001)
B(a=Red|m=Blue)	-0.001	-0.001
	(0.001)	(0.001)
B(others-deceive)	0.009***	0.009***
	(0.001)	(0.001)
Female		-0.078**
		(0.033)
Age		0.001
		(0.001)
Higher education		0.043
		(0.048)
Observations	1,210	1,182

Notes: Marginal effects from a probit regression in the Hidden Evasion experiment. 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. 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Consistent with the findings just reported, senders are 14.8 percentage points more likely to choose the deceptive option in PARTIAL compared to DIRECT ($p = 0.001$) and 13.6 percentage points more likely to do so in SILENCE compared to DIRECT ($p = 0.002$). The marginal effect of IGNORANCE is insignificant compared to DIRECT ($p = 0.327$). Beliefs about other senders' behaviour have a significant positive effect on the likelihood of choosing the deceptive option ($p < 0.001$). Controlling for gender, age, and education leaves the results

essentially unchanged (compare Table 5, column 2).[17] The fact that IGNORANCE is statistically indistinguishable from DIRECT suggests the falsehood cost is higher in IGNORANCE than the other evasion treatments.

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

We next compare how often the deceptive option is chosen across the three evasion treatments. We find no support for Hypotheses 2 and 3 (focusing on differences between the evasion treatments) as the proportion choosing the deceptive option does not significantly differ across any of these pairwise comparisons. The deception rate in IGNORANCE is not significantly different from PARTIAL ($\chi^2(1, 603) = 1.421$, $p = 0.233$) or SILENCE ($\chi^2(1, 605) = 0.602$, $p = 0.438$). Similarly, PARTIAL is statistically indistinguishable from SILENCE ($\chi^2(1, 602) = 0.113$, $p = 0.737$). We summarize these findings in Result 2 which suggests the precise language of evasion does not matter.

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

4.1.3. Senders' beliefs

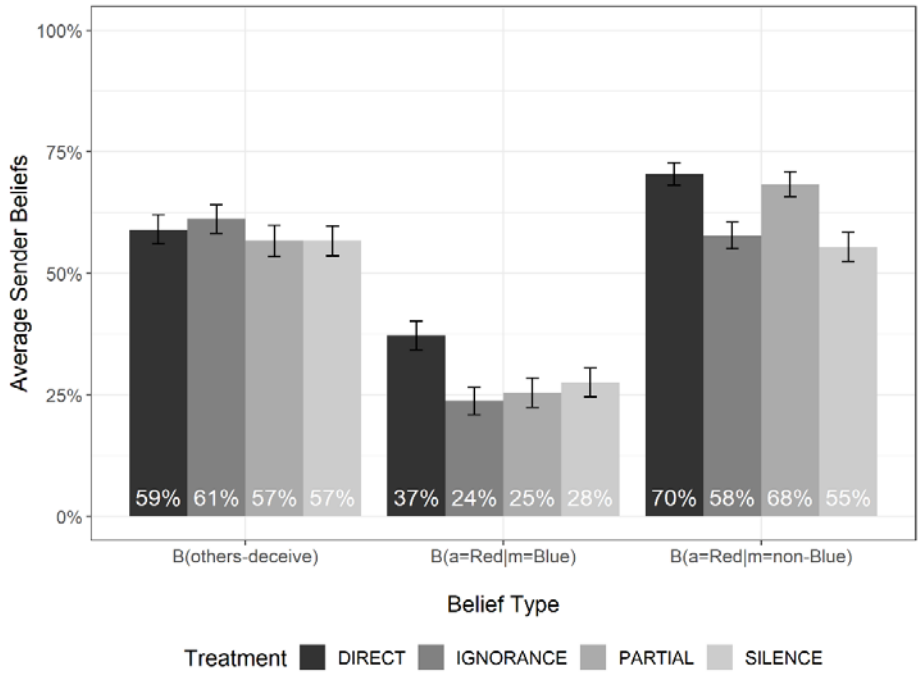
Figure 4 presents average sender beliefs across the four treatments, for all senders, irrespective of the sender's choice (deceptive or truthful) since this did not affect the distribution of beliefs (see Appendix B for the analysis of belief distributions across treatments and decisions).

We find no difference across treatments in the senders' beliefs about the likelihood that other senders would deceive ($H(3) = 5.471$, $p = 0.140$). However, the beliefs about how receivers would react to senders' messages differ. When comparing senders' beliefs about how likely receivers are to choose Red after the truthful message (m=Blue), we find a significant difference across treatments ($H(3) = 56.048$, $p < 0.001$). Specifically, senders believe that receivers are most likely to reward truthful senders in DIRECT (see Appendix B for the results of the pairwise comparisons). This, however, does not mean it is 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 (m=non-Blue) is significantly higher than after the Blue message (DIRECT: $t(304) = 16.619$, $p < 0.001$; IGNORANCE: $t(302) = 18.695$, $p < 0.001$; PARTIAL: $t(299) = 19.379$, $p < 0.001$; SILENCE: $t(301) = 13.566$,

[17] We find a significant effect of gender: females are 7.8 percentage points less likely to choose the deceptive option than men ($p = 0.017$). The effect is not systematic, since there is no evidence of a gender effect in the Open Evasion experiment. In any case, our study was not built to investigate gender effects.

$p < 0.001$). According to their beliefs, senders should always make the deceptive choice to maximise their earnings. As before, this is not what happens, suggesting that deception incurs psychological costs so that senders forego some earnings to avoid them.

Figure 4. Average sender beliefs across treatments in Hidden Evasion



*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).

We also find significant differences across treatments in senders' estimates of the likelihood receivers will choose Red after each deceptive message ($H(3) = 79.065$, $p < 0.001$). In particular, senders believe that receivers are significantly less likely to choose Red after the IGNORANCE and SILENCE message compared to the DIRECT and PARTIAL message (see Appendix B for the pairwise comparisons). This supports our assumption that the influence cost of deception is lower in IGNORANCE and SILENCE than in PARTIAL or DIRECT.

Result 3. *When evasion is non-verifiable, senders believe that receivers are more likely to choose the action implied by the message when the message is a direct lie or a partial truth than when keeping silent or feigning ignorance.*

4.2. Open Evasion experiment

4.2.1. Sample characteristics

For the Open Evasion experiment, our sample consists of 1,204 participants (63% female) randomly distributed across the four treatments. Their average age was 36.6 years, with 89% having completed higher education (college or above). Table 6 depicts summary statistics for

the sample demographics across the treatments. The last row displays test statistics for the null hypothesis of perfect randomization. As with the Hidden Evasion experiment, there is no evidence that the demographic characteristics of the participants are unbalanced across treatments; we also control for demographic variables in all regressions.

Table 6. Sample characteristics and randomization check in Open Evasion

Treatment (N)	Age	Female	Higher education
DIRECT (303)	35.7 (0.72)	0.653 (0.03)	0.919 (0.02)
IGNORANCE (305)	37.0 (0.72)	0.597 (0.03)	0.877 (0.02)
PARTIAL (297)	36.9 (0.74)	0.640 (0.03)	0.875 (0.02)
SILENCE (299)	36.6 (0.74)	0.645 (0.03)	0.884 (0.02)
	$H(3) = 2.58$, $p = 0.460$	$\chi^2(3, 1198) = 2.79$, $p = 0.425$	$\chi^2(3, 1179) = 3.80$, $p = 0.284$

*Note*s. This table reports means and standard errors (in parenthesis) in each treatment of the Open Evasion 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).

4.2.2. Senders' message choice

Figure 5 depicts the deception rate in each treatment.[18] As in the Hidden Evasion experiment, the lowest deception rate is in DIRECT. The pattern in the comparisons between DIRECT and the evasion treatments is also similar to the Hidden Evasion experiment with one exception: the deception rate in SILENCE is no longer significantly higher compared to DIRECT ($\chi^2(1, 602) = 0.641$, $p = 0.423$; $d = 0.07$). The deception rate in PARTIAL is significantly higher than in DIRECT ($\chi^2(1, 600) = 4.113$, $p = 0.043$; $d = 0.17$), while the deception rate in IGNORANCE remains statistically indistinguishable from DIRECT ($\chi^2(1, 608) = 0.828$, $p = 0.363$; $d = 0.08$).

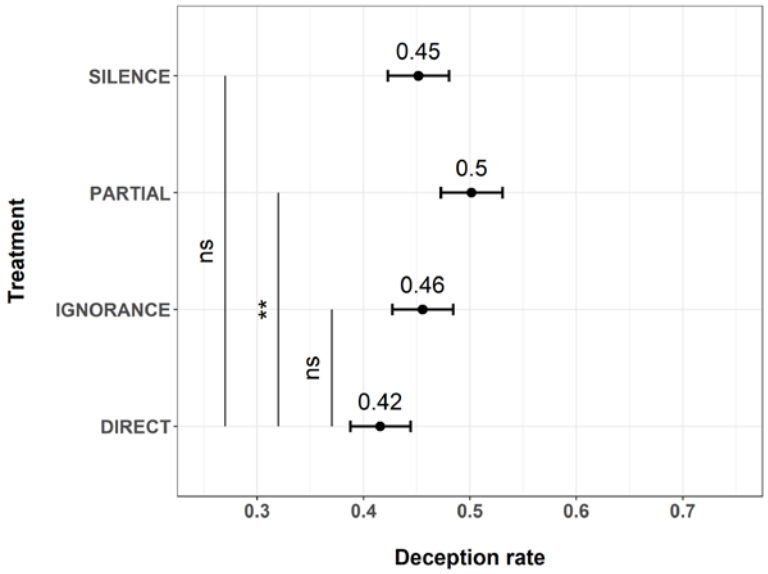
As in the Hidden Evasion experiment, we conduct a probit analysis predicting choice of the deceptive option from experimental treatment and control variables. The results reported in Table 7 corroborate our main findings. After controlling for beliefs (Table 7, column 1), senders are 10.9 percentage points more likely to choose the deceptive option in PARTIAL compared to DIRECT and this difference is statistically significant ($p = 0.018$). The marginal effect of IGNORANCE ($p = 0.903$) and SILENCE ($p = 0.575$) is insignificant. As in the Hidden

[18] As in the Hidden Evasion experiment, in DIRECT we used three different versions for the automated message coming from the uninformed sender (the versions used in the three evasion treatments). Before analysing this treatment as one, we test for any effect on the decision to lie coming from the specific automated message associated with the uninformed sender. A Chi-square test comparing the deception rate across the three versions of DIRECT suggests no significant differences ($\chi^2(2, 303) = 1.870$, $p = 0.393$). We therefore pool across the three versions of this treatment and analyse them as a unitary set of observations for our main hypothesis testing.

Evasion experiment, beliefs about how likely others are to deceive have a significant positive effect on choice of the deceptive option ($p < 0.001$). We also find a negative effect of beliefs about how likely receivers are to choose Red after receiving the message that the segment is Blue ($p = 0.062$). Controlling for demographics does not change anything (Table 7, column 2).

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

Figure 5. Deception rate across treatments in Open Evasion



*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 $p > 0.10$.

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 have meant that the differences observed in the Hidden Evasion experiment (Result 1) were only due to differences in social image costs. Result 4 instead suggests that the psychological cost of deception is lower for partial truth than for lying even after controlling for social image concerns. We therefore find partial support for Hypothesis 4a.

Next, we conduct pairwise comparisons of the evasion treatments to investigate whether the language of evasion matters when evasion is verifiable. It does not. The deception rate in IGNORANCE is statistically indistinguishable from that in PARTIAL ($\chi^2(1, 602) = 1.096$, $p = 0.295$) and SILENCE ($\chi^2(1, 604) = 0.001$, $p = 0.982$), while deception rates in PARTIAL and SILENCE do not differ either ($\chi^2(1, 596) = 1.310$, $p = 0.253$), a result in line with the findings in the Hidden Evasion experiment (Result 2). This suggests that the sender

incurs similar costs by staying silent or declaring ignorance if they were to be found out, or that any differences are not large enough to be captured by our setting.

Result 5. *When evasion is verifiable, the deception rate does not significantly differ across the three evasion treatments.*

Table 7. Probit analysis of choosing the deceptive option in Open Evasion

	Dependent variable: Choice of deceptive option	
	(1)	(2)
IGNORANCE	-0.006	-0.006
	(0.046)	(0.047)
PARTIAL	0.109**	0.113**
	(0.046)	(0.047)
SILENCE	0.026	0.022
	(0.047)	(0.048)
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		0.015
		(0.034)
Age		0.001
		(0.001)
Higher education		-0.002
		(0.052)
Observations	1,204	1,172

Notes: Marginal effects from a probit regression in the Open Evasion experiment. 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. 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Next, we test Hypothesis 5 by comparing the average deception rate in each treatment of the Open Evasion experiment, with that in the corresponding treatment of the Hidden Evasion experiment. The deception rates across all treatments are lower in Open Evasion than in Hidden Evasion, although only for SILENCE is the pairwise comparison statistically significant – 45% vs 54% ($\chi^2(1, 601) = 4.677$, $p = 0.031$; $d = 0.18$). 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).[19] An overall analysis shows that the deception rate significantly decreases with 8.5 percentage points in Open Evasion compared to Hidden Evasion ($p < 0.001$, see Table B13 in Appendix B for the full regression table).[20]

Result 6. *When evasion is verifiable (Open Evasion), the deception rate is lower than when it is not (Hidden Evasion).*

4.2.3. Senders' beliefs

Figure 6 presents the average sender beliefs across the four treatments in the Open Evasion experiment. The elicited beliefs were the same as in the Hidden Evasion experiment. Whether the sender chose the deceptive or the truthful message did not significantly affect the distribution of beliefs so, here, we report average beliefs irrespective of senders' decision (see Appendix B for the analysis of belief distributions across treatments and decisions).

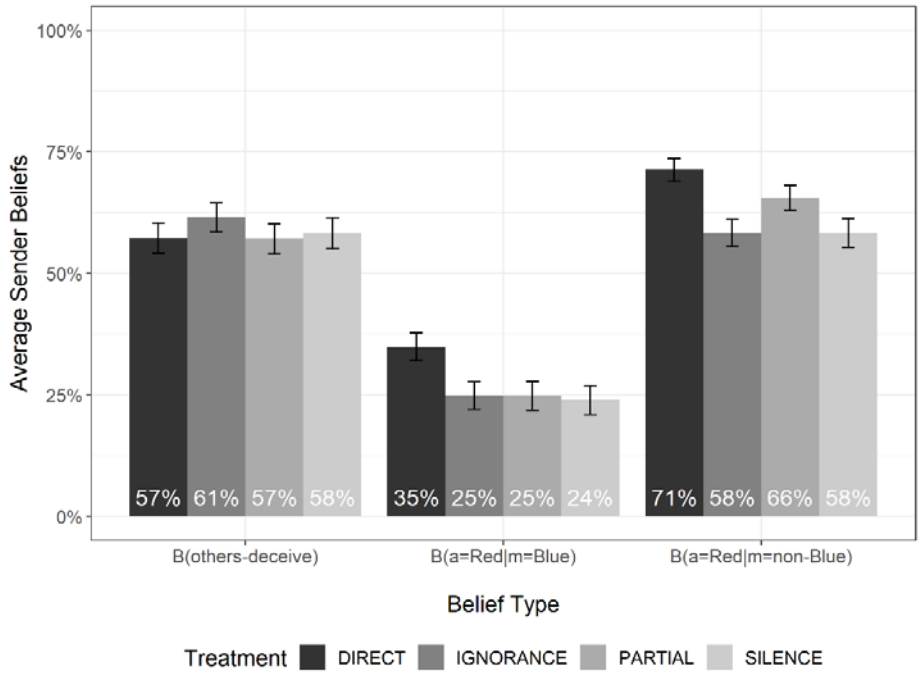
Figure 6 depicts a belief distribution very similar to that in the Hidden Evasion experiment. We find no difference across treatments in the average sender's belief about the likelihood that other senders would engage in deception B(others-deceive) ($H(3) = 4.802$, $p = 0.187$). However, the average belief about receiver's likelihood to choose the action Red after observing the Blue, truthful message (B(a=Red|m=Blue)) is significantly higher in DIRECT compared to the evasion treatments ($H(3) = 46.581$, $p < 0.001$). Senders believe that the deceptive option is more profitable than the truthful one also when evasion can be verified, since the average belief about the likelihood that the receiver will choose the Red action after the deceptive message (non-Blue) is always significantly higher than after the Blue message (DIRECT: $t(302) = 18.532$, $p < 0.001$; IGNORANCE: $t(304) = 17.805$, $p < 0.001$; PARTIAL: $t(296) = 18.892$, $p < 0.001$; SILENCE: $t(298) = 17.744$, $p < 0.001$). Finally, in what the average sender's beliefs about the likelihood the receivers will choose Red after the deceptive message is concerned, we find again significant differences across treatments ($H(3) = 60.308$, $p < 0.001$). This further strengthens the robustness of Result 3.

[19] Result 6 should be interpreted with some caution because the Hidden Evasion and Open Evasion experiments were not conducted simultaneously. Hidden Evasion 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 the Open Evasion experiment, 7 weeks after, to isolate the role of social image (since this only made sense if differences were observed in the Hidden Evasion experiment). 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: $t(2373) = 0.513$, $p = 0.608$; Female: $\chi^2(1, 2402) = 0.473$, $p = 0.492$; Higher education: $\chi^2(1, 2370) = 1.686$, $p = 0.194$).

[20] Note the overall analysis was not pre-registered.

Result 7. *When evasion is verifiable, senders believe that receivers are more likely to choose the action implied by the message when the message is direct or a partial truth than when keeping silent or feigning ignorance.*

Figure 6. Average sender beliefs across treatments in Open Evasion



Notes. 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).

5. Conclusion

We study the use of a variety of deceptive communications in the context of a sender-receiver game, where an informed party can benefit from deceiving an uninformed counterpart. We compare the rate at which a deceptive option is chosen in four treatments: the benchmark case of direct lies, and three evasions including feigned ignorance, partial truth and silence. By design the only reason for some deceptive communications to be chosen more than others is variation in their psychological costs. Possible determinants of those costs include the cost of intentionally leading the other party to a false belief (deception), of uttering a false statement or lie (falsehood), of changing the beliefs of the other party in a manner against their interest (influence), and of being perceived as a deceiver (social image).

Consistent with previous results, we find that senders do not always choose the deceptive option, either via direct lies or via evasion. Nevertheless, the deceptive option is more frequently chosen when it takes the form of evasion rather than direct lying. This suggests evasion has a lower psychological cost. Indeed, even after eliminating the possibility of

plausible deniability from evasion, some types of evasion are still chosen more frequently than direct lies, indicating that the preference for evasion is not only due to social image costs, but is also driven by intrinsic costs.

A possibility we considered is that senders expect receivers to react more gullibly to evasive messages. We rule this out with additional evidence from incentivized beliefs showing there is no indication that senders believe the receiver will be more likely to choose the option best for the sender under evasion than under a direct lie. This gives further support to our proposal that a direct lie incurs a greater psychological cost than evasion.

Our findings suggest that deception might be more widespread than suggested by previous estimates which relied mostly on paradigms where only direct lies are allowed, as in most of the lying literature documented in the meta-analyses of Abeler et al. (2019) and Gerlach et al. (2019). Consequently, the focus on outright lies might be too narrow, since many people might refrain from direct lies, yet engage in evasion due to its lower psychological cost.

For organizations and policy makers, our results suggest 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 have an incentive to 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 mandate on statements that contain instrumental information is important to reduce such deceptive communications.

Importantly, 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 individuals are unlikely to be held accountable, as in the case of evasion. Thus, enforcing deterrence policies that rely on reputation, might not be helpful, and could even backfire. As has been recently shown by Tergiman and Villeval (2021) in a market setting between informed managers and uninformed

investors where only some types of direct lies can be detected, increasing reputation does not make managers lie less, but switch from detectable to deniable lies. This is consistent with our finding that senders deceived more in the Hidden Evasion experiment, where evasion was not detectable, than in the Open Evasion experiment, where it was.

Of course, our study is only 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 factors relate to what can be called the "menu" of deception. 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 along with truth-telling and direct lying. Different contexts will render different evasions more or less beneficial, partly (but not entirely) due to their being more or less credible and detectable. "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"[21] is best used when an event has occurred long ago. The differing psychological costs associated with each item on the wide menu of deception 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).

[21] See for instance https://www.politico.com/story/2017/06/25/washington-defense-trump-russia-239914 for instances of "faulty" memories.

These menu effects show how in most everyday interactions, communication is richer than our pre-constructed restricted messages. People are creative in getting away with deception, while maintaining a favourable self and social image. Similarly, participants in our setting had to actively indicate their preference for silence, while, in reality, this is often a passive act. However, if anything, what all the above considerations imply is that evasion might be even less psychologically costly and therefore more prominent in more naturalistic environments. We leave these possibilities open for future research.

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.

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.

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

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

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

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.

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.

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.

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.

Gillespie, E., (2008). "Stemming the Tide of Greenwash: How an Ostensibly 'Greener' Market Could Pose Challenges for Environmentally Sustainable Consumerism," *Consumer Policy Review, 18(*3), 79.

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.

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.

Krawczyk, M. (2019). What should be regarded as deception in experimental economics? Evidence from a survey of researchers and subjects. *Journal of Behavioral and Experimental Economics, 79*, 110-118.

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.

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.

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.

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. (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.

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.

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

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.

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

Appendix A. Proofs

Lemma 1

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 $g - C$. This is not a profitable deviation when $C > g - l$. Hence, if this treatment is met $C > g - l$, it is optimal for the sender to truthfully reveal the state (i.e., to use $m^* = Blue$ as their equilibrium strategy).

Corollary 1

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 naïve 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:

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

Since $g > l$, 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.

Proposition 1

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

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

$$\begin{cases} \mu_{R^N}(\theta_1 = Red \mid m = Red) = 1; \\ \mu_{R^N}(\theta_1 = Red \mid m = Blue) = 0; \\ \mu_{R^N}(\theta_1 = Red \mid m = x) = p; \end{cases}$$

Following these beliefs, the naïve receiver's optimal action when $m = Blue$ is $a_{R^N} = Blue$, while when $m = Red$, the naïve receiver's optimal action is $a = Red$; but, when $m = x$, the naïve 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

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

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

Following these beliefs, the naïve receiver's optimal action when $m = Blue$ is $a_{R^N} = Blue$, while when $m = Red$, the naïve receiver's optimal action is $a_{R^N} = Red$; when $m = x$, the naïve 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 naïve 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.

Appendix B. Additional Analyses

Are average sender beliefs different across treatments?

In the following tables we present the results of multiple comparison tests (Tukey HSD) for differences in mean senders' 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).

Hidden Evasion experiment

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

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 choosing Red after the deceptive message (non-Blue) in Hidden Evasion

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 Hidden Evasion

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

Open Evasion experiment

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

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 choosing Red after the deceptive message (non-Blue) in Open Evasion

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	4.96	1.00
SILENCE – PARTIAL	-7.2	-12.3	-2.23	0.00

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

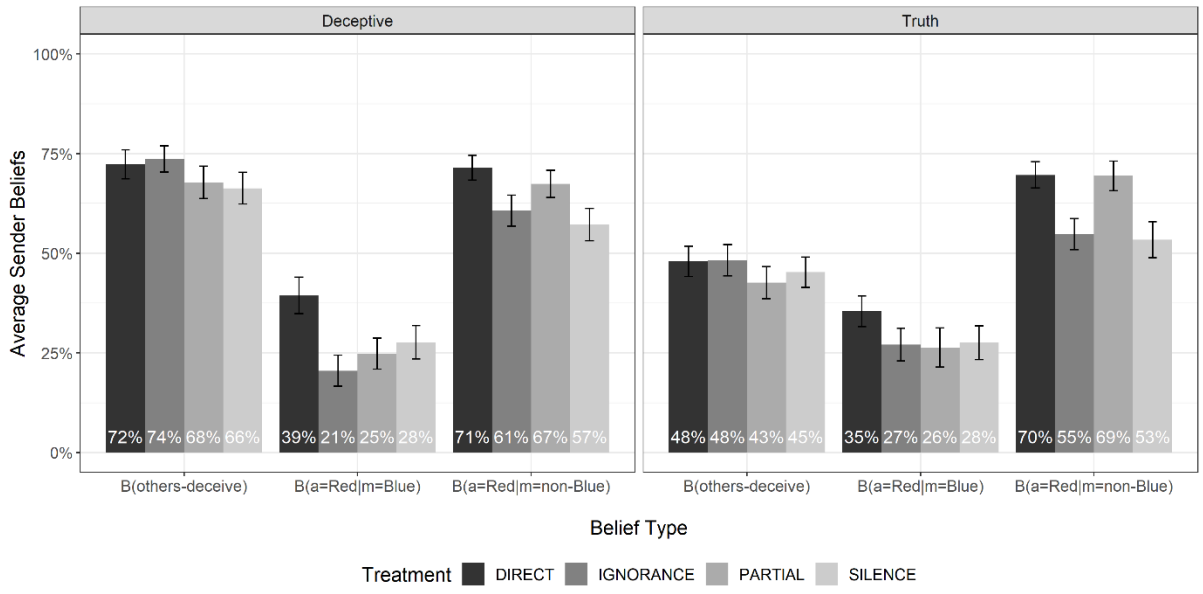
Treatments Compared	Mean Difference	95% Confidence Interval		Adjusted p-value
		Lower Bound	Upper Bound	
IGNORANCE – DIRECT	4.2	-1.4	9.8	0.22
PARTIAL – DIRECT	-0.2	-5.8	5.5	1.00
SILENCE – DIRECT	1.0	-4.7	6.6	0.97
PARTIAL – IGNORANCE	-4.4	-10.0	1.3	0.20
SILENCE – IGNORANCE	-3.2	-8.9	2.4	0.46
SILENCE – PARTIAL	1.1	-4.6	6.8	0.96

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

Hidden Evasion 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 Hidden Evasion



*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 B7. ANOVA results of senders' beliefs about the likelihood of receivers choosing Red after the truthful message (Blue) in Hidden Evasion across treatments and decision

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 B8. ANOVA results of senders' beliefs about the likelihood of receivers choosing Red after the deceptive message (non-Blue) in Hidden Evasion across treatments and decision

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

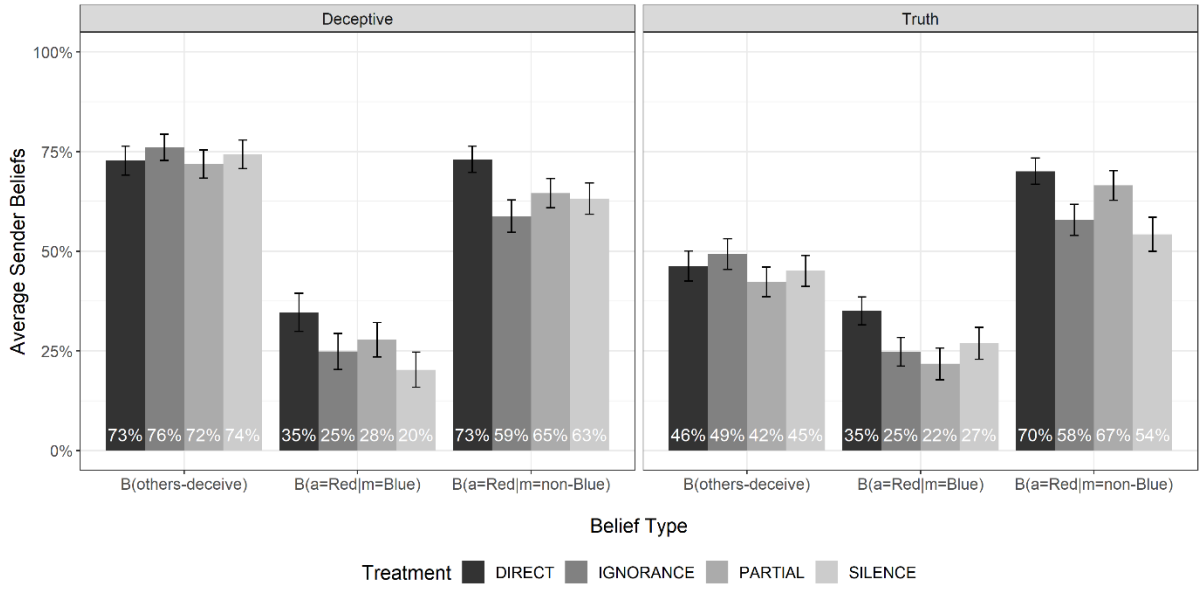
Table B9. ANOVA results of senders' beliefs about the likelihood of other senders choosing the deceptive message(non-Blue) in Hidden Evasion across treatments and decision

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	3	50897	16966	30.7	0.00
Decision	1	1714	1714	3.1	0.08
Treatment x Decision	3	2517	839	1.5	0.21
Residuals	1202	664350	553	NA	NA

Open Evasion 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 Open Evasion



*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 B10. ANOVA results of senders' beliefs about the likelihood of receivers choosing Red after the truthful message (Blue) in Open Evasion across treatments and decision

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 B11. ANOVA results of senders' beliefs about the likelihood of receivers choosing Red after the deceptive message (non-Blue) in Open Evasion across treatments and decision

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 B12. ANOVA results of senders' beliefs about the likelihood of other senders choosing the deceptive message (non-Blue) in Open Evasion across treatments and decision

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

Are deception rates in the evasion treatments lower in the Open Evasion experiment compared to the Hidden Evasion one?

Table B13. Probit analysis of deception rates across Open and Hidden Evasion

	Dependent variable:				
	Choice of deceptive option				
	(Overall)	(DIRECT)	(IGNORANCE)	(PARTIAL)	(SILENCE)
Open Evasion	-0.850***	-0.045	-0.071	-0.080*	-0.147***
	(0.022)	(0.044)	(0.045)	(0.045)	(0.045)
B(a=Red|m=non-Blue)	0.001	-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)	0.011***	0.011***	0.012***	0.011***	0.010***
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.035	-0.086*	-0.051	-0.010	0.025
	(0.024)	(0.047)	(0.047)	(0.047)	(0.047)
Age	0.001	0.000	-0.001	0.004*	0.001
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Higher education	0.020	-0.035	-0.111	0.127*	0.087
	(0.035)	(0.076)	(0.068)	(0.070)	(0.068)
Treatment FE	Yes				
Observations	2,354	596	593	581	584

Notes: This table reports marginal effects from probit regressions for each treatment. The dependent variable is whether the chosen message is deceptive (1 if yes, 0 if not). Open Evasion is a dummy for the Open Evasion 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

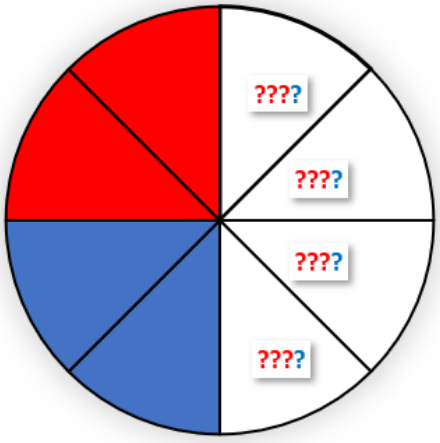
Appendix C. Additional Survey

C1. The Survey

Survey Procedures. The survey was conducted online in November 2018 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

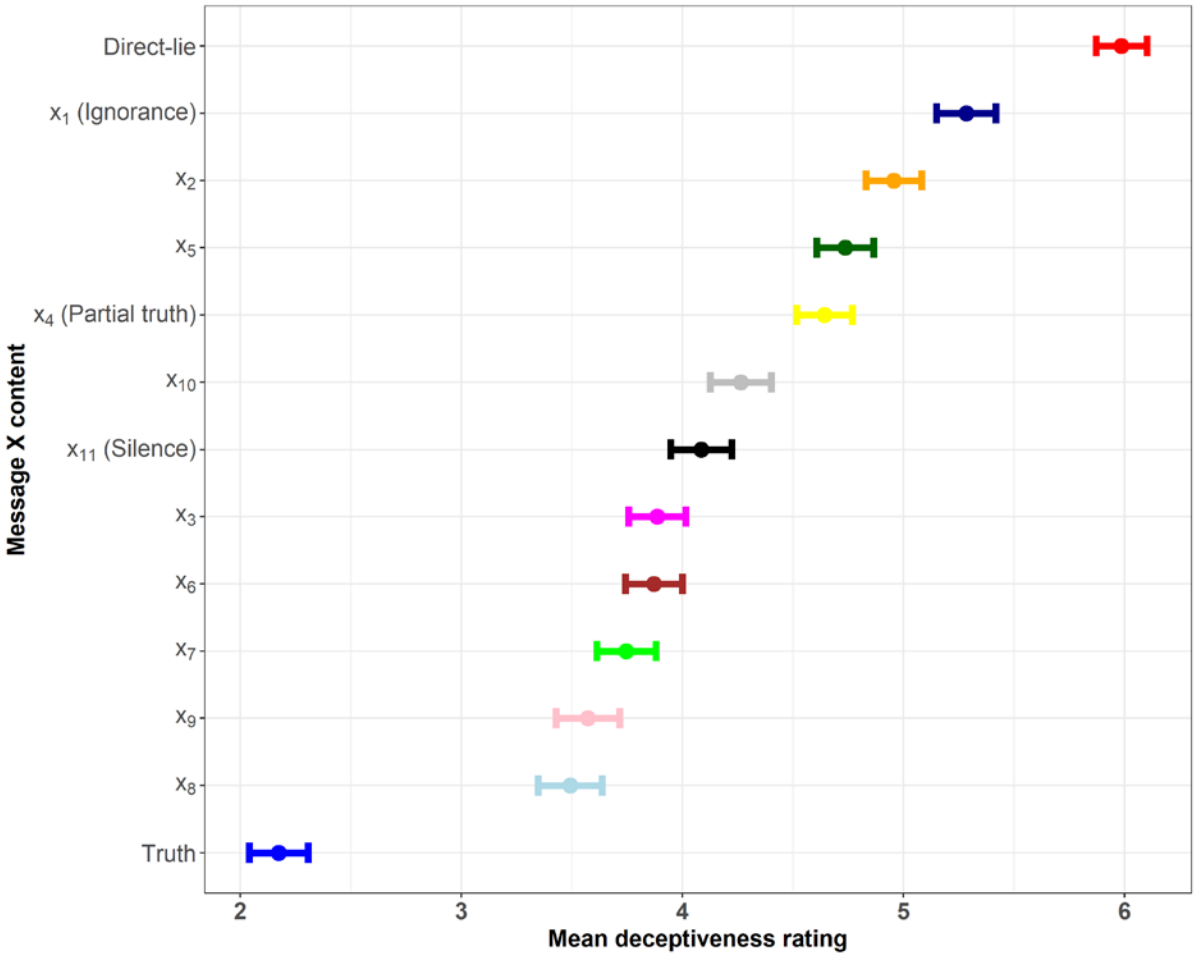


Table C1. T-test results for the comparison of all evasive messages with truth-telling and direct lying

Evasive message	Comparison with truth	Comparison with direct lie
"I do not know the colour of the segment"	$t(200) = -14.804$, $p < 0.001$	$t(200) = 4.683$, $p < 0.001$
"A hidden colour segment was chosen"	$t(200) = -13.619$, $p < 0.001$	$t(200) = 7.108$, $p < 0.001$
"The segment is either RED or BLUE"	$t(200) = -10.122$, $p < 0.001$	$t(200) = 11.925$, $p < 0.001$
"The segment was more likely to be RED than BLUE"	$t(200) = -13.618$, $p < 0.001$	$t(200) = 8.469$, $p < 0.001$
"The segment is more likely to be RED than BLUE"	$t(200) = -13.162$, $p < 0.001$	$t(200) = 7.975$, $p < 0.001$
"There are more RED segments"	$t(200) = -10.153$, $p < 0.001$	$t(200) = 11.703$, $p < 0.001$
"There are both visible and hidden colour segments"	$t(200) = -8.969$, $p < 0.001$	$t(200) = 12.003$, $p < 0.001$
"The current year is 2018"	$t(200) = -6.962$, $p < 0.001$	$t(200) = 12.963$, $p < 0.001$
"Today is Friday"	$t(200) = -7.606$, $p < 0.001$	$t(200) = 12.954$, $p < 0.001$
"Today is Tuesday"	$t(200) = -11.132$, $p < 0.001$	$t(200) = 9.455$, $p < 0.001$
"" (Keep silent: a blank message containing no information)	$t(200) = -10.582$, $p < 0.001$	$t(200) = 10.257$, $p < 0.001$

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

Comparison	
PARTIAL – IGNORANCE	$t(200) = -3.444$, $p < 0.001$
SILENCE – IGNORANCE	$t(200) = -6.783$, $p < 0.001$
SILENCE – PARTIAL	$t(200) = -3.237$, $p = 0.001$

C2. Experimental Instructions for the Survey

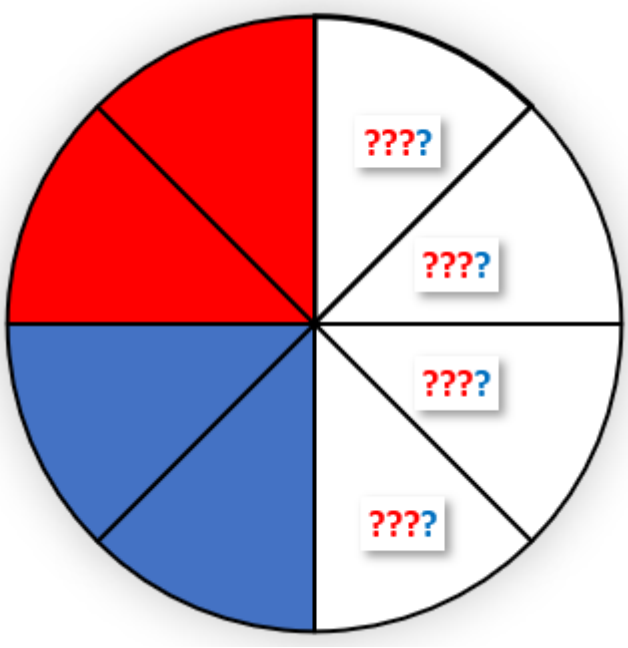
Below are the instructions for the survey. We provide the instructions from the perspective of Person A (the sender), and we use brackets ({}) to indicate the changes from the perspective of Person B (the receiver).

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.



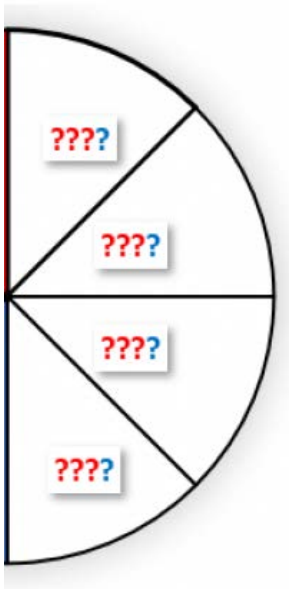
Four segments on the spinner are visibly coloured, two **RED** and two **BLUE**. The other four are hidden colour segments that have a white flap concealing the colour. Under the flap, three of these hidden colour segments are **RED** and one is **BLUE**. There is no way to know what colour is the segment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

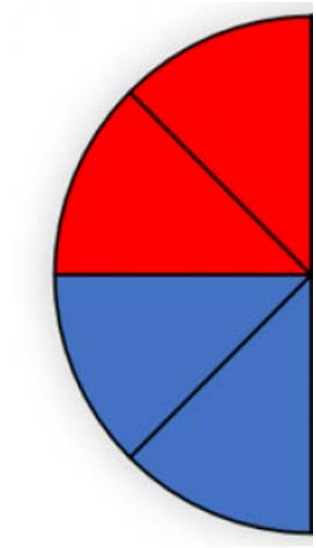
----------(page break) ----------



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?

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

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

- o Yes
- o No
- o 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?

- o Yes
- o No
- o 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?

- o Yes
- o No
- o 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?

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

Q2. What is your age?

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

Q3. What is your marital status?

- o Single, never married
- o Married or domestic partnership
- o Divorced
- o Widowed
- o Separated
- o 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.

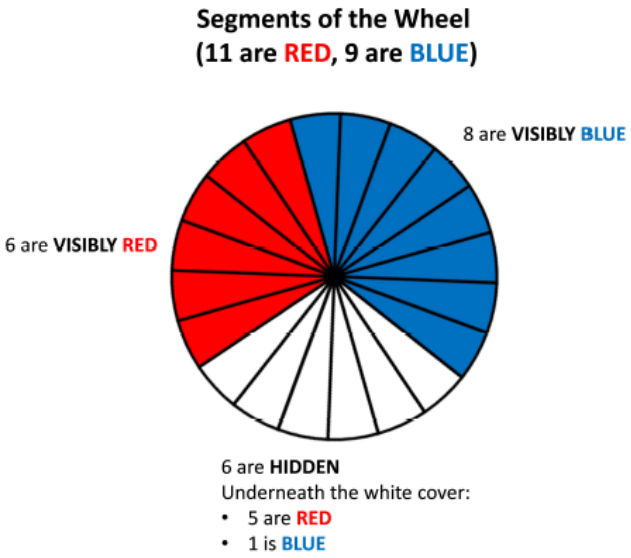
What follows is a description for **both** roles. You will learn your role after you have studied this description.

----------(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.



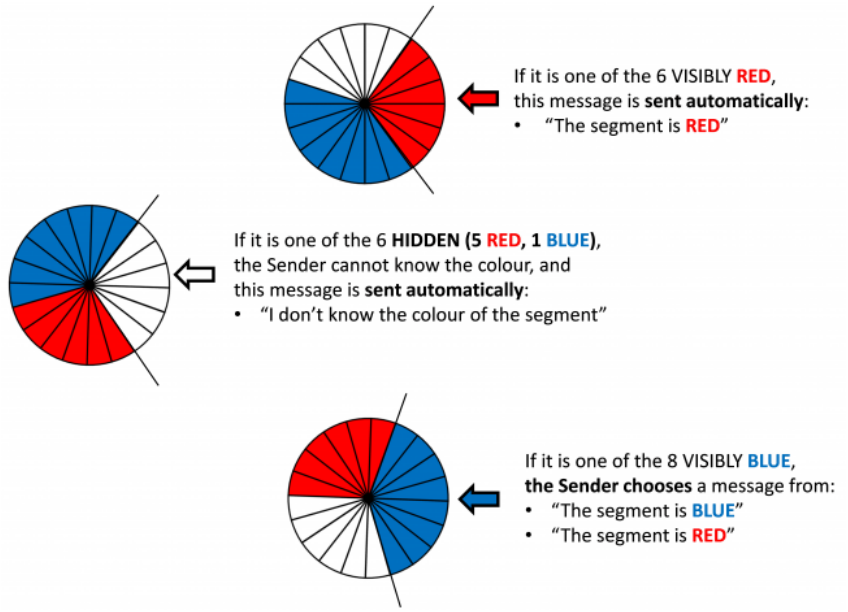
The Sender will observe the spin and its outcome, but the Receiver will not. Note that if a hidden colour segment is selected, the Sender cannot know whether the segment is **RED** or **BLUE**.

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:

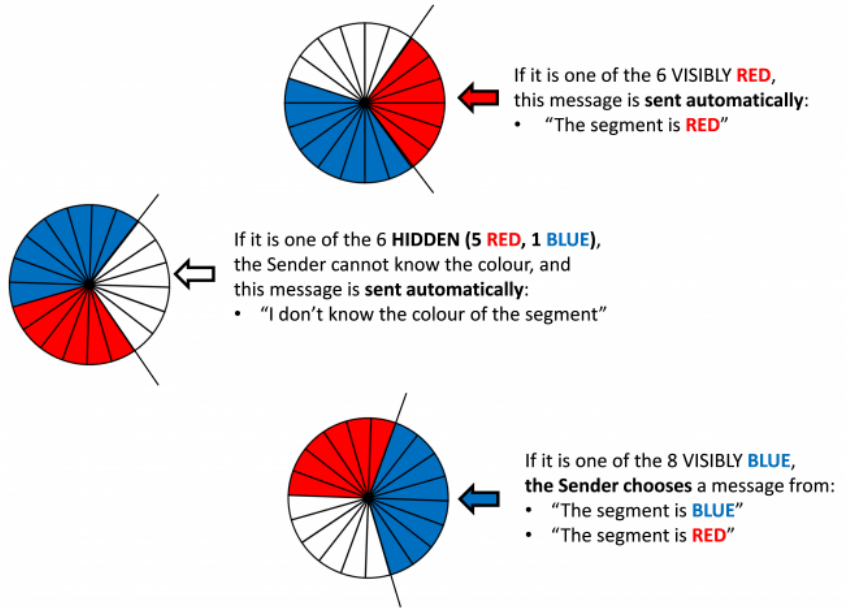
	Sender's bonus	**Receiver's bonus**
The segment is **RED** and the Receiver guesses **RED**	£2	£2
The segment is **RED** and the Receiver guesses **BLUE**	£1	£1
The segment is **BLUE** and the Receiver guesses **RED**	£2	£1
The segment is **BLUE** and the Receiver guesses **BLUE**	£1	£2

----------(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

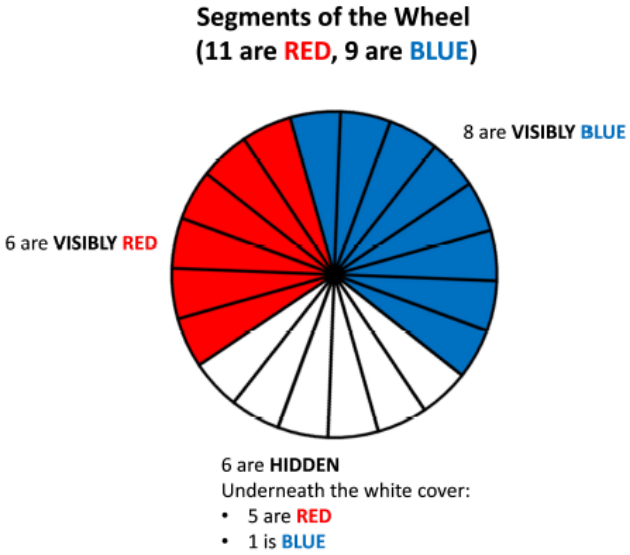
Your role in this study is that of: **Sender.**

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:

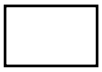
- "The segment is **BLUE**"
- "The segment is **RED**"

----------(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**."



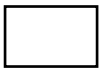
----------(page break) ----------

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**."



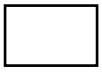
----------(page break) ----------

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.