It’s Not A Lie If You Believe It: Lying and Belief Distortion Under Norm-Uncertainty

Cristina Bicchieri\textsuperscript{a}, Eugen Dimant\textsuperscript{a,b,\ast\ast}, Silvia Sonderegger\textsuperscript{c}

\textsuperscript{a}University of Pennsylvania and Behavioral Ethics Lab
\textsuperscript{b}Identity and Conflict Lab
\textsuperscript{c}University of Nottingham and CeDEx

Abstract

This paper focuses on norm-following considerations as motivating behavior when lying opportunities are present. To obtain evidence on what makes it harder/easier to lie, we hypothesize that subjects might use belief-manipulation in order to justify their lying. We employ a two-stage variant of a cheating paradigm, in which subjects’ beliefs are elicited in stage 1 before performing the die task in stage 2. In stage 1: a) we elicit the subjects’ beliefs about majoritarian (i) behavior or (ii) normative beliefs in a previous session, and b) we vary whether participants are (i) aware or (ii) unaware of the upcoming opportunity to lie. We show that belief manipulation happens, and takes the form of people convincing themselves that lying behavior is widespread. In contrast with beliefs about the behavior of others, we find that beliefs about their normative convictions are not distorted, since believing that the majority disapproves of lying does not inhibit own lying. These findings are consistent with a model where agents are motivated by norm-following concerns, and honest behavior is a strong indicator of disapproval of lying but disapproval of lying is not a strong indicator of honest behavior. We provide evidence that supports this hypothesis.

Keywords: Cheating, Experiment, Lying, Social Norms, Uncertainty

JEL: C72, C91, D8, D9

\ast\ast This work benefited from discussions with Pierpaolo Battigalli, Roland Bénabou, Valentina Bosetti, Alexander Cappelen, Gary Charness, Christine Exley, Tobias Gesche, Sandy Goldberg, Peter Graham, David Henderson, Agne Kajackaite, Michel Maréchal, Gloria Origgi, Guido Tabellini, Bertil Tungodden, Joël van der Weele, and Roberto Weber. We also thank participants at the 2017 North American Economic Science Association Meeting, the Conference on Epistemic Norms as Social Norms, the 2018 Social Epistemology Network Event (SENE), and the 2018 Norms and Behavioral Change Workshop for helpful feedback, as well as the input received from seminar attendees and conversations at Bocconi University, UCSD, the University of Arizona, and University of Zurich.

\ast\ast Corresponding author

Email addresses: cb36@sas.upenn.edu (Cristina Bicchieri), edimant@sas.upenn.edu (Eugen Dimant), Silvia.Sonderegger@nottingham.ac.uk (Silvia Sonderegger)

This version: January 30, 2019
1. Introduction

The perennial tension between honest and dishonest behavior has been the focus of much attention in the experimental literature, and recent works (such as Dufwenberg Jr and Dufwenberg Sr, 2016; Khalmetski and Sliwka, 2017; Abeler et al., 2018; Feess and Kerzenmacher, 2018; Gneezy et al., 2018a) have emphasized the role of image concerns (i.e., the desire to appear honest) as a driver for honest behavior. This paper takes a complementary approach and focuses on norm-following considerations. Our definition of norm is taken from Bicchieri (2006): a norm of honesty exists if enough people are believed to abstain from lying and to disapprove of it. Existing research has shown that social norms can be powerful motivators in environments in which they are clearly defined and communicated to the individuals. (see Cialdini et al., 1990; Bicchieri, 2006, 2016; Fehr and Schurtenberger, 2018). In real life, however, there may be uncertainty as to which norms apply to (or are followed in) a specific situation, or even uncertainty as to the interpretation of a specific norm (e.g. the many interpretations of fairness in Babcock et al., 1995). When this is the case, people form beliefs about the relevant norm by analogy with previous experiences, or by drawing upon shared cultural narratives. This process of belief formation is susceptible to self-serving manipulation. A number of recent empirical studies have shown that, in ambiguous situations, people often choose to entertain beliefs that justify evading costly pro-social behavior (Dana et al., 2007; Di Tella et al., 2015; Exley, 2015; Gneezy et al., 2018b). These are what the literature calls “motivated beliefs” (Bénabou, 2015; Gino et al., 2016).

This paper draws on the notion that people may entertain self-serving beliefs to justify lying. An advantage of relying on belief manipulation rather than on other methods, such as providing subjects with information about peer behavior in previous sessions, is that it allows the study of the effect of large belief variations, something that might not be possible otherwise (since the outcomes of previous sessions might not contain enough variation).

We employ a two-stage variant of the ‘die-under-the-cup’ paradigm (Shalvi et al., 2011b; Fischbacher and Föllmi-Heusi, 2013), in which subjects’ beliefs are elicited in stage 1 before
performing the die task in stage 2. In stage 1, we elicit the subjects’ beliefs about majoritarian (i) behavior or (ii) normative beliefs in a previous session, and we vary whether participants are (i) aware or (ii) unaware of the upcoming opportunity to lie. Our hypothesis is that individuals might engage in belief manipulation in order to justify lying in the subsequent task when aware that they will face a lying opportunity shortly. By contrast, belief manipulation should not occur when subjects are unaware of the upcoming task. We compare these two treatments to identify what type of beliefs make lying more likely.

We build a theoretical model in which individuals may belong to one of three types: 1) Unconditionally honest - who never lie; 2) Unconditional liars - who always lie for personal gain; and 3) Conditional norm followers - who refrain from lying if they believe that most people are honest and disapprove of lying, but would lie otherwise. In this sense, a norm may exist but not be followed at a particular time if expectations are not met Bicchieri (2006, 2016). Uncertainty about the norm is modelled by assuming that the exact shares of unconditional liars and unconditionally honest individuals are uncertain. However, the shares of these two types clearly have implications for norm following. For conditional norm-followers, this uncertainty about “what most people do” generates an opportunity to engage in belief-manipulation, convincing themselves that an honesty norm does not apply to their particular situation, in order to lie more easily.

Our findings indicate that belief manipulation happens, and results in a greater incidence of lying in the subsequent die tossing task. Importantly, manipulation takes the form of people convincing themselves that lying behavior is widespread. This is in line with the predictions of our norm-based model. Intuitively, if lying behavior is believed to be common, this rules out that an honesty norm might apply. A conditional norm-follower who believes that lying is common will thus find it easier to lie, since this does not involve disobeying the norm. We also argue that our findings suggest that image considerations are not relevant in this setup. That’s because image-motivated subjects who care about appearing honest will find it easier to lie when they believe lying behavior to be rare, not widespread. Intuitively, when lying is uncommon, an observer is less likely to interpret a winning report as dishonest, since the majority of winning reports are indeed truthful. As shown by previous research, social image concerns may play a role when deciding when to

---

For example, we may believe that a majority of people disapprove of lying, but observe widespread dishonesty. In this case, the incongruence between beliefs and observed behavior leads us to misbehave, too.
lie. In our setup, however, they appear to be overridden by norm-based motives, reaffirming the importance of norms in driving behavior.

In order to investigate the role of normative (as opposed to empirical) considerations in driving lying or honest behavior, we also run a treatment in which, in stage 1, we elicit beliefs about what the majority believes should be done rather than about how the majority behaves. Our second finding is that, in contrast with beliefs about what others do, beliefs about their normative convictions are not distorted. Self-serving belief manipulation is not necessary, since believing that the majority disapproves of lying does not inhibit own lying. These findings are consistent with a norm-based model in which widespread honest behavior is a strong indicator of disapproval of lying (and thus that a norm of honesty exists and is followed), but the opposite doesn’t hold. Widespread disapproval of lying is not necessarily a strong indicator that most people behave honestly. To confirm this hypothesis, we ran a follow up experiment in which new subjects had to choose the statement they most agreed with after being provided information about the outcome of the die task. We compared two conditions. In the first condition, subjects were informed that the majority of individuals in the die experiment had refrained from lying, and were asked to guess how many subjects in that experiment also disapproved of lying. In the second condition, subjects were informed that the majority of individuals in the die experiment disapproved of lying, and were asked to guess how many subjects in that experiment also refrained from lying. Our findings strongly support our hypothesis. People interpret honest behavior as a strong indicator of disapproval of lying. By converse, expressing disapproval of lying is not seen as a strong guarantee of honesty.

2. Related literature

Our paper contributes to the literature on lying behavior. Abeler et al. (2018) conducted a thorough meta-study of the existing work on this topic, running additional experiments in order to discriminate among different possible theoretical accounts. One of their experiments (which is very relevant for us) studies the effect of shifting the subjects’ beliefs on lying behavior in the die task. To change beliefs, the authors employed suitable “anchors” in the form of a brief description of a “potential” experiment with high/low incidence of winning number reports. Although the anchors were successful in affecting beliefs, this had no effect on subsequent lying behavior, in contrast with our own findings. Our reading of this null result is that it is consistent with an environment in which op-
posing considerations (norm- or image-based) may cancel each other out. A reason why image concerns are absent in our experiment is that, differently from Abeler et al. (2018), it was conducted online, and thus involved no direct human exchange. Recent research (such as e.g., Cohn et al., 2018) has shown that machine-mediated interactions strongly reduce image-related considerations compared to interactions that directly involve human contact. Hence, although our results indicate that norm-following concerns are dominant in our setup, this does not imply that image concerns are not important. A significant body of research (including Abeler et al., 2018; Gneezy et al., 2018a) has provided convincing evidence that these matter. Our contribution should be rather seen as underscoring the importance of norm-based concerns in driving behavior. Our paper is also related to Diekmann et al. (2015), who compare a treatment in which subjects are informed about behavior in a previous session with a control condition without information feedback, finding that information triggers more lying compared to the control. However, they also find that, within the treatment session, behavior changed little with the specific nature of the previous session subjects were informed about.

A key novelty of our paper with respect to existing literature is that it employs a different method, based on self belief-manipulation, rather than factual evidence about previous sessions or anchors based on “potential” experiments. This has the advantage of generating sufficient variation in beliefs, while avoiding the use of hypothetical scenarios (which may or may not shift individual beliefs about real world-experiments).

Finally, our finding that empirical expectations (about what others do) play a more important role in giving (or not) “licence” to lie than normative expectations (about the normative beliefs of others) shares similarities with Bicchieri and Xiao (2009). In a dictator game experiment, they find that empirical expectations about the actions of other dictators matter more for individual behavior than normative expectations about what other dictators think ought to be done. A contribution of our paper with respect to previous work is that it provides a theoretical rationale for why empirical expectations matter more than normative expectations. Our explanation is based on an asymmetry in the inferences we draw from normative versus empirical information. An ongoing discussion in existing literature (e.g., Eriksson et al., 2015) is the extent to which learning about common behavior triggers inferences about what is approved of. Our hypothesis is that,

---

2Snowberg and Yariv (2018) discuss the robustness of behavior on mTurk versus laboratory settings.
3Normative expectations are second-order beliefs about the normative beliefs of others Bicchieri (2006)
if the majority behaves honestly, this is a strong indication that the majority also disapproves of lying, and thus an honesty norm exists and is followed. By converse, if the majority disapproves (or says it disapproves) of lying, this does not necessarily imply that most people behave accordingly. It is well possible that a norm of honesty exists but is not followed in spite of widespread disapproval (see also Bicchieri et al., 2018). To test this hypothesis, we designed an additional experiment, in which subjects were given one piece of information, either empirical or normative (“in a previous experiment the majority did not lie/disapproved of lying”) and were asked to report their normative or empirical expectations. Our findings strongly confirm our hypothesis.

The remainder of the paper is organized as follows. In Section 2, we describe the experimental design, while Section 3 presents a theoretical model where individuals are motivated by norm-following, and derive testable hypotheses. In that Section, we also briefly describe the predictions that would emerge in a model based on social image considerations. The results are presented in Section 4. Section 5 concludes.

3. Experimental Design

3.1. General Procedure

For our main experiment, we present data collected from 990 participants across 8 different conditions (we present a detailed breakdown of observations per treatment variation at the bottom of the respective figures in the results section). In addition, we collected data from 100 participants for the pre-experimental survey to obtain the truthful empirical and normative information, and another 203 participants for post-experimental survey to investigate the direction of belief distortion. All data was collected on Amazon Mechanical Turk. The mean age of the participants was 35.5 years and 49.8% of them were female. The average duration of the experiment from start to finish was 6 minutes and the average hourly payoff was $9.50 (including a show-up fee equivalent to $5 per hour), which is well above average mTurk pay (Hara et al., 2018).

4To ensure high quality data collection on mTurk, we utilize a combination of CAPTCHAs and sophisticated screening questions to avoid pool contamination. We applied the following restrictions to the participant pool: participants had to be in the U.S., approval rate was greater than 99%, and they had not taken this study before.

5In fact, recent evidence suggests that pay rates above what is typically considered an ‘ethical’ mTurker wage among social scientists (about $6) does not further increase performance in the realm of attention or engagement (Andersen and Lau, 2018).
3.2. Treatments

Our treatment conditions vary along two dimensions:

I) Whether belief elicitation concerned normative or empirical expectations

II) Whether participants were aware, prior to the belief elicitation phase, that they would play the dice task

Variations in (I) and (II) occurred between subjects. We refer to the treatment combinations as depicted below in Figure 1 and will discuss the detailed experimental procedure in the next section.
3.3. Detailed Procedure

Following the treatment randomization, the remainder of the experiment consists of two parts: the belief-elicitation phase and the dice task.\(^6\)

**Part I: Belief Elicitation Regarding Behavior/Beliefs of Others in Dice Task**

The first stage of the experiment consisted of a belief-elicitation procedure. We asked participants to report their beliefs about – depending on the treatment – the truth of alternative empirical or normative information. This procedure has the advantage of focusing subjects on a norm of honesty, but at the same time leaves them uncertain as to whether it is followed Cialdini et al. (1990) The combination of norm-focusing and uncertainty allows us to study the mechanism of belief distortion and subsequent cheating.

Upon signing a consent form, reading the instructions (see Appendix for screenshots), and passing a number of comprehension questions, participants were first presented with the incentivized belief-elicitation task. The belief task asked participants to indicate which of two mutually exclusive statements presented to them was true. The truthfulness of the statements was based on the results from the pre-experimental survey. We used data from a trial session that included questions regarding the appropriateness of lying on the task. From this sample, we collected both empirical and normative information about the frequency and appropriateness of lying and used them as part of the incentivized belief elicitation in the main experiment. At this point, depending on the exact treatment (details below), participants may or may not have already been aware that they will be engaging in the dice task with a cheating opportunity following the belief elicitation.

When presented with two mutually exclusive statements in the belief elicitation, participants had to choose one statement. A correct answer increased the participants’ payoff by $0.25. Importantly, the accuracy of the guesses was revealed only at the very end of the experiment and thus participants were not made aware of the actual truthfulness of the presented statements before the lying opportunity. This procedure was necessary in order to ensure that whether a norm of honesty applies to the situation remained uncertain (i.e. participants were not sure if their selected statement was correct) and did not directly affect participant behavior in the die task. The statements presented to the participants varied across treatments (see Table 1) and the exact wording is presented below:

---

\(^6\)From the perspective of the participant, the different parts of the experiment had no names to avoid potential priming.
Normative Treatment:

Please read the following statements and determine whether you believe them to be true or false. Which statement is true?

“In a similar study, most people said it is OK to lie for your own benefit.”

or

“In a similar study, most people said it is not OK to lie for your own benefit.”

Empirical Treatment:

Please read the following statements and determine whether you believe them to be true or false. Which statement is true?

“In a similar study, most people lied for their own benefit.”

or

“In a similar study, most people did not lie for their own benefit.”

It is important to note that this paper does not examine the impact on behavior of providing new information, i.e., we do not compare a situation where there is norm uncertainty versus one in which subjects are informed that a norm applies and is followed. Instead, we are interested in how people may manipulate their own beliefs in a self-interested manner. Across all treatments, we always elicit the participants’ beliefs regarding only one aspect (empirical or normative) that is instrumental in determining whether a norm of honesty applies or not. Belief elicitation alone is enough to produce our results.

Part II: Dice Task

After submitting their guess, participants were presented with the dice task. Participants clicked on a button to roll the electronic 6-sided dice and saw the outcome of the roll on their screen. Following the roll, participants were asked to write the outcome of the roll into an input field. Participants were told that there was no deception in the study, that the roll generator was fair and its outcome untraceable by the experimenter. Reporting a “5” yielded a payoff of $0.25 (the equivalent of $2.5 per hour), while reporting any other number yielded a zero payoff. Afterwards, participants received the respective payments
and were asked to complete a post-experimental questionnaire.

In order to study the relevant mechanisms at hand, the dice task was employed in one of two ways. In the Cheating Possibility Known (CPK) treatments, the dice task was public knowledge and announced before the belief elicitation phase. In the Cheating Possibility Unknown (CPU) treatments, the subsequent dice task was announced after the belief elicitation phase. This fine distinction allows us to test the behavioral mechanisms.

Importantly, to make treatments comparable, participants were always explained the mechanics of the dice task at the beginning of all treatments, that is before the belief elicitation phase. This ensured that participants knew which task the presented empirical and normative information referred to. What varied between CPK and CPU conditions was the explicit mentioning of whether the participants themselves would engage in the task after the belief elicitation. Note that this ensures that any potentially observable belief-distortion mechanism cannot be explained by demand effects, since, by design, their existence would merely produce a level effect and be unable to explain differences within the same treatment. For the purpose of comparison, we also ran a baseline condition (Control) in which participants only played the dice task. In this sense, the baseline condition is closer to a simplified version of a CPU condition where no belief distortion was possible.

4. Theoretical examination

In this section, we present a simple theoretical model of belief manipulation in the spirit of Bénabou and Tirole (2006, 2011, 2016) accounts of “motivated” belief distortion. At the heart of our argumentation is the idea that conditional norm followers may strategically distort their beliefs in order to engage in norm-transgressing behavior. Our modeling of norm-following is grounded in the literature on norms and in particular in the work by Bicchieri (2006).

Consider a setup where individuals belong to one of three types: (i) Unconditional Liars (UL) – incur no psychological cost from lying. (ii) Unconditional Honest (UH) – incur a prohibitively high cost from lying and, as a result, never lie in a given situation. (iii) Conditional Norm Followers (CN) – incur a psychological lying cost equal to $\theta > 0$ if the share of individuals who

(a) tell (or would tell) the truth (empirical requirement), and

(b) believe that lying is wrong (normative requirement)
are majoritarian, and no lying cost otherwise. CN types can thus be thought of as norm followers in the sense of Bicchieri (2006), while \( \theta \) can be seen as arising from cognitive dissonance generated from lying whilst knowing that lying is a violation of the norm (as defined by (a) and (b)). Let the share of UL be denoted as \( \alpha_{UL} \) and the share of UH as \( \alpha_{UH} \); the share of CN is \( 1 - \alpha_{UL} - \alpha_{UH} \).

The precise values of \( \alpha_{UH} \) and \( \alpha_{UL} \) are not perfectly observed and depend on the specific nature of the situation at hand.\(^7\) There are two possibilities: with probability \( q \in (0, 1) : \alpha_{UH} = h \) and \( \alpha_{UL} = l \) (state H); with probability \( 1 - q : \alpha_{UH} = l \) and \( \alpha_{UL} = h \) (state L), where \( h > 1/2 > l \) and \( 1 - h - l > 0 \). Our assumptions ensure that that, when state H occurs, the share of truthtellers is always majoritarian, while it is minoritarian otherwise. They also ensure that the share of CN types is always strictly positive.

**Relationship Between Empirical Compliance and Normative Beliefs**

With probability \( g \in (0, 1] \) UH types believe that lying is morally wrong, while with remaining probability \( 1 - g \) they believe lying is morally acceptable. Similarly, the probability that individuals of type UL disapprove of lying in a given situation is \( r \in (0, 1] \), while the probability that they believe lying to be acceptable in is \( 1 - r \). The following table provides a summary of the possible cases (states of the world) that may arise, with their prior probabilities. In the table, \( a = T \) (resp., \( L \)) indicates the action of telling the truth (lying), and \( m = W \) (resp., \( R \)) indicates the moral conviction that lying is wrong (right).

<table>
<thead>
<tr>
<th>Prior probability</th>
<th>Majoritarian behavior</th>
<th>Majoritarian moral conviction</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>( qg )</td>
<td>( a = T )</td>
<td>( m = W )</td>
<td>state H1</td>
</tr>
<tr>
<td>( q(1 - g) )</td>
<td>( a = T )</td>
<td>( m = R )</td>
<td>state H2</td>
</tr>
<tr>
<td>( (1 - q) r )</td>
<td>( a = L )</td>
<td>( m = W )</td>
<td>state L1</td>
</tr>
<tr>
<td>( (1 - q) (1 - r) )</td>
<td>( a = L )</td>
<td>( m = R )</td>
<td>state L2</td>
</tr>
</tbody>
</table>

\(^7\)The underlying idea is that the specific nature of the situation at hand will mould the lying costs of “unconditional” individuals, and this in turn will determine the share of UH and UL. The use of the term “unconditional” refers to the fact that the lying costs of these types do not depend on majoritarian behavior/normative beliefs (while type CN’s do), and this implies that, for a given situation, their lying behavior can be taken as fixed. However, this does not preclude that their lying costs may differ depending on the specific details of the dilemma they are presented with. We treat the mechanism through which this happens as a “black box”, since our main focus of interest are type CN.
Lying Decision

Since UL agents always lie and UH never do, the decision to lie or not only concerns CN. Importantly, CN types incur a positive psychological cost of lying (equal to $\theta > 0$) only when the state of the world is H1, since this is the only state where the majority of individuals satisfy both the empirical (the majority doesn’t lie) and normative (the majority thinks that lying is wrong) requirements. Suppose that a CN individual believes that the state is H1 with probability $p$, and let $\mu > 0$ be the monetary transfer obtained from lying (clearly enough, in this setup lying may only take the form of dishonestly reporting the winning number). The individual will lie if

$$\mu - \theta p > 0 \rightarrow \frac{\mu}{\theta} > p$$  \hspace{1cm} (2)

(where $\theta p$ is the expected lying cost) and will tell the truth otherwise. In what follows, we will assume that

$$\mu < \theta$$  \hspace{1cm} (3)

implying that a CN individual who believes that the state to be H1 with certainty will prefer to tell the truth.

Belief Elicitation

At $t = 0$, before the lying/not lying decision, subjects are asked to report their beliefs about the share (majoritarian/minoritarian) of

(i) people who lie (Empirical treatment), or

(ii) people who believe that lying is morally wrong (Normative treatment).

To answer the belief elicitation question, subjects are forced to engage in information processing – they retrieve information from past experiences, historical evidence and other relevant sources, and process it into a fully-formed belief. These beliefs are then used at $t = 1$, when subjects evaluate whether the return from lying offsets or not the psychological cost it may involve. Note that, for simplicity, the monetary incentives present in the experimental belief elicitation task are ignored in this stylized model (they are explicitly accounted for in the Robustness section presented in the Theoretical Appendix).

---

*We adopt the convention that, if indifferent, a CN agent tells the truth and does not manipulate his own beliefs. This is immaterial for our results.*
Consider first the empirical treatment. Suppose that, at \( t = 0 \), the subject’s belief is that the majority of people tell the truth (denoted as \( b = b_T \)). From table (1), it is straightforward to see that
\[
p(b_T) = \Pr(\text{state } H1 \mid b = b_T) = \Pr(\text{state } H1 \mid \text{majority select } a = T) = g
\]
On the other hand, if a subject believes that the majority of individuals lies (denoted as \( b = b_L \)), then the corresponding posterior is,
\[
p(b_L) = \Pr(\text{state } H1 \mid b = b_L) = \Pr(\text{state } H1 \mid \text{majority select } a = L) = 0.
\]
Consider now the normative treatment. Suppose that, at \( t = 0 \), the subject’s belief is that the majority of people believe that lying is morally wrong (denoted as \( b = b_W \)). Using Bayesian updating, we can compute
\[
p(b_W) = \Pr(\text{state } H1 \mid b = b_W) = \Pr(\text{state } H1 \mid \text{majority think } m = W) = \frac{gg}{gg + (1 - g)r}
\]
Similarly, we can compute posterior beliefs when the subject believes that the majority of people believe that lying is morally right (denoted as \( b = b_R \)).
\[
p(b_R) = \Pr(\text{state } H1 \mid b = b_R) = \Pr(\text{state } H1 \mid \text{majority think } m = R) = 0.
\]
The posterior beliefs (4) to (7) are computed assuming that, when updating, subjects are unaware that their beliefs about the majoritarian action or moral conviction might have been self-manipulated. The Robustness section presented in the Appendix discusses what might happen when this assumption is relaxed (full awareness).

**Self-Serving Information Processing**

We now turn to information processing at \( t=0 \). As we have seen, beliefs formed at the elicitation stage affect the subject’s posterior beliefs about the likelihood of state \( H1 \) and, thus, the decision to lie or not. This opens the door to the possibility that, at the belief formation stage, the individual may gain from engaging in self-serving manipulation. Gino et al. (2016) use the term **motivated Bayesian** to capture the notion that, when processing and encoding past experiences, historical narratives etc. into beliefs, people may do so in a biased way, in order to generate a self-serving interpretation of reality.
(what the literature calls “motivated beliefs”). Manipulation takes the form of ignoring or underweighting “unfavorable” evidence, or conveniently “massaging” the inferences being drawn, in a direction that suits the decision maker’s material interests. Our analysis follows this approach and assumes that CN agents behave as motivated Bayesians. Similar to Di Tella et al. (2015), we do not model the belief manipulation process explicitly, and instead adopt a black-box approach (see Bénabou and Tirole 2015 for a discussion of possible underlying manipulation processes). Formally, belief manipulation occurs if and only if (i) the subject is aware of the cheating task that will follow at t=1 and (ii) belief manipulation advances the subject’s material interests, in the sense that

$$\mu \cdot (\|b_{\text{manipulated}} - \|b_{\text{not manipulated}}\| > 0$$

where $b_{\text{not manipulated}}$ and $b_{\text{manipulated}}$ indicate unbiased and biased beliefs, and $\|b\|$ is an indicator function that takes value 1 if, given belief $b$ the agent lies in the lying task, and takes value 0 otherwise. Hence, at t=0, UN agents choose their beliefs by focusing exclusively on material payoff. Intuitively, the lying cost $\theta$ is salient at t=1, when the actual lying occurs, but not at t=0, when beliefs are formed. This is in line with existing models of belief manipulation that build on the notion of a “split self” with partially conflicting interests.

**Timing**

The timing of the game is as follows:

- **t=0** Belief elicitation task. The task triggers the formation of belief $b$ about majoritarian behavior (in the empirical treatment) or majoritarian moral conviction (in the normative treatment). During the belief formation phase, subjects may engage in self-serving belief manipulation (if they are aware of the lying task ahead).

- **t=1** Lying/truth-telling task. CN individuals compute expected lying costs and decide whether to lie or not. Payoffs are realized.

**Equilibrium**

Borrowing the terminology of Bénabou (2015), an intra-personal equilibrium (Perfect

---

9UL and UH types do not engage in belief manipulation since their behavior in the lying task is independent of their beliefs.
Bayesian equilibrium in the game between his date-0 and date-1 selves) for CN individuals satisfies the following conditions:

1. At \( t=1 \), subjects compute the probability that the state is \( H \) by Bayesian updating from their belief \( b \). They tell the truth if the posterior belief, \( p(b) \), satisfies \( p(b) \geq \frac{\mu}{\theta} \) and lie otherwise.

2. At \( t=0 \), subjects (when aware of the cheating task that will follow) choose to engage in belief manipulation iff, taking \( t=1 \) behavior as given, condition (8) above holds.

**Cheating Possibility Known (CPK) Condition**

We first characterize the equilibrium of the game in the CPK condition, in which subjects are informed at the outset about the cheating opportunity they will face later in the game. At \( t=0 \), subjects may therefore find it optimal to distort their beliefs, in order to modify their actions at \( t=1 \). As described in (8), in our model, belief manipulation will occur if it induces the subject to lie in a situation where, in the absence of manipulation, he wouldn’t lie. The only types of manipulation that may possibly satisfy this requirement take the form of convincing oneself that (i) the majority lie (empirical treatment), or (ii) the majority considers lying to be acceptable (normative treatment). Manipulating beliefs in the opposite direction (by convincing oneself that the majority doesn’t lie or does not approve of lying) would not advance the agent’s material interests, since it would induce him to judge state \( H1 \) more likely, and would therefore make him less willing to lie.\(^{10}\)

**Proposition 1:** (Nature of belief manipulation) *In equilibrium, belief manipulation may only take the form of: (i) inducing belief \( b_T \) instead of \( b_L \) *(empirical treatment)*, or (ii) inducing belief \( b_R \) instead of \( b_W \) *(normative treatment).*

In what follows, we accordingly consider the benchmark case where the underlying state is \( H1 \), so that, in the absence of belief manipulation, the agent would naturally come to the conclusion that the the majority of individuals tell the truth (empirical) or disapprove of lying (normative).\(^{11}\) The predictions obtained in the other possible states of the world are discussed in the Theoretical Appendix.

---

\(^{10}\)Formally, that’s because \( p(b_T) = g > 0 = p(b_L) \) and \( p(b_W) = \frac{qg}{qg + (1-q)r} > 0 = p(b_R) \).

\(^{11}\)As will become clear in the Empirical Results section, this also corresponds to the true state observed in the data.
Suppose that belief at \( t = 0 \) in the absence of belief manipulation is \( b_{\text{no manip}} \), so that \( p = p(b_{\text{no manip}}) \) at \( t = 1 \). In the empirical treatment, \( b_{\text{no manip}} = b_T \), while in the normative treatment \( b_{\text{no manip}} = b_W \). Optimal behavior at \( t = 1 \) is then given by:

\[
\begin{align*}
  a &= L \text{ if } p(b_{\text{no manip}}) < \frac{\mu}{\theta} \\
  a &= T \text{ if } p(b_{\text{no manip}}) \geq \frac{\mu}{\theta}
\end{align*}
\]

(9)

where \( p(b_{\text{no manip}}) \) is given by \( p(b_T) \) (given in (4)) in the empirical treatment, and is equal to \( p(b_W) \) (given in (6)) in the normative treatment. Alternatively, the subject may engage in self-serving belief manipulation, which results in \( b_{\text{manip}} = b_L \) (empirical treatment) or \( b_{\text{manip}} = b_R \) (normative treatment), triggering \( p = p(b_L) \) or \( p = p(b_R) \) at \( t = 1 \). Since \( p(b_L) = p(b_R) = 0 \), this ensures that the agent will then lie. Summing up, the net return from self-serving beliefs at \( t = 0 \) is equal to

\[
\begin{align*}
  0 & \text{ if } p(b_{\text{no manip}}) \leq \frac{\mu}{\theta} \\
  \mu & \text{ if } p(b_{\text{no manip}}) > \frac{\mu}{\theta}
\end{align*}
\]

(10)

This implies that, if \( p(b_{\text{no manip}}) \leq \frac{\mu}{\theta} \), the agent will not engage in belief manipulation, since this is not needed in order to lie at \( t = 1 \). On the other hand, if \( p(b_{\text{no manip}}) > \frac{\mu}{\theta} \), the agent finds it optimal to manipulate his beliefs at \( t = 0 \), in order to trigger lying at \( t = 1 \).

**Proposition 2:** (CPK condition) In the CPK condition, CN agents always lie at \( t=1 \) in both the empirical and the normative treatment. At \( t=0 \), they manipulate their beliefs iff

\[ p(b_{\text{no manip}}) > \frac{\mu}{\theta} \]

(11)

where (i) \( b_{\text{no manip}} = b_L \) (empirical treatment), and (ii) \( b_{\text{no manip}} = b_W \) (normative treatment), and do not manipulate otherwise.

**Cheating Possibility Unknown (CPU) Condition**

We now consider what happens in the CPU condition. Clearly enough, in this case at \( t = 0 \) the agent has no incentive to distort his beliefs, since he does not anticipate the future lying opportunity. At \( t = 0 \), he thus processes information unbiasedly in both treatments. The condition for agents to lie at \( t = 1 \) is thus \( p(b_T) < \frac{\mu}{\theta} \) (empirical treatment) and \( p(b_W) < \frac{\mu}{\theta} \) (normative treatment). Comparing these with the predictions obtained in the CPK condition, we obtain the following.
**Proposition 3 (CPU condition)**  If \( p(b_T) > p(b_W) \) then lying in the normative CPU treatment (weakly) exceeds lying in the empirical CPU treatment, while the opposite holds if \( p(b_T) < p(b_W) \).

**Hypotheses**

This Section describes the testable implications of the theory when the underlying true state is H1.\(^\text{12}\) We start with a definition.

**Definition** We say that *belief manipulation in the form of x* occurs in a given treatment (empirical/normative) if the share of subjects reporting belief x in CPK is higher (and the share reporting \( y \neq x \) is lower) than in CPU.

**Hypothesis 0** If belief manipulation occurs in a given treatment (empirical/normative), then lying in the CPK condition of that treatment must be higher than in the CPU condition.

Hypothesis 0 follows from our basic premise that subjects may engage in belief manipulation only when this advances their material interests, by making it easier for them to lie at t=1.

**Hypothesis 1**

a) In the empirical treatment: if belief manipulation occurs, it must be in the form of “the majority lie”.

b) In the normative treatment: if belief manipulation occurs, it must be in the form “the majority approves of lying”.

Hypothesis 1 follows from Proposition 1.

**Hypothesis 2** In the CPK condition, the share of subjects who report the winning number is the same in both the empirical and normative treatment.

This follows from Proposition 2.

**Hypothesis 3**

a) \( p(b_T) > p(b_W) \): If belief manipulation occurs in only one treatment, then it must be the empirical treatment.

\(^{12}\text{As we discuss in the Empirical Results section, our data indicate that this was indeed the underlying state of the world in our setup. The Theoretical Appendix discusses the predictions of the theory for the other possible states of the world.}\)
b) $p(b_T) < p(b_W)$: If belief manipulation occurs in only one treatment, then it must be the normative treatment.

Hypothesis 3 also arises from Proposition 2.

**Hypothesis 4** In the CPU condition:

a) $p(b_T) > p(b_W)$ If the share of subjects reporting the winning number is higher in one of the two treatments, then it must be the normative treatment.

b) $p(b_T) < p(b_W)$. If the share of subjects reporting the winning number is higher in one of the two treatments, then it must be the empirical treatment.

Hypothesis 4 follows from Proposition 3.

Recall that $p(b_T)$ is the posterior belief that the majority disapproves of lying, conditional on the majority telling the truth, while $p(b_W)$ is the posterior belief that the majority tells the truth, conditional on the majority disapproving of lying. The nature of the relationship between $p(b_T)$ and $p(b_W)$ is an empirical question. To address this question, we ran an additional experiment explicitly designed for this purpose, which is described below in the Empirical Results section.

**A Model of Social Image**

Our theoretical model focuses on agents who, at $t=1$, are motivated by norm-following concerns. A number of recent papers – Dufwenberg Jr and Dufwenberg Sr (2016); Khalmetski and Sliwka (2017); Abeler et al. (2018); Gneezy et al. (2018a) – have instead emphasized the role played by image-related motivations in driving truthtelling or lying behavior. It is therefore important to understand if the predictions described above would also emerge in a model where agents are motivated by image concerns rather than norm following. This is what we do in this section, by considering a model that is inspired by Gneezy et al. (2018a)’s model of social image.

Consider a setup where individuals belong to one of three types: (i) Unconditional Liars (UL) – incur no lying costs and always lie when this increases their material payoff (ii) Unconditional Honest (UH) – incur a prohibitively high psychological cost from lying and therefore never lie (iii) Image Conscious (IC) – in addition to material payoff, they (a) suffer a psychological cost $t > 0$ from lying, and (b) derive utility from their social image, which, as in Abeler et al. (2018); Gneezy et al. (2018a) “reputation for honesty” model and Khalmetski and Sliwka (2017), is defined as the probability that their report is
interpreted as being honest by an outside observer. The underlying idea is that, for IC types, being viewed as honest is an intrinsically valued part of their social identity. The key difference between this model and the main model is that conditional agents – type (iii) – are driven by image-considerations, rather than norm-related considerations (types (i) and (ii) are identical in both models). The utility of an IC type is given by:

\[
\begin{align*}
\delta \rho_0 - t & \text{ if report a losing number} \\
\mu + \delta \rho_1 - t & \text{ if report the winning number}
\end{align*}
\] (12)

where \( \delta > 0 \) is the weight given to social image, \( \rho_0 \equiv \Pr(\text{honest report} \mid \text{report} = \text{losing #}) \) and \( \rho_1 \equiv \Pr(\text{honest report} \mid \text{report} = \text{winning #}) \) are the social images that derive from reporting a losing and a winning number, and || is an indicator function that takes value of 1 if the report is dishonest, and zero otherwise. In what follows, we assume \( t \geq \delta \). This implies that, at \( t=1 \), type IC always prefers to report truthfully when he draws the winning number. Note also that type UH always report honestly, while type UL always report the winning number independently of their true draw. This implies that \( \rho_0 = 1 \), since all losing reports are truthful. As in the main model, we assume that the precise values of the shares of type UH and UL (\( \alpha_{UH} \) and \( \alpha_{UL} \)) are not perfectly observed. With probability \( q \in (0,1) \): \( \alpha_{UH} = h \) and \( \alpha_{UL} = l \) (state H); with probability \( 1-q \): \( \alpha_{UH} = l \) and \( \alpha_{UL} = h \) (state L), where \( h > 1/2 > l \) and \( 1 - h - l > 0 \). This implies that, in state H, the majority of subjects provide truthful reports, while in state L the majority of subjects provide a dishonest report (they report the winning number in spite of drawing a losing number).

**Social image**

We follow Bénabou and Tirole (2011) and take the view that, when assessing the likelihood that a report may or may not be truthful, the external observer knows the true underlying state (H or L). Letting \( \rho_i^1 \) denote the social image from reporting the winning number in state \( i \in \{H, L\} \), we prove in the Theoretical Appendix that,

**Lemma 1’**: \( \rho_H^1 > \rho_L^1 \).

---

13 Image-related concerns also play a key role in and Bénabou and Tirole (2011) model of laws and norms, in which agents derive utility from generating more favorable beliefs about their underlying type (in this setup: lying costs). Applied to the present context, that model would yield ambiguous predictions that depend on the precise features of the environment, and that are not easily testable.
Intuitively, when there are many liars (state L), the social image from reporting the winning number is low, since this is likely to be a lie. When it is known that the majority of people tells the truth (state H), on the other hand, a winning number report is likely to be truthful, and thus yields a higher social image. Image conscious agents are thus less inclined to lie when they believe that the state is L, in contrast with unconditional norm followers, who, as we have seen, would always lie in that case.

**Lying Decision**

Consider a CN individual who has drawn a losing number and who believes that the state is H with probability $p$. It is easy to show that, at $t=1$, the individual will lie and report the winning number if $p > \hat{p}$, where $\hat{p} = 1 - \frac{\mu - t - \delta(1 - \rho_H)}{\delta(\rho_H - \rho_L)}$, and will tell the truth otherwise.\(^{14}\) In what follows, we rule out the trivial cases $\hat{p} \geq 1$ and $\hat{p} \leq 0$ – in which IC agents always lie or always tell the truth independently of the state of the world – and we accordingly focus on the case where $\hat{p} \in (0, 1)$.\(^{15}\)

**Belief Manipulation**

As in the main setup, at $t=0$ belief manipulation occurs if and only if (i) the subject is aware of the cheating task that will follow at $t=1$ and (ii) belief manipulation advances the subject’s material interests.

**Equilibrium**

An intra-personal equilibrium for IC individuals satisfies: 1) At $t=1$, subjects compute the probability $p(b)$ that the state is H by Bayesian updating from their belief $b$. They tell the truth if $p(b) \leq \hat{p}$ and lie otherwise. 2) At $t=0$, subjects (when aware of the cheating task that will follow) choose to engage in belief manipulation iff, taking $t = 1$ behavior as given, condition (8) holds. 3) Given the underlying state of the world and the equilibrium behavior of IC types at $t=1$, the social images $\rho_0$ and $\rho_1$ are derived by Bayesian updating.

**Empirical Predictions**

We start off by analyzing the CPK condition. Recall that belief manipulation will occur if it induces the subject to lie in a situation where, in the absence of manipulation, he wouldn’t lie. Consider then the empirical treatment. It is straightforward to see that $p(b_T) = 1$ while $p(b_L) = 0$. Hence, at $t=1$, an IC type lies when $b = b_T$ and tells the

\(^{14}\)The expected payoff from lying is $\mu - t + \delta p \rho_H + \delta(1 - p) \rho_L$. When $p > \hat{p}$, this exceeds $\delta$, the payoff from a truthful report (since $\rho_0 = 1$).

\(^{15}\)This is guaranteed when $\mu \in (t + \delta(1 - \frac{1}{1+\delta}), 1 + t + \delta(1 - \frac{1}{1+\delta}))$. 

20
truth when \( b = b_L \). This implies that, differently from the norm-based model, in the social image model the only type of belief manipulation that may occur at \( t=0 \) takes the form of convincing oneself that the majority doesn’t lie! Intuitively, when most agents tell the truth an observer is likely to interpret a winning report as honest. By contrast, when most agents lie, reporting the winning number is likely to be interpreted as a sign of dishonesty, and thus delivers a low social image. Consider now the normative treatment. If \( g > r \), then \( p(b_W) > p(b_R) \) and hence the social image model predicts that the only type of belief manipulation that may occur takes the form of convincing oneself that the majority disapproves of lying.\(^{16} \) Recall that \( g \) is the probability that Type UH (Unconditionally Honest) disapprove of lying, and \( r \) is the equivalent probability for type UL (Unconditional Liars). Although the precise values of \( g \) and \( r \) are an empirical question, we conjecture that \( g > r \), and accordingly assume this to be the case in what follows.

**Hypothesis 1’**

a) In the empirical treatment: if belief manipulation occurs, it must be in the form of “the majority does not lie”.

b) In the normative treatment: if belief manipulation occurs, it must be in the form “the majority disapproves of lying”.

In order to ensure comparability with the predictions of our model, in what follows we consider the case where the underlying state is \( H_1 \), i.e. the majority tell the truth and believe lying to be wrong. As we have seen, when this is the case IC subjects do not need to engage in belief manipulation in order to lie. This implies that,

**Hypothesis 2’:** The shares of
(i) each of the possible stated beliefs and
(ii) subjects reporting the winning number
take the same values in all conditions: empirical CPK, normative CPK, empirical CPU and normative CPU.

To sum up, our analysis has shown that a social image model of lying a’ la Gneezy et al. (2018a) would deliver qualitatively different predictions to our norm-based model. Intuitively, while the norm-based model predicts that conditional agents are more inclined to lie when they believe lying to be common, the social image model yields the opposite

\(^{16} \)That’s because, from Bayesian updating, We have \( p(b_W) = \frac{gq}{gq+(1-q)r} \) and \( p(b_R) = \frac{(1-g)q}{(1-g)q+(1-q)(1-r)} \).
prediction. As a result, the two models generate different empirical implications.

5. Empirical Results

Our analysis will vary by the extent of knowledge regarding the upcoming lying opportunity (CPK vs. CPU), as well as the type of belief elicitation (empirical vs. normative). Because the CPK and CPU conditions are the same except for the knowledge about the subsequent die task, any difference in belief distributions between these two treatments indicates active belief distortion.

We will unpack our findings in multiple steps. First, we identify the “true” state of the world as revealed by the data, showing that, indeed, a norm for honesty as defined in Bicchieri (2006) (state H1) applied to our setup. Second, since hypotheses 3 and 4 are conditional on the relationship between \( p(b_T) \) – the belief that the majority disapproves of lying when the majority tell the truth – and \( p(b_W) \) – the belief that the majority tells the truth when the majority disapprove of lying – we discuss the additional experiment we designed to this purpose, and spell out how its findings inform hypotheses 3 and 4. Third, we discuss the results from our main experiment, comparing beliefs and lying behavior in the different treatments (normative/empirical) and conditions (CPK/CPU), and relating the findings to our hypotheses.

5.1. True State of the World

We first report the outcomes from a trial session that included questions regarding the appropriateness of lying on the task, and that was used to incentivize belief elicitation in the main experiment based on a total of 100 participants. The data indicate that the majority of individuals (83%) disapproved of lying, and refrained from lying (only 37% reported the winning number, which suggests a lying rate of approx. 21%). This latter finding is further corroborated by the data from our main experiment, where the number of winning reports across all treatments was 35%). This indicates that the true state of the world corresponded to H1: the majority of people disapproved of lying and did not lie.

5.2. Experiment on the Relationship between \( p(b_T) \) and \( p(b_W) \)

To test the relationship between \( p(b_T) \) and \( p(b_W) \), we designed a simple and incentive-compatible experiment that we run on mTurk.\(^{17}\) As before, participants were explained the

\(^{17}\) n=300, 2 conditions in between-subjects design, $0.50 show-up fee, 99% approval rate, U.S. residents.
setup of the original experiment and were then randomly allocated to one of two treatment variations. In each treatment, participants received an hypothetically true empirical or normative information, namely (1) empirical (“the majority did not lie”) or (2) normative (“the majority did not approve of lying”) and were then asked to guess the exact percentage of: (i) those who approve of lying (after being told (1) or (ii) those who did lie (after being told 2). This approach allows us to elicit normative (posterior) beliefs from empirical information and vice versa. In particular, the answer to (i) delivers an estimate of $p(b_T)$, namely the probability that a norm for honesty applies when the majority tells the truth, while the answer to (ii) delivers an estimate to $p(b_W)$, namely the probability that a norm for honesty applies when the majority disapproves of lying.

The belief elicitation was incentive compatible, and the information relied on the effective behavior in our previous experiment. Importantly, participants were only presented with one statement and only provided one guess:

- “The majority of participants did not lie for their own benefit. How many participants approved of lying?”
- “The majority of participants did not approve of lying. How many participants lied for their own benefit?”

Figure 2 illustrates our findings and the results are clear: observed empirical information strongly affects the normative belief, but not vice versa. When participants are told that the majority of participants did not lie, they infer that the majority (77.48%) disapprove of lying (Figure 2, left panel). We do not observe the reverse to the same degree (Figure 2, right panel): When participants are told that the majority disapproves of lying, they infer that only 47.65% will be honest. The difference is highly significant with $p<0.001$ (Mann-Whitney U). Thus, these results indicate that, in our setup, norm-inference is much stronger in one direction (empirical $\rightarrow$ normative) than in the other (normative $\rightarrow$ empirical).

In concluding that $p(b_T) > p(b_W)$, we can restate hypotheses 3 and 4 as follows:

**Hypothesis 3** If belief manipulation occurs in only one treatment, then it must be the empirical treatment.

---

18In both conditions on average participants think that “majority” means 71.44% and 71.84%, respectively.
Hypothesis 4 In the CPU condition, if the share of subjects reporting the winning number is higher in one of the two treatments, then it must be the normative treatment.

5.3. Main Experiment: Beliefs and Lying Behavior

Figure 3 summarizes our key results. First, by inspecting the empirical condition, we see that belief distortion indeed occurs, and takes the form of subjects convincing themselves that the majority lies. When participants did not know about the upcoming cheating opportunity (CPU), 38.1% of participants said that they believe that a majority of previous participants lied. However, when the opportunity to lie was known (CPK), 62.7% of participants reported this state, with the difference being highly statistically significant (EPT, p < 0.001). This offers compelling evidence for active belief distortion because the only difference between both conditions is whether or not one’s own lying opportunity could be anticipated at the time of the belief elicitation. Our theoretical model predicts that belief distortion should translate into higher lying rates overall. This is exactly what we observe (EPT, 40.5% vs. 21.4%, p < 0.01). These findings are thus in line with hypotheses 0 and 1.

Our results also indicate that, differently from the empirical condition, in the normative condition, distortion did not occur. Regardless of whether the cheating opportunity was announced at the time of belief elicitation, the fraction of participants stating that the majority of previous participants approved of lying remains invariant (34.9% vs. 39.4%,
EPT, p=0.4575). As expected, the absence of belief distortion also leads to an absence of a significant difference in lying behavior (38.0% vs. 40.2%, EPT, p=0.7216). In sum, the empirical analysis indicates that belief manipulation occurred only in the empirical treatment, which in line with our hypothesis 3. The significant difference in belief manipulation between empirical and normative treatment can be explained by the asymmetry in norm inference discussed above. When we believe that most people do not lie, we also infer that they disapprove of lying, a powerful inducement to be honest. On the contrary, believing that most people disapprove of lying does not necessarily lead us to conclude that they also refrain from lying. The norm against lying in this case may be weakened by the concurring
belief that people disobey it, providing a conditional follower with a reason to transgress.

To test hypothesis 2, we compare the incidence of lying in empirical and normative CPK (i.e., when subjects anticipate the cheating opportunity and thus may choose to distort beliefs). In the normative treatment, as discussed, our results indicate that CN subjects were equally at ease lying independently of their beliefs, and thus felt no need to engage in belief manipulation. Conversely, in the empirical treatment CN subjects did feel the need to manipulate their beliefs, in order to lie more easily. In both cases, though, the final outcome should be the same, namely, that CN subjects lied. As predicted, the reported shares of the winning number are statistically indistinguishable with 40.5% (empirical) vs 40.2% (normative).

Hypothesis 4 predicts that, in the CPU condition (when subject do not anticipate the upcoming lying opportunity), the share of reported winning numbers in the normative treatment should be higher than in the empirical treatment. This follows since the lack of belief manipulation inhibits lying behavior in the empirical treatment (where, as we have seen, when offered the opportunity NC subjects do choose to manipulate beliefs in order to facilitate subsequent lying), but not in the normative treatment (where belief manipulation is not needed for lying to occur). Again, the data provide support for our hypothesis: 21.4% (empirical) vs 37.9% (normative), which is highly significant with p<0.01.

We substantiate our previous results through the lens of a Logit regression analysis, examining both stated beliefs and lying behavior. We ensure the robustness by including a battery of controls (age, gender, risk (SOEP), and CRT score). The regressions fully confirm the previous results: when comparing our empirical CPU and CPK conditions, both lying behavior and the reported beliefs are significantly different. Conversely, and in line with our previous analysis and theoretical examination, neither is true when comparing the normative CPU and CPK conditions. We also provide robustness checks for hypotheses 2 and 4 with respect to the lying rates between empirical and normative conditions: the former suggests that the lying rates between empirical CPK and normative CPK would be indistinguishable, whereas the latter suggests that lying would be more prevalent in normative CPU compared to empirical CPU. Our regression results suggest exactly this and are displayed on the right-hand side of Panel B of Table 1.
Table 1: Logit Regression Analysis of Reported Beliefs and Lying Behavior

<table>
<thead>
<tr>
<th>Panel A</th>
<th>DV: Reported Beliefs</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (Baseline: CPK Condition)</td>
<td>Empirical CPU ( (I = \text{Majority Lies}) )</td>
<td>-1.005***</td>
<td>-0.945***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Normative CPU ( (I = \text{Majority Does Not Approve of Lying}) )</td>
<td>-0.192</td>
<td>-0.250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>252</td>
<td>252</td>
<td>256</td>
<td>255</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B</th>
<th>DV: Lying Behavior</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment (Baseline: CPK Condition)</td>
<td>Empirical CPU ( (I = \text{Majority Lies}) ) | Normative CPU ( (Hypothesis 2) )</td>
<td>0.401***</td>
<td>0.425***</td>
<td>0.987</td>
<td>1.031</td>
</tr>
<tr>
<td></td>
<td>Normative CPU ( (I = \text{Majority Does Not Approve of Lying}) ) | Normative CPU ( (Hypothesis 4) )</td>
<td>0.913</td>
<td>0.837</td>
<td>2.246***</td>
<td>2.089***</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>252</td>
<td>252</td>
<td>256</td>
<td>255</td>
<td>253</td>
</tr>
</tbody>
</table>

Panel A and Panel B (Self-Serving Conditions): Logit regressions (odds ratios reported) with robust standard errors clustered at the individual level. Control variables include Age, Gender, Risk (SOEP), CRT score. Constant estimated but not displayed. Significance levels: *p < 0.10, **p < 0.05, ***p < 0.01.

Finally, we also report on conditional lying behavior.\(^{19}\) Figure 4 indicates that, in the empirical CPK treatment, the percentage of participants reporting the winning number was significantly higher for those who said that most people lie than for those who said the opposite (50.6% vs. 23.4%, \(p=0.0104\)). The same holds for CPU (31.3% vs. 15.4%, \(p=0.0351\)), but, in line with our model, the difference in the likelihood of reporting the winning number for different beliefs (essentially, a measure of the predictive power of beliefs for lying behavior) is substantially larger in empirical CPK (50.6%-23.4%=27.2%) than in empirical CPU (31.3%-15.4% = 15.9%). Intuitively, that’s because, in the CPK treatment, CN subjects purposefully manipulate their beliefs in order to lie in the subsequent task, which strengthens the correlation between lying behavior and holding the belief that the majority of people lie.

\(^{19}\)Because we run pairwise comparisons, we take into account that such multiple comparisons cause an inflation of the type-I-error. To counteract this inflation, we capitalize on the false discovery rate correction as introduced by Benjamini and Hochberg (1995), which has shown to be superior to Bonferroni with respect to statistical power.
In the normative treatment, we find that, generally speaking, the percentage of participants reporting the winning number was higher for those who said that most people approve of lying than for those who said the opposite.\textsuperscript{20} The key statistic for our purposes, however, is again the extent to which the predictive power of beliefs for lying behavior

\textsuperscript{20}Whilst our model provides no specific predictions in this respect (since unconditional subjects are treated as a “black box”), it is plausible that this could be due to unconditional truth-tellers being more likely than unconditional liars to believe that most people disapprove of lying.
changes as we move from CPU to CPK. The data show that the difference in the probability of reporting the winning number when holding different beliefs is the same across the two conditions (normative CPK: 52.0% -32.5% =19.5%, normative CPU: 51.1% - 31.1%=20%). This fits well with our theoretical account since, as we have seen, in the normative treatment, CN subjects did not engage in belief manipulation and, as a result, there is no reason to expect the correlation between beliefs and lying behavior to differ in CPK and CPU.²¹

6. Discussion and Conclusion

Lying is defined as asserting something that is false with the intention to lead someone else to believe it (Isenberg, 1968). One of the main reasons lying is bad is that it diminishes trust between people, making life and social interactions difficult and time-consuming. Consequently, in many societies there is a strong norm against lying, and a shared understanding of what counts for extenuating reasons to lie (e.g., to save lives, to confound an enemy, among others). To exist, a social norm against lying requires the following conditions: a) we have to expect that most people in our reference network do not lie (empirical expectation), b) we have to expect that most people in our reference network disapprove of lying (normative expectation), and c) we must have a preference for not lying conditional on these expectations (Bicchieri, 2006). Having a conditional preference implies that if our expectations were to change, then we may stop following the norm (at least temporarily). For example, we may come to realize that a sizable number of people violate it, or that transgressions are no longer met with disapproval or punishment. In that case, the norm may lose its grip on us.

Norm violation often provides a benefit to the transgressor. If we can convince ourselves that a norm does not apply to the situation we are in, then we have reason to transgress it. Often, we face uncertainty about the norm because it is not fully clear what other people are doing or have done in the same situation. Uncertainty about what the norm is or whether it is presently followed can be solved in a self-serving way if we can provide reasonable justifications for norm-violation (Bicchieri and Chavez, 2013). For example, evidence consistent with the desired behavior often receives preferential treatment (Kunda, ²¹All of our results are fully robust to a regression analysis that mimics our previous exercise (with and without controls). The results are available upon request.)
Or, if one must decide what to believe to be true about a norm, one may give more weight to selfish motivations, since there is uncertainty as to what belief is true (Schweitzer and Hsee, 2002). Finally, it is sometimes possible to give a subjective interpretation of a norm (e.g. fairness). If so, people will choose the interpretation that lowers the difference between what is normatively required and what one does (Spiekermann and Weiss, 2016). The "best" action is a function of an epistemic state, and one often chooses whatever information makes selfish actions appropriate. As a consequence, in situations where we either think a norm is not presently followed or it is not clear what the norm is, we often find justifications for self-serving behavior.

Much of the behavioral literature on lying assumes that the norm against lying has been internalized, becoming a moral norm (Bicchieri, 2006). In this case individuals have an unconditional preference for telling the truth: that is, they will be insensitive to what other people do or approve of. This social insensitivity does not mean that such a moral agent will not lie. Instead, when she does, she will experience an internal conflict. For a moral agent, lying has internal costs, threatening one’s self-image as a moral being (Klein et al., 2017; Mazar et al., 2008). The act of lying produces ethical dissonance, experienced as a conflict between right and wrong behaviors (Rabin, 1994; Barkan et al., 2012). So, if someone commits a moral violation, they must engage in ‘ethical maneuvering’ to find a compromise between profit and self-image (Shalvi et al., 2015). In this case, individuals look for self-serving justifications to provide reasons not only for lack of pro-sociality (Di Tella et al., 2015; Exley, 2015), but also potentially for unethical behavior. The internalized norm hypothesis predicts that explicitly making a situation ethically salient intensifies the threat to one’s self-image, decreasing the power of self-justifications (Mazar et al., 2008). Conversely, when ethical boundaries are blurred, we observe more unethical behavior (Pittarello et al., 2015; Shalvi et al., 2011a).

Our experimental design varies both the nature of belief elicitation (empirical – about what others do – versus normative – about what others think should be done) and the timing at which participants learn about their own opportunity to engage in lying. In the "Cheating Possibility Known" (CPK) condition, participants know prior to the belief elicitation that they will have the opportunity to lie on the task, while in the "Cheating Possibility Unknown" (CPU) condition, participants learn about the opportunity to lie only after the belief elicitation has taken place. Our working assumption is that belief distortion takes place whenever individuals are aware of their own lying opportunity at the
time of belief elicitation. If we observe a change in the belief distribution, we would also assume that this translates into higher rates of lying. Our experiment yields a number of interesting results. We find a significant relationship between lying and expressing a belief that most people lie. We also find strong evidence of belief distortion; individuals choose to believe that most people lie because this belief is aligned with their own preferred subsequent action: lying. This, in turn, leads to higher cheating rates overall, which supports the notion that the reason for individuals to distort their beliefs is to increase cheating. Conversely, when belief elicitation concerns what others approve of, we do not observe belief distortion. As argued (and conclusively shown in a follow-up study), this is driven by the fact that normative and empirical information vary in their signaling content: words are cheap, but actions are costly, which directly affects the decision to lie and distort one’s beliefs. From a policy perspective, we recommend sending unambiguous information about what behavior is common in a particular environment. If the common behavior is negative, the best option would be identifying subgroups that behave in a positive way and broadcasting their behavior. As to normative messages, the results are mixed. Words and deeds often differ, and stressing what people approve of may not necessarily induce good responses. The best option seems to be a combination of congruent empirical and normative information, as positive and unambiguous as possible (see Schultz et al., 2007).
References


Appendix

I. Additional Results Main Experiment

This section consists of two parts. First, we perform a robustness exercise in which we mirror the main experiment but change the beneficiary of the lie. For our purposes this means that instead of lying for one’s own benefit (self-serving conditions), the beneficiary of the lie is now a charity (UNICEF) and present the results in Figure A.1. 22

In the second section, we examine the properties of conditional liars for the other-regarding conditions under both empirical and normative uncertainty. The results are presented in Figure A.2. In line with our theoretical framework, we observe no relationship between cheating and stated beliefs.

I.1: Other-Regarding Conditions: Beliefs and Lying Behavior

We now turn to the complementary analysis for the other-regarding condition (where the beneficiary of the lie is a charity). As predicted, behavior looks vastly different along all dimensions: beliefs, lying rates, and conditional lying rates. In fact, we do not observe any significant belief distortion, neither in the empirical condition (51.6% vs. 46.0%, EPT, p=0.3777) nor in the normative condition (47.3% vs. 42.5%, EPT, p=0.4433). As expected the absence of belief distortion also leads to an absence in significant changes of lying behavior (Empirical: 25.4% vs. 31.0%, EPT, p=0.3270; Normative: 24.8% vs. 29.1%, EPT, p=0.4353).

Noteworthy, we observe low overall levels of lying in the purely other-regarding conditions. We attribute this to both the absence of self-serving motives as well as the induced uncertainty about the appropriateness of altruistic behavior, which corroborates existing literature (Di Tella et al., 2015; Exley, 2015). Seemingly, this reasoning is even stronger when pure altruism results from a deviant act. This finding is in line with our hypothesis $H_3$ in that lying for others provides enough of an independent (altruistic) justification that there should be no need for belief distortion. The norm-based hypothesis predicts that one will choose a justification for not following a norm of most people lie if it is in one’s

---

22We operationalized the data collection by altering the order in which the same participant would make a decision in the self-serving and other-regarding condition. Although participants were made aware of this only after the first decision was made, in an attempt to retain a clean comparison, here and throughout the paper we only look at the first decision made. We consider the spillover effects of decisions in a companion paper.
self-interest to violate it (Bicchieri and Chavez, 2013), but will need no belief distortion if violating it benefits others.

We substantiate our previous results through the lens of a Logit regression analysis, examining both stated beliefs and lying behavior. We ensure the robustness by including a battery of controls (age, gender, risk (SOEP), and CRT score). The regressions fully confirm the previous results.

---

**Figure A.1:** Results from belief elicitation (top panel) and lying behavior (bottom panel) for the self-serving conditions (participant is beneficiary of the lie). Results are broken down for the Empirical (left panel) and Normative (right panel) conditions. Dotted red line illustrates the expected value of 16.67%. Whiskers indicate 95% confidence intervals.
Table A2: Logit Regression Analysis of Reported Beliefs and Lying Behavior

<table>
<thead>
<tr>
<th>Panel A</th>
<th>DV: Reported Beliefs</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>(Baseline: CPK Condition)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical CPU</td>
<td>(1 = Majority Lies)</td>
<td>0.223</td>
<td>0.272</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.253)</td>
<td>(0.260)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normative CPU</td>
<td>(1 = Majority Does Not Approve of Lying)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.193</td>
<td>0.372</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.252)</td>
<td>(0.271)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>252</td>
<td>252</td>
<td>256</td>
<td>256</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B</th>
<th>DV: Lying Behavior</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>(Baseline: CPK Condition)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical CPU</td>
<td>(1 = Majority Lies)</td>
<td>0.759</td>
<td>0.831</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.214)</td>
<td>(0.240)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normative CPU</td>
<td>(1 = Majority Does Not Approve of Lying)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.802</td>
<td>0.646</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.227)</td>
<td>(0.196)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>252</td>
<td>252</td>
<td>256</td>
<td>256</td>
<td></td>
</tr>
</tbody>
</table>

Panel A and Panel B (Other-Regarding Conditions): Logit regressions (odds ratios reported) with robust standard errors clustered at the individual level. Control variables include Age, Gender, Risk (SOEP), CRT score. Constant estimated but not displayed. Significance levels: *p < 0.10, **p < 0.05, ***p < 0.01.

I.2: Other-Regarding Conditions: Conditional Lying

In line with our theoretical foundation and in contrast to the previous findings, we do not observe any significant relationship between stated beliefs and subsequent lying behavior when a charity is the beneficiary of a lie, neither in the empirical nor in the normative condition. The results are presented in Figure A.2.

Uncertain Empirical Information: When empirical information is uncertain, cheating rates are statistically independent of the stated belief, both in the CPK condition (33.6% vs. 26.3%, p=0.2025) and the CPU condition (24.8% vs. 20.0%, p=0.8773).

Uncertain Normative Information: In the same vein, we do not observe any significant relationship between stated beliefs and lying behavior when the charity is the beneficiary of the lie. This is true for both the CPK condition (24.8% vs. 33.8%, p=0.1261) and the CPU condition (19.8% vs. 27.7%, p=0.1428).\(^{23}\)

\(^{23}\)All of our results are fully robust to a regression analysis that mimics our previous exercise (with and without controls). The results are available upon request.

37
Figure A.2: Results from conditional lying behavior for the self-serving conditions (individual is beneficiary of the lie). Results are broken down for the Empirical (top panel) and Normative (bottom panel) conditions. Whiskers indicate 95% confidence intervals. Stars indicate significant differences at the conventional levels of *$p < 0.1$, **$p < 0.05$, and ***$p < 0.01$ and are false discovery rate corrected.
II. Screenshots Main Experiment

Exemplarily, we present below the screenshots for the empirical conditions. Normative treatment is identical, except for the statements phrased normatively. Screenshots from all treatment variations are available upon request.
Thank you for choosing to participate in this study! It is important that you read all of the instructions carefully to maximize your earnings. This is a survey based study where you will be asked to answer questions and complete simple tasks. At the end of the study you will be asked to fill out a short questionnaire and demographic questions.

It is expected that this study will take approximately 10 minutes. There is no deception in this study. Everything you see or read is true.

You begin the survey with an endowment of $0.50 and will have the opportunity to earn more based on your answers. You will be presented with a set of statements and asked to determine which statement you believe to be true. Guessing correctly will earn you an additional $0.25.

You will then roll an electronic 6-sided dice. This roll generator is fair and unrecorded. The experimenters will not know what number you rolled and the rolls are generated at random. You will report what you rolled and that will determine what, if any, additional earnings you will receive.

You may stop participating in this study at any time; however, you will not receive your payment unless you complete the entire survey through to the end. A payment code will be sent to you upon completion. Be sure to click the continue buttons at the bottom of each page after to continue through the survey.

This is a question to test your comprehension of the instructions:

Is the dice you will roll recorded by the experimenter?

☐ Yes
☐ No
Please read the following statements and determine whether you believe them to be true or false. If you answer correctly you will earn $0.25 more in bonus payment.

Which statement is **TRUE**?

- [ ] "In a similar study most people did not lie for their own benefit."
- [ ] "In a similar study most people lied for their own benefit."

A roll of 1, 2, 3, 4, or 6 does not result in additional earnings for you. A roll of 5 earns you an additional $0.25.

What number did you roll?

- [ ] 1
- [ ] 2
- [ ] 3
- [ ] 4
- [ ] 5
- [ ] 6
Please read the following statements and determine whether you believe them to be true or false.
If you answer correctly you will earn $0.25 more in bonus payment.

Which statement is TRUE?

- *In a similar study most people lied for the benefit of a charity.*
- *In a similar study most people did not lie for the benefit of a charity.*

A roll of 1, 2, 3, 4, or 6 does not result in earnings for you.
A roll of 5 will result in a $0.25 donation to the United Nations International Children’s Emergency Fund (UNICEF).

What number did you roll?

- 1
- 2
- 3
- 4
- 5
- 6
What is your gender?

- Male
- Female

What is the highest level of school you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate (high school diploma or equivalent including GED)
- Some college but no degree
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Doctoral degree
- Professional degree (JD, MD)

How old are you in years?

- 

How do you see yourself:

Please indicate on the scale if you are generally a person who is fully prepared to take risks or do you try to avoid taking risks?

0 means: "not at all willing to take risks"
10 means: "very willing to take risks"

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
</table>

Risk Willingness
What is your general opinion of charitable organizations?

'1' = I do not support the work of charitable organizations.  
'10' = I fully support the work of charitable organizations.

What is your general opinion of the United Nations International Children's Emergency Fund (UNICEF)?

'1' = I do not support the work of UNICEF.  
'10' = I fully support the work of UNICEF.
7. Theoretical Appendix

7.1. Additional results: Other regarding condition

Suppose that the beneficiary of the lie is a charity rather than the subject as in the condition analyzed in the main text. If a CN subject lies, he may incur a psychological cost but also experiences a warm glow, denoted by $\gamma \mu \geq 0$, where $\mu$ is the payment received by the charity and $0 \leq \gamma \leq 1$ is a measure of altruistic concern towards the charity. Note that the lie now generates no direct monetary return for the agent. Two observations are in order. First, recall that belief manipulation in our model is motivated by personal monetary returns from lying at $t = 1$. Since the direct monetary return from lying is now zero, it follows that the agent will never choose to distort his beliefs at $t = 0$.

Second, optimal behavior at $t = 1$ is thus given by:

\[
\begin{align*}
    a &= L \text{ if } p(b) < \frac{\gamma \mu}{\theta} \\
    a &= T \text{ if } p(b) \geq \frac{\gamma \mu}{\theta}
\end{align*}
\]

(13)

Compared with its equivalent in the self-serving condition (given by (9)) the requirement for lying to occur is now more stringent, since $\mu$ is multiplied by $\gamma \leq 1$. This implies,

**Proposition** (Other-regarding condition)

a) Belief manipulation never occurs in the other regarding condition, implying that CPK and CPU conditions yield the same amount of lying.

b) Comparing across treatments, if $p(b_T) > p(b_W)$ then lying in the normative treatment (weakly) exceeds lying in the empirical treatment, while the opposite holds if $p(b_T) < p(b_W)$.

c) Within each treatment (empirical/normative), lying in the self-serving condition (weakly) exceeds lying in the other regarding condition.

**Proof** In text above.

7.2. Robustness

TBA

---

1This argument is strengthened if we explicitly take into account the monetary incentives present in the belief elicitation task, since in that case belief manipulation results in a monetary (opportunity) cost at $t = 0$. 

45
7.3. Underlying state different from H1

TBA

7.4. Social Image model

Proof of lemma 1': Let $\eta \in [0,1]$ denote the probability that a type IC who draws a losing number dishonestly reports the winning number. By Bayesian updating, $\rho_1 = \frac{1}{6} + \frac{1}{6}[\alpha_{UL} + \eta(1 - \alpha_{UH} - \alpha_{UL})]$, decreasing in $\eta$. Substituting for $\eta = 0$ (resp. $\eta = 1$) in these expressions, we obtain the upper (lower) bound for $\rho_1$. This proves that $\rho_1 \in [\frac{1}{1+5(1-\alpha_{UH})}, \frac{1}{1+5\alpha_{UL}}]$. To prove that $\rho_1^H > \rho_1^L$, note that since $h > 1/2$, the lower bound for $\rho_1^H$, $\frac{1}{1+5(1-h)}$, exceeds the upper bound for $\rho_1^L$, namely $\frac{1}{1+5h}$. 