CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2019-07

Cristina Bicchieri, Eugen Dimant, Silvia Sonderegger

August 2019

It's Not A Lie If You Believe It: On Norms, Lying, and Self-Serving Belief Distortion

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

The most recent version of the working paper can always be downloaded following this link.

It's Not A Lie If You Believe It: On Norms, Lying, and Self-Serving Belief Distortion

Cristina Bicchieri[a], Eugen Dimant[a,b,**], Silvia Sonderegger[c]

[a] *University of Pennsylvania and Behavioral Ethics Lab*
[b] *Identity and Conflict Lab*
[c] *University of Nottingham and CeDEx*

Abstract

This paper focuses on belief distortion in the context of lying decisions. We employ a two-stage variant of the "dice under the cup" paradigm, in which subjects' beliefs are elicited in stage 1 before performing the dice task in stage 2. In stage 1, we elicit the subjects' beliefs about (i) majoritarian behavior or (ii) majoritarian normative beliefs in a previous session, and, in order to identify self-serving belief distortion, we vary whether participants are aware or unaware of the upcoming opportunity to lie in the dice task. We find that belief distortion occurs, but only with a specific kind of beliefs. When subjects are aware of the dice task ahead, they convince themselves that lying behavior is widespread in order to justify their lying. In contrast with beliefs about majority behavior, we find that beliefs about the extent to which lying is disapproved of are not distorted. Believing that the majority disapproves of lying does not inhibit own lying. These findings are consistent with a model where agents are conditional norm-followers, and where honest behavior is a strong indicator of disapproval of lying, but disapproval of lying is not a strong indicator of honest behavior.

Keywords: Cheating, Experiment, Lying, Social Norms, Uncertainty
JEL: C72, C91, D8, D9

*This work benefited from discussions with Johannes Abeler, Pierpaolo Battigalli, Roland Bénabou, Valentina Bosetti, Alexander Cappelen, Gary Charness, Christine Exley, Fabio Tufano, Tobias Gesche, Sandy Goldberg, Peter Graham, David Henderson, Agne Kajackaite, Michel Maréchal, Daniele Nosenzo, Gloria Origgi, Kai Ou, Guido Tabellini, Bertil Tungodden, Roberto Weber, and Joël van der Weele. 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), the 2018 Norms and Behavioral Change Workshop, and the 12th NYU-CESS Conference for helpful feedback, as well as the input received from seminar attendees at Bocconi, Princeton, Nottingham, East Anglia, Oxford, Pennsylvania, San Diego, and Zurich.

**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: August 16, 2019

Jerry tries to find a way to trick a lie detector

George Costanza: "Jerry, just remember...it's not a lie if you believe it!"

Seinfeld, TV Show, "The Beard" (Season 6, Episode 16, 1995)

1. Introduction

Social norms can be powerful motivators (Cialdini et al., 1990; Bicchieri, 2006, 2016; Bicchieri and Dimant, 2019). In some cases, however, there may be uncertainty as to which norms apply to a specific situation, or even whether a norm that might apply is followed at all in a given context. For instance, even if there is a norm against lying, some lies might be considered innocuous ("white lies") and therefore permissible, while others are clearly not (Erat and Gneezy, 2012). Between these two extremes is a grey area characterized by uncertainty about how (un-)acceptable lying might be and what behavior is common in a given situation.[1] When people are uncertain, they form beliefs by drawing upon similar experiences or shared cultural narratives (Shiller, 2017; Bénabou et al., 2018a), which is susceptible to self-serving distortion. Recent experimental studies have shown that, in ambiguous situations, people may choose to entertain beliefs that justify evading costly pro-social behavior (Babcock et al., 1995; Di Tella et al., 2015; Exley, 2015; Gneezy et al., 2018c).[2] The literature refers to these as "motivated beliefs distortions" (Bénabou, 2015; Bénabou and Tirole, 2016; Gino et al., 2016).

We connect these streams of literature and investigate belief distortion at the intersection of social norms and deviant behavior (lying behavior in particular). For this novel context, we employ a variant of the 'dice under the cup' paradigm (Shalvi et al., 2011; Fischbacher and Föllmi-Heusi, 2013). This task (and variations of it) has been shown to predict rule violations in natural settings, and has been used as a proxy to measure the prevalence of rule violations across societies (Cohn et al., 2015; Gächter and Schulz, 2016; Hanna and Wang, 2017; Cohn and Maréchal, 2018; Cohn et al., 2019).

We have two primary goals, which will improve our understanding of the underlying

[1] Lying is defined as asserting something that is false with the intention to mislead (Isenberg, 1968, see also Sobel, 2019).

[2] Also see Chen and Gesche (2017); Ging-Jehli et al. (2019). For cases in which risk and/or ambiguity are arguably absent see Exley and Kessler (2018); Dimant et al. (2019).

drivers of lying versus honest behavior that have not yet been investigated in the existing literature. First, we want to answer whether self-serving belief distortion aimed at evading costly pro-social behavior, which has been identified in other domains, also applies to lying.

Second is a more conceptually important departure from the existing literature. Because different forms of beliefs are central to the concept of social norms, we aim to identify which of those are more commonly distorted to justify lying behavior. When choosing their actions, people form beliefs about what others do/have done in the same situation (empirical expectations), and what others approve of (normative expectations) (Bicchieri, 2006, Bicchieri and Dimant, 2019). Since individuals engage in motivated belief distortion with the precise intent of justifying their choices, it is important to understand the kind of belief distortion that yields behavioral change. This is vital from a policy perspective and informs more effective reforms and behavioral interventions, e.g. nudges (Benartzi et al., 2017).[3]

We employ a two-stage variant of the 'dice-under-the-cup' paradigm, where subjects' beliefs are elicited in stage 1 before subjects perform a dice task in stage 2.[4] In the dice task, reporting a "5" yields a positive monetary payoff while reporting any other number yields a zero payoff. In stage 1, we employ an incentive-compatible belief elicitation protocol, in which we randomly elicit subjects' beliefs on either the honest/dishonest behavior adopted by the majority of subjects in a previous session, or the approval of honest/dishonest behavior by the majority of previous participants. Consistent with the aforementioned literature, the belief elicitation process induces subjects to retrieve, process, and interpret information from past experiences and shared narratives, in order to form their beliefs. We utilize a host of treatments to identify belief distortion and introduce a theoretical model to derive precise behavioral predictions. Our pivotal experimental intervention is to randomly vary whether participants are aware or unaware of an upcoming opportunity to lie about their beliefs. We hypothesize that, when subjects are aware of a lying opportunity, the belief formation process may become subject to distortion aimed at justifying lying in the subsequent task. By contrast, distortion will not occur when subjects form their beliefs

[3] An advantage of using subjects' own belief distortion to study the effect of different kinds of beliefs on behavior is that it may avoid some of the limitations that arise with alternative belief-shifting methods. Arguably, providing subjects with information about the pervasiveness of lying or other deviant behavior (e.g., Bicchieri et al., 2019b; Dimant, 2019) in previous sessions may incur the problem that these information sets may not contain enough variation to generate different behavioral responses. These limitations are absent (at least in principle) in the case of subjects' own belief distortion.

[4] While our main examination focuses on self-serving belief distortion, in the appendix we also look at belief manipulation in an other-regarding domain as a robustness test of our theoretical examination.

before finding out that they will soon engage in the dice task. In this case, beliefs will be unbiased. We compare these two treatments to test whether belief distortion occurs and, if it does, to identify which beliefs are more apt to be distorted, and how belief distortion affects behavior. We also consider a baseline treatment in which belief elicitation is omitted.

To inform our empirical investigation, we build a theoretical model that borrows from Bicchieri's (2006) theory of social norms: in order to exist, a social norm against lying requires the following conditions: a) people expect that most individuals in their reference network do not lie (empirical expectation), b) people expect that most individuals in their reference network disapprove of lying (normative expectation), and c) people have a preference for not lying conditional on these expectations.[5] Having a conditional preference implies that if our expectations change, then we might stop following the norm (at least temporarily); e.g., we might realize that a sizable number of people violate it, or that transgressions are no longer disapproved of or punished, which might lessen the norm's influence. Intuitively, believing that lying is common – or that most people do not disapprove of it – rules out that an honesty norm is followed in this context. Then, a conditional norm-follower would find it easier to lie, since the condition for obeying the norm is not in effect (for recent experimental evidence, see Bicchieri et al., 2019a).

Our theoretical model aggregates different streams of literature into an optimal beliefs model in the tradition of Brunnermeier and Parker (2005); Bénabou and Tirole (2006a); Bénabou (2015); Bénabou and Tirole (2016). We assume that individuals belong to one of three types. The first two types – **Unconditional Honest** (UH) and **Unconditional Liars** (UL) – are (socially) unconditional, in the sense that their behavior in a given situation does not depend on what they believe others do and/or deem appropriate. These unconditional types are introduced because they are argumentatively reasonable and also for technical reasons as they allow to evade the issue of multiple equilibria, though their role is limited. The third type, **Conditional Liars** (CL), is the focus of our analysis. These individuals are conditional norm-followers, in the sense that they refrain from lying if they believe that most people are honest *and* disapprove of lying, but will lie if they believe that either of these conditions are not met. We assume the exact share of Unconditional Honest and Unconditional Liars is a random, unobserved variable (the state of the world). This generates uncertainty about the dominant norm and, for Conditional Liars, creates

[5] By reference network we mean the people that matter to the decision maker in a specific situation (Bicchieri, 2016). In our case, these are other participants in the same task.

an opportunity to engage in self-serving belief distortion to facilitate lying in the dice task.

Our empirical results are multi-faceted and consistent with our theoretical examination. In what follows, we present an abbreviated outlook of the most important findings. Our first finding is that, in the empirical treatment, subjects engage in belief distortion: when aware of the upcoming lying task, they are more likely to believe that lying is widespread and are more likely to lie, than when their beliefs are formed *before* they become aware of the lying task ahead, as expected. Conversely, in the normative treatment, we find that the subjects' elicited beliefs are independent of whether or not they are aware of the upcoming lying opportunity, which suggests that belief distortion does *not* occur. Lying rates are indistinguishable in both cases, and are as high as they are in the empirical treatment with belief distortion. Our interpretation is that, in the normative treatment, self-serving belief distortion is *not* necessary to induce lying, since believing that the majority disapproves of lying does not inhibit one's own lying. This stands in contrast with the empirical treatment, where we find that belief distortion, in the form of convincing oneself that lying is widespread, is needed to facilitate lying in the subsequent task.

These findings are consistent with our norm-based model under the condition that there is an asymmetry between what we can infer from empirical as opposed to normative information: 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 does not hold. Widespread disapproval of lying is not necessarily a strong indicator that most people behave honestly. Even if most people express disapproval of lying, a norm against lying may not be followed.

To confirm this hypothesis, we ran a follow-up experiment in which new subjects chose the statement they most agreed with after being provided information about either the lying behavior or the approval of lying behavior of participants in a previous dice task experiment. We compared two conditions in which we essentially provided one truthful part of the social norm information (empirical or normative) to infer the beliefs about the other part (normative or empirical): in the first condition, subjects were informed that the majority of individuals in the dice experiment refrained from lying, and were asked to guess how many subjects in that experiment disapproved of lying. In the second condition, subjects were informed that the majority of individuals in the dice experiment disapproved of lying, and were asked to guess how many subjects in that experiment refrained from lying. Our findings support the hypothesis that, while people interpret honest behavior as a strong indicator of disapproval of lying, disapproval of lying is not seen as a strong

guarantee of honesty; 'walking the talk' is a more costly signal and thus more predictive than cheap talk of normative (dis)approval.

Coming back to our main experiment, when comparing all treatments, we find that the empirical treatment where beliefs are elicited *before* subjects become aware of the upcoming dice task (and are thus unbiased) yields the following insight: there, reports of the winning number are substantially *lower* than in all other treatments, where lying rates are essentially equivalent. Our interpretation is that, once subjects have formed unbiased beliefs, this prevents them from engaging in further belief distortion, even when they are eventually confronted with a lucrative lying opportunity in the form of the dice task. In that treatment, subjects are consequently less prone to misreport the winning number than in all other treatments.[6] This suggests that belief distortion occurs during the process of belief formation, and not subsequently, when beliefs are already formed and one would need a good reason for belief change (e.g., new compelling evidence). This finding is consistent with literature on "belief stickiness," which shows that, once people form their beliefs, they find it hard to change them even when doing so would be materially beneficial (e.g., Falk and Zimmermann, 2017; Gneezy et al., 2018b,c).

Finally, when looking at behavior in the baseline condition in which belief elicitation is omitted and subjects are immediately engaged in the dice task, we find that lying rates are the same as when beliefs are elicited and subjects are aware of the lying task ahead. This suggests that taking part in the experimental belief elicitation task is not necessary for subjects to engage in belief distortion. Self-serving belief distortion may also occur when subjects are not explicitly asked by the experimenter to state their beliefs.

The remainder of the paper is organized as follows. In Section 2, we describe our contribution to the literature and present the experimental design in Section 3. To derive testable hypotheses, we develop a novel theoretical model of belief distortion in Section 4. The results are presented in Section 5. In Section 6, we discuss a number of possible alternative models, including the possibility that belief distortion costs may depend on the nature of the belief to be distorted, the the theory of moral dissonance as formalized by Rabin (1994), image motivations (Bénabou and Tirole, 2006b; Gneezy et al., 2018a; Abeler et al., 2018), and pure conformity. Section 7 concludes.

[6] This occurs for empirical beliefs since, as explained, in that treatment (but not in the normative treatment) conditional liars have to engage in self-serving belief distortion to justify lying in the dice task.

2. Contribution to the Literature

Our paper contributes to the growing literature on self-serving belief distortion, a concept extensively documented in the domain of overconfidence – see e.g. Zimmermann (2019) for an important recent contribution. This literature focuses on distortions of beliefs about one's own ability, and typically abstracts from the implications of belief distortion for the subject's subsequent behavior. The exception is Schwardmann and van der Weele (2019), who document that people may strategically choose to hold overconfident beliefs in order to become more persuasive in subsequent interactions.

The notion that self-serving strategic considerations may be primary motivators of belief distortion is present in a number of recent works. Di Tella et al. (2015), for example, show that people may distort their beliefs about the actions of their opponents in order to justify taking money from them, while in Exley (2015) beliefs are distorted in order to donate less to charity. Gneezy et al. (2018b,c) document similar instances of self-serving belief distortion that justify evading costly pro-social behavior. These are all examples of what may be called *forward looking* belief distortions, aimed at changing the subjects' subsequent actions in a way that improves their material interests.[7]

This paper looks at this type of belief distortion within the context of lying decisions. An important novelty of our investigation is in the nature of the beliefs we focus on – empirical and normative expectations – which have not been considered by previous literature, and are natural in the context of lying. Our rich empirical strategy allows us to uncover which beliefs are subject to self-serving distortion, and consequently to identify the kind of beliefs that inhibit or facilitate lying. This has clear relevance for policy design aimed at mitigating rule-violating behavior. To augment our contribution, we present a novel theoretical model of belief distortion that builds upon related streams of literature (Bénabou and Tirole, 2006a; Bicchieri, 2006) and effectively explains our empirical findings. In an extensive discussion section, we also argue that our results are not easily reconcilable with a number of alternative theoretical frameworks presented in the literature.

On a conceptual level, our paper also contributes to the vast literature on lying (see Abeler et al. 2018 and Gerlach et al. 2019 for exhaustive meta-analyses). Recent empirical and theoretical contributions to this literature (e.g., Abeler et al., 2018; Gneezy et al.,

[7] Another branch of the literature (such as Saucet and Villeval, 2019) focuses instead on *backward looking* belief distortion, that documents selective recalls of past actions.

2018a) have highlighted that image or reputation concerns explain a number of stylized facts not easily explained by other models. We view these findings as complementary to our analysis (see Section 6 for further discussion).[8]

Our contribution with respect to the lying literature is to show how, and why, self-serving belief distortion can affect lying behavior and which types of norm-related beliefs play a crucial role in facilitating such behavior. We present a theoretical and behavioral rationale for why empirical expectations (what others do) may play a more important role in giving (or not giving) "licence" to lie than normative expectations (what others approve of). An ongoing discussion in existing literature (e.g., Eriksson et al., 2015) concerns the extent to which learning about common behavior leads to inferences about what is commonly approved of. We hypothesize that, if the majority behaves honestly, then this is a strong indication that the majority also disapproves of lying – thus an honesty norm exists and is followed. Conversely, if the majority disapproves (or says that they disapprove) of lying, this does not necessarily imply that most people behave accordingly. It is likely that a norm of honesty exists but is not followed in spite of widespread disapproval (see also Bicchieri et al., 2019b). 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 inferred normative or empirical expectations.

Finally, our model also adds to the theoretical work on norms (for a discussion of game-theoretic accounts of social norms, see Bicchieri and Sontuoso, 2018). Much of the behavioral literature assumes that the norm against lying has been internalized so that people become desensitized to what others do or approve of (Bicchieri, 2006). In contrast, we focus on the behavior of conditional norm-followers, highlighting how these individuals

[8] One of the experiments conducted by Abeler et al. (2018) employed different anchors (in the form of the description of a "potential" experiment with very high/low incidence of winning reports) to shift the subjects' beliefs about lying shares. Although the anchors were successful in shifting the subjects' beliefs, the study finds no effect on subsequent lying behavior. This stands in contrast with our finding that belief distortion is associated with higher lying rates. The null result by Abeler et al. (2018) is consistent with an environment in which opposing considerations (for instance, norm-following and reputation for honesty, further discussed in Section 6) cancel each other out. A distinctive feature of our study is that, differently from their experiment, it was conducted online, and thus involved no direct human exchange. This might have substantially reduced or even eliminated image-based considerations (see e.g., Cohn et al., 2018; Bolton et al., 2019) Consistent with this is the finding by Dimant et al. (2019) showing that explicit norm-nudge interventions exhibited no effect in altering lying behavior using an online setting and design in which subjective and objective risk was absent by design.

may want to distort own beliefs to affect their subsequent behavior. Thus, our model follows the tradition of the game theoretic literature on strategic information revelation in the context of norms – e.g., Bénabou and Tirole, 2006a; Sliwka, 2007; Adriani and Sonderegger, 2009, 2018; Adriani et al., 2018 – and applies its insights to belief distortion.

3. Experimental Design

3.1. General Procedure

Our data collection is subdivided into several self-serving experiments – results of which are presented in the main body of this paper – and several other-regarding experiments – the results of which presented in the Appendix and intended as robustness checks of the theory. We present the aggregate data collection below and a detailed breakdown of observations per treatment at the bottom of the respective figures in the result sections. For the data analyzed in the main body of the text, we capitalize on:

- Pre-experimental data collection to obtain the truthful empirical and normative information for subsequent incentive-compatible belief elicitation → 100 data points
- Data collection for the main belief-distortion and lying experiment (Baseline plus all normative and empirical CPK and CPU treatments) → 724 data points
- An additional experiment to examine specific theory predictions → 300 data points

Nested in a non-interactive setting of our study, all data was collected on Amazon Mechanical Turk (MTurk), which is particularly well suited for our design nd has been similarly used in recent research (Kuziemko et al., 2015; DellaVigna and Pope, 2017). Literature related to our research agenda has successfully used variants of cheating paradigms and norm-nudge interventions on MTurk (Bolton et al., 2019; Bicchieri et al., 2019a). Importantly, recent literature indicates the robustness, generalizability, and reproducibility of laboratory findings (Arechar et al., 2018; Coppock et al., 2018; Snowberg and Yariv, 2019). To 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%. We used online tools to test IPs for low quality respondents.

Any residual noise would be uncorrelated with the treatments and work against us.[9] 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 participants earned on average \$0.95 equivalent to an average hourly payoff of \$9.50 (including a show-up fee of \$0.5, equivalent to \$5 per hour) and is well above average MTurk pay (Hara et al., 2018).[10]

3.2. Treatments

We introduce two between-design dimensions along which our experiment varies:

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

We refer to the treatment combinations as depicted in Figure 1.

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.[11]

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, majoritarian behavior (empirical treatment) or majoritarian normative convictions (normative treatment) in a previous session.

Upon signing a consent form, reading the instructions (see Appendix for screenshots), and passing a number of comprehension questions, participants were first presented with

[9] To achieve sufficient statistical power, our data collection was informed by a pre-test also used in Dimant et al. (2019) (for details, see pre-registration documents on AsPredicted.org #23244 and #23283). At least 110 observations per treatment arm had to be collected to achieve a moderate effect size of approximately 0.39, power of 0.8, and an alpha of 0.05. To account for noise in the data that may render some data points invalid, we aimed to collect about 125 usable data points per treatment.

[10] Noteworthy, 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).

[11] The different parts of the experiment had no names to avoid potential priming.

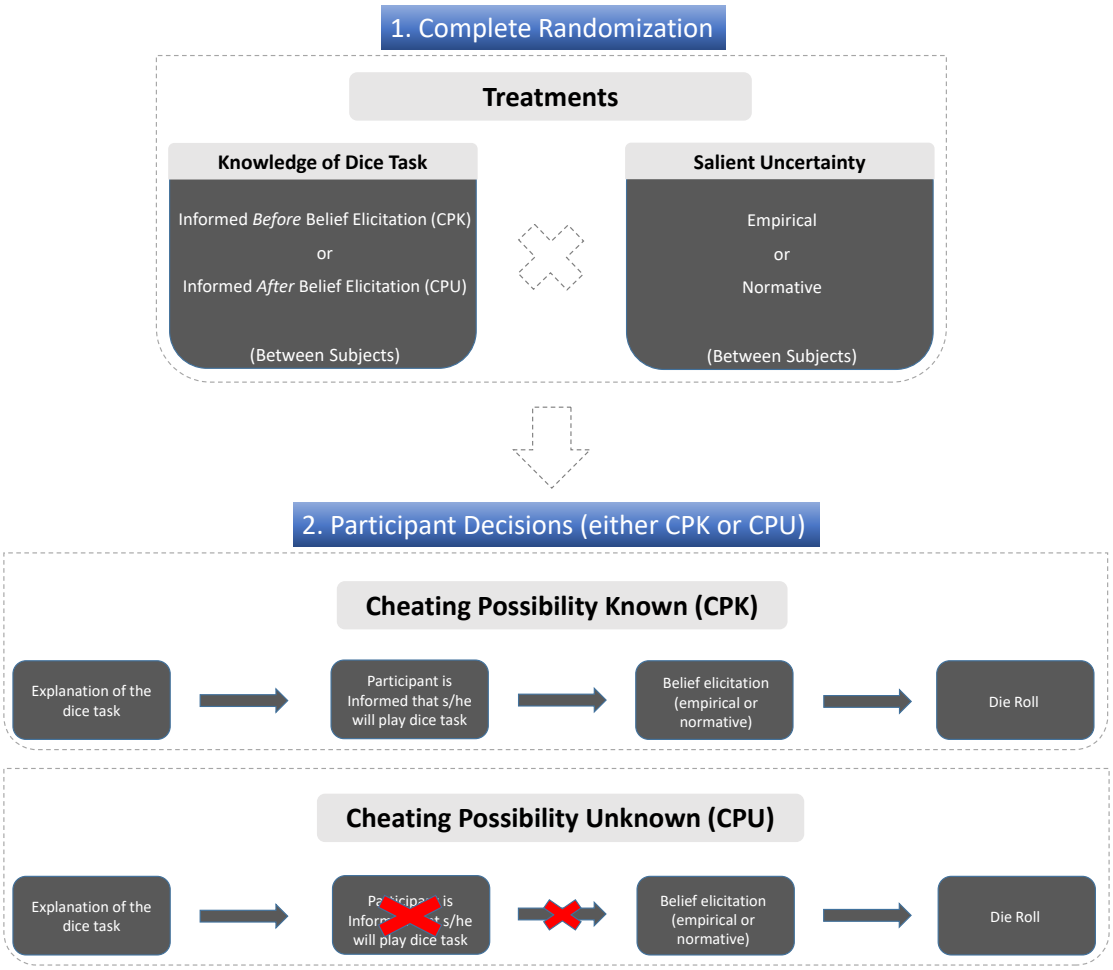


Figure 1: Experimental procedure.

the incentivized belief-elicitation task. The belief task asked participants to indicate which of the 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 the information 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, par-

ticipants had to choose one statement. A correct answer increased the participants' payoff by \$0.25 (equivalent to \$2.5 per hour). 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 dice task. The statements presented to the participants varied across treatments (see Figure 1) and the exact wording is presented below:

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 of providing new information on behavior; 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 (as partially addressed by Dimant et al., 2019). Instead, we are interested in how people may distort their beliefs in a self-interested manner. Aside from the Baseline, in all treatments the participants' beliefs regarding only one aspect (empirical or normative) that is instrumental in determining whether a norm of honesty applies or not were elicited.

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.

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 whether belief distortion occurs and its influence on subsequent behavior.

Importantly, to make treatments comparable, participants were always explained the mechanics of the dice task at the *beginning* of all treatments, i.e. *before* the belief elicitation phase. This ensured that participants knew which task the presented empirical and normative information referred to. What varied between the CPK and CPU conditions was the explicit mention of whether the participants themselves would engage in the task after the belief elicitation. 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 in which participants only played the dice task without having been asked to express their belief in a previous stage.

4. Theoretical Examination

In this section, we present a simple theoretical model of belief distortion in the spirit of Bénabou and Tirole (2006a, 2011, 2016) which, as we will show, goes a long way in rationalizing our empirical findings. We start by building a general setup that models belief distortion abstracting from the precise reasons underpinning the conditionality of

subjects' behavior on their expectations. We then add more structure, proposing a norm-based theoretical account of why expectations may matter for subjects' behavior. All our results are proved in the text unless otherwise specified.

Conditional and unconditional individuals

We consider a setup where individuals belong to one of three types: (i) Unconditional Liars (UL) – always lie if this generates a positive monetary return for them. (ii) Unconditional Honest (UH) – incur a prohibitively high cost from lying and, as a result, do not lie. (iii) Conditional Liars (CL) – may choose to lie or not depending on what they believe about the underlying state of the world. Differently from the other types, conditional liars may have an interest in distorting their beliefs in order to affect their future behavior.

Let the share of UL be denoted as α_{UL} and the share of UH as α_{UH}; the share of CL is $1-\alpha_{UL}-\alpha_{UH}$. The precise values of α_{UH} and α_{UL} are not perfectly observed and depend on the specific nature of the situation at hand.[12] There are two possibilities:

- with probability $q \in (0,1)$: $\alpha_{UH} = h$ and $\alpha_{UL} = l$;
- with probability $1-q$: $\alpha_{UH} = l$ and $\alpha_{UL} = h$,

where $h > 1/2 > l$ and $1-h-l > 0$. Our assumptions ensure that when the share of unconditional truth-tellers is high, the majority of subjects tell the truth (independently of what CL do), while when the share of unconditional truth-tellers is low, the majority of subjects lie.

States of the world

A state of the world is defined as a pair (a, m), where a indicates majoritarian behavior and m indicates the normative convictions held by the majority. We assume that all individuals belonging to the same unconditional type hold the same normative convictions. With probability $g \in (0,1]$ all UH types disapprove of lying, while with remaining probability $1-g$ they believe lying is morally acceptable. Similarly, the probability that individuals

[12] The underlying idea is that the specific nature of the situation at hand will influence 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 behavior of these individuals is not motivated by their beliefs about the state of the world. 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 CL.

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 states of the world, 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 normative conviction that lying is wrong (right).

Prior probability	Majoritarian behavior	Majoritarian normative conviction	State
qg	$a = T$	$m = W$	state H1
$q(1-g)$	$a = T$	$m = R$	state H2
$(1-q)\,r$	$a = L$	$m = W$	state L1
$(1-q)\,(1-r)$	$a = L$	$m = R$	state L2

(1)

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 unacceptable (*Normative treatment).*

Belief about majoritarian behavior elicited in the empirical treatment is denoted as $b_a \in \{b_T, b_L\}$, while belief about majoritarian normative conviction elicited in the normative treatment is $b_m \in \{b_R, b_W\}$. To answer the belief elicitation question, subjects must 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 CL subjects decide whether or not to lie.[13]

[13]For simplicity, the monetary incentives present in the belief elicitation task are ignored in this stylized model (and explicitly accounted for in the robustness section presented in the appendix)

Lying decision

Since UL agents always lie and UH never do, the decision to lie or not only concerns CL. In what follows, we focus exclusively on lies that generate a positive monetary return for the subject, and take the form of dishonestly reporting the winning number in order to obtain the monetary transfer $\mu > 0$.[14] Let u_{S} denote the non-monetary (dis-)utility from lying for a CL type when the state is S. Since subjects cannot observe the state, they choose their behavior by computing the expected return from lying. The expected non-monetary component of utility from lying is,

$$E\left[u(b)\right] = u_{\mathrm{H1}} \Pr(\mathrm{H1} \mid b) + u_{\mathrm{H2}} \Pr(\mathrm{H2} \mid b) + u_{\mathrm{L1}} \Pr(\mathrm{L1} \mid b) + u_{\mathrm{L2}} \Pr(\mathrm{L2} \mid b)$$

where $\Pr(\mathrm{S} \mid b)$ is the probability that the subject assigns to state S given belief b (more on this below). At t=1 conditional liars choose to lie if,

$$E\left[u(b)\right] + \mu > 0 \tag{2}$$

where the left hand side of (2) is the total expected return from the lie, and tell the truth otherwise.

Consistency requirement

At $t = 1$, CL's decision to lie or not depends on their beliefs about the underlying state of the world, as described in (2). We impose the following consistency requirement. Given b_a or b_m formed at $t = 0$, beliefs about the likelihood of state S occurring are derived from Bayes' rule, as follows:

$$\Pr(\mathrm{S} \mid b_a) = \frac{\Pr(a \mid \mathrm{S})\ \Pr(\mathrm{S})}{\Pr(a)} \text{ and } \Pr(\mathrm{S} \mid b_m) = \frac{\Pr(m \mid \mathrm{S})\ \Pr(\mathrm{S})}{\Pr(m)}]$$

This requirement imposes restrictions on the beliefs that subjects may hold about the state of the world.[15] If, for instance, g is large, then this implies that a subject who has formed the belief that the majority tell the truth at $t = 0$ also needs to acknowledge that the

[14] As will become clear below, subjects will never choose to distort beliefs to engage in lying behavior that generate zero or negative monetary returns (such as reporting a losing number when drawing the winning number). These types of lies are therefore ignored here.

[15] For convenience, we assume that subjects are unaware that their beliefs b_a or b_m formed at $t = 0$ may have been distorted, although the model can be generalized to account for this possibility.

majority are likely to disapprove of lying, and, consequently, that the state of the world is likely H1.[16] More precisely, in the empirical treatment, the consistency requirement implies that $t = 1$ beliefs satisfy: $\Pr(\text{L1}|\ b_T) = \Pr(\text{L2}|\ b_T) = \Pr(\text{H1}|\ b_L) = \Pr(\text{H2}|\ b_L) = 0$ and

$$\begin{cases} \Pr(\text{H1} \mid b_T) = g \\ \Pr(\text{H2} \mid b_T) = 1 - \Pr(\text{H1} \mid b_T) \\ \Pr(\text{L1} \mid b_L) = 1 - r \\ \Pr(\text{L2} \mid b_L) = 1 - \Pr(\text{L1} \mid b_L) \end{cases}$$

In the normative treatment, $t = 1$ beliefs satisfy: $\Pr(\text{H2}|\ b_W) = \Pr(\text{L2}|\ b_W) = \Pr(\text{H1}|\ b_R) = \Pr(\text{L1}|\ b_R) = 0$ and

$$\begin{cases} \Pr(\text{H1} \mid b_W) = \frac{qg}{(1-q)r+qg} \\ \Pr(\text{L1} \mid b_W) = 1 - \Pr(\text{H1} \mid b_W) \\ \Pr(\text{H2} \mid b_R) = \frac{q(1-g)}{q(1-g)+(1-q)(1-r)} \\ \Pr(\text{L2} \mid b_R) = 1 - \Pr(\text{H2} \mid b_R) \end{cases}$$

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 different states of the world 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 belief distortion. 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"). Distortion 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

[16] We follow previous literature such as Brunnermeier and Parker (2005); Oster et al. (2013) and assume that updating from belief b follows Bayes rule. Recent literature has highlighted that, in a number of settings (e.g., overconfidence) people stray away from the Bayesian benchmark (Zimmermann, 2019; Harrison and Swarthout, 2019). Clearly enough, our theoretical predictions would continue to apply for more general updating processes so long as we impose some minimal consistency requirements, such as, e.g., that if the majority tell the truth then it is likely that the majority also disapprove of lying.

follows this approach and assumes that conditional agents behave as motivated Bayesians.[17] Similar to Di Tella et al. (2015), we do not model the belief distortion process explicitly, and instead adopt a black-box approach (see Bénabou and Tirole, 2016 for a discussion of possible underlying distortion processes). Formally, belief distortion occurs if and only if (i) the subject is aware of the cheating task that will follow at t=1 and (ii) belief distortion advances the subject's material interests, in the sense that,

$$\mu \cdot (\|_{b \text{ distorted}} - \|_{b_{\text{not distorted}}}) > 0 \tag{3}$$

where $b_{\text{not distorted}}$ and $b_{\text{distorted}}$ indicate unbiased and biased beliefs, and $\|_b$ is an indicator function that takes value 1 if, given belief b at $t = 0$, condition (2) holds – implying that, at $t = 1$, the agent lies – and takes value 0 otherwise.[18] Hence, at t=0, conditional agents choose their beliefs by focusing exclusively on material payoff. This is in line with existing models of belief distortion that build on the notion of a "split self" with partially conflicting interests.[19]

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 normative convictions (in the normative treatment). During the belief formation phase, subjects may engage in self-serving belief distortion (if they are aware of the lying task ahead).
- t=1 Lying/truth-telling dice task. Conditional subjects compute the expected total return from lying and from telling the truth, and choose their action. Payoffs are realized.

Equilibrium

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

[17] UL and UH types do not engage in belief distortion since their behavior in the lying task is independent of their expectations.

[18] Our setup thus implicitly assumes that beliefs distortion involves no direct costs. This is a parsimonious benchmark, but is not essential. Allowing for a belief distortion cost $c > 0$ would leave our results unchanged, with the exception of Propositions 0 and 1, which would have to be recast as "In the baseline and the CPK conditions, conditional agents lie whenever their return from doing do exceeds the belief distortion cost".

[19] Section 6 discusses the alternative possibility that belief distortion may be aimed at improving the subject's psychological well-being.

Bayesian equilibrium in the game between his date-0 and date-1 selves) for conditional individuals satisfies the following conditions:

1. At t=1, subjects compute the probabilities of different states of the world by Bayesian updating from their $t = 0$ belief b. They lie if expression (2) holds, and tell the truth otherwise.

2. At t=0, subjects (when aware of the cheating task that will follow) choose to engage in belief distortion iff, taking $t = 1$ behavior as given, condition (3) above holds.

Baseline condition

For benchmark purposes, we first describe what the model predicts for the baseline case, in which the belief elicitation stage $t = 0$ is omitted. This implies that, at $t = 1$, when subjects have to form beliefs about the state of the world in order to decide how to behave, they are free to direct this process in a self-serving manner, without the "straitjacket" imposed by the the need to be consistent with previously formed beliefs. As a result, we expect that conditional subjects will lie.

Proposition 0 (Baseline condition): *In the baseline, conditional liars always lie.*

Cheating Possibility Known (CPK) Condition

We now consider 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, conditional liars may thus find it optimal to distort their beliefs, in order to modify their actions at t=1. As described in (8), our model predicts that the only case in which belief distortion will *not* occur is when distortion is not necessary to justify, and hence induce, cheating in the dice task. There are thus two possibilities: (1) belief distortion at $t = 0$ is necessary to induce cheating at $t = 1$, in which case beliefs are distorted at $t = 0$, or (2) belief distortion is not necessary for cheating. Note that in both cases the final outcome is the same, namely lying at $t = 1$.

Proposition 1 (CPK condition): *In the CPK condition, conditional liars always lie in both the empirical and the normative treatment.*

This implies that the CPK condition will induce the same amount of lying as the baseline.

Cheating Possibility Unknown (CPU) Condition

Next, we investigate the CPU condition. Clearly enough, in this case subjects have no incentive to distort their beliefs, since they do not anticipate the future lying opportunity. At $t = 0$, they thus process information unbiasedly in both treatments. This implies that, different from the CPK and baseline conditions, in the CPU condition it is possible that conditional liars may end up not cheating at $t = 1$. This happens when the unbiased beliefs elicited at $t = 0$ are such that condition (2) does not hold. The implication is that,

Proposition 2 (CPU condition): *Lying rate in the CPU condition is either,*

(a) the same as in the CPK condition (if there is no belief distortion in CPK), or

(b) lower than in the CPK condition (if there is belief distortion in CPK).

Conditional Liars are Norm Followers

We now add more structure about the underlying motivations of conditional liars. The setup we consider is grounded in the work by Bicchieri (2006). We assume that conditional liars are norm followers, in the sense that they 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 one should not lie (*normative requirement*)

are majoritarian, and no lying cost otherwise. The psychological cost θ can be seen as arising from the cognitive dissonance generated from lying while knowing that lying is a violation of the norm (as defined by (a) and (b)). Importantly, this cost is contingent on the state of the world being H1, since this is the only state where a majority of individuals satisfy both the empirical (the majority doesn't lie) and normative (the majority thinks that one should not lie) requirements. A conditional subject who believes that the state is H1 with probability p will lie if,

$$\mu - \theta p \rightarrow \frac{\mu}{\theta} > p \tag{4}$$

and will tell the truth otherwise.[20] We assume that $\mu < \theta$, implying that a CL individual who believes the state to be H1 with certainty will prefer to tell the truth (if this was not

[20] We adopt the convention that, if indifferent, a conditional subject tells the truth and does not distort his own beliefs. This is immaterial for our results.

the case then conditional subjects would always lie independently of their beliefs about the state of the world).

Let $p(b)$ denote the probability that the subject ascribes to state H1 when his $t = 0$ belief is b. As we have seen, in the empirical treatment the choice of b affects beliefs about the state of the world as follows: $p(b_T) > 0$ and $p(b_L) = 0$; similarly, in the normative treatment: $p(b_W) > 0$ and $p(b_R) = 0$. From (4), a low posterior belief p makes it more likely that a conditional subject will lie at $t = 1$. If $p(b) = 0$, he will lie for sure. Given $p(b_L) = 0$ and $p(b_R) = 0$, this implies that the only type of distortion that may take place in the norm model must take the form of conditional subjects convincing themselves that (i) the majority lie (empirical treatment), or (ii) the majority considers lying to be acceptable (normative treatment). Distorting beliefs in the opposite direction (by convincing themselves that the majority doesn't lie or does not approve of lying) would not advance material interests, as it would induce the belief that state H1 is more likely, thus making the subjects less willing to lie at $t = 1$. This leads to,

Proposition 3 (Nature of belief distortion in the norm model): *In equilibrium, belief distortion may only take the form of: (i) inducing belief b_L instead of b_T (empirical treatment), or (ii) inducing belief b_R instead of b_W (normative treatment).*

Note that, while subjects can choose what to believe when engaging in belief distortion, their unbiased beliefs will reflect the true underlying state of the world. In what follows, we focus on the benchmark case where the underlying state is H1, so that, in the absence of belief distortion, a subject would come to the conclusion that the the majority of individuals tell the truth (empirical) or disapprove of lying (normative). In turn, by Proposition 3, this implies that conditional subjects might be tempted to engage in belief distortion, in order to avoid being honest in the subsequent dice task. As will become clear in the Empirical Results section, H1 was actually the true underlying state in our experiment, and is thus the relevant state to consider if we want to test the predictions of the theory with our experimental data.[21]

In the CPK condition, we know by Proposition 1 that conditional subjects always lie. Consider then the CPU condition. In that case, beliefs at $t = 0$ are not distorted, and are thus given by $b = b_T$ in the empirical treatment and $b = b_W$ in the normative treatment (since the underlying state is H1). Lying will occur at $t = 1$ if $p(b_T) < \frac{\mu}{\theta}$ in the empiri-

[21] Predictions obtained in the other possible states of the world are discussed in the Appendix.

cal treatment, and if $p(b_W) < \frac{\mu}{\theta}$ in the normative treatment. This leads to the following proposition,

Proposition 4 (Beliefs and behavior in the norm model): *Suppose that the underlying state is* H1. *If* $p(b_T) > p(b_W)$ *then: (i) Lying in normative CPU (weakly) exceeds lying in empirical CPU; (ii) Because of (i), the incentive to engage in belief distortiom in empirical CPK is (weakly) stronger than in normative CPK. The opposite predictions hold if* $p(b_T) < p(b_W)$.

Hypotheses

This Section describes the testable implications of the theory.

Belief distortion: *We say that belief distortion in the form of x occurs if the share of subjects reporting belief x in CPK is higher than in CPU.*

Our first two hypotheses follow from our basic premise that conditional subjects will, whenever possible, engage in belief distortion when this advances their material interests, by making it easier for them to lie at t=1.

Hypothesis 1: *In the CPK condition, the share of subjects who report the winning number in both the empirical and normative treatment is the same, and is equal to that in the baseline treatment.*

Hypothesis 1 follows from Propositions 0 and 1. Intuitively, in the CK and the baseline treatments, nothing prevents conditional subjects from engaging in belief distortion whenever needed. As a result, they always lie in the dice task.

Hypothesis 2: *If belief distortion occurs, the share of subjects reporting the winning number in CPU is lower than in CPK. If belief distortion does not occur, the share of reports of the winning number in CPU and CPK is the same.*

Hypothesis 2 follows from Proposition 2. Contrary to CPK, in the CPU treatment $t = 0$ beliefs are formed unbiasedly, and may therefore inhibit lying. The only case where CPK and CPU may deliver the same incidence of dishonest behavior is if belief distortion is actually *not needed* for conditional subjects to lie in the dice task.

Hypothesis 3: *If belief distortion occurs, it is in the form of:*

a) "the majority lie"(empirical treatment), or

b) "the majority approves of lying" (normative treatment).

Hypothesis 3 follows from Proposition 3. The next hypothesis deals with comparisons between normative and empirical treatments in the CPK and CPU conditions.

Hypothesis 4: *Suppose that the underlying state is* H1.

a) If $p(b_T) > p(b_W)$, the following must hold:

1. *In CPK, if belief distortion occurs in only one treatment, then it must be the empirical treatment.*

2. *In CPU, if the share of subjects reporting the winning number is higher in one treatment, then it must be the normative treatment.*

b) If $p(b_T) < p(b_W)$: in (i) it must be the normative treatment, while in (ii) it must be the empirical treatment.

Hypothesis 4 arises from Proposition 4. Recall that $p(b_T)$ is the probability that a subject ascribes to the majority disapproving of lying when the majority tell the truth, while $p(b_W)$ is the probability ascribed to the majority telling the truth when the majority disapprove 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.

5. Empirical Results

Our analysis varies 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 dice task, any difference in belief distributions between these two treatments indicates active belief distortion.

We unpack our findings in multiple steps. First, we identify the "true" state of the world as revealed by the data, showing that a norm for honesty as defined in Bicchieri (2006) (state H1) applies to our setup. Second, since Hypothesis 4 is 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 Hypothesis 4. Third, we discuss the results from our main experiment, comparing beliefs and lying 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 (37% reported the winning number, which suggests a lying rate of approximately 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%. Thus, the true state of the world corresponded to H1, in which 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.[22] As before, participants were explained the setup of the original experiment and were then randomly allocated to one of two treatment variations. In each treatment, participants received empirical or normative information about a previously-run experiment, namely (1) empirical ("the majority did not lie") or (2) normative ("the majority did not approve of lying") and were then asked to guess whether (i) the majority of participants (dis)approved of lying (after being told 1) or whether (ii) the majority of participants did (not) lie (after being told 2). This approach allows us to elicit normative (posterior) expectations from empirical information and vice versa. The answer to (i) delivers an estimate of $p(b_T)$, namely the probability that a norm of 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 of honesty applies when the majority disapproves of lying. Importantly, participants were only presented with one statement and only provided one guess, which was incentivized based on our thruthfully collected data.

[22] n=300, 2 conditions in between-subjects design, \$0.50 show-up fee, 99% approval rate, U.S. residents. See Section IV in the Appendix for detailed instructions.

Figure 2 illustrates our findings and the results are clear: empirical information strongly affects the normative expectation, but not vice versa. When participants are told that the majority of participants *did not lie*, 77.48% infer that the majority 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*, only 47.65% infer that the majority is honest. Thus, these results indicate that, in our setup, inference is much stronger in one direction (empirical → normative) than in the other (normative → empirical). In concluding that $p(b_T) > p(b_W)$, we can restate hypothesis 4 as follows:

Hypothesis 4 *(i) In CPK, if belief distortion occurs in only one treatment, then it must be the empirical treatment. (ii) In CPU, if the share of subjects reporting the winning number is higher in one treatment, then it must be the normative treatment.*

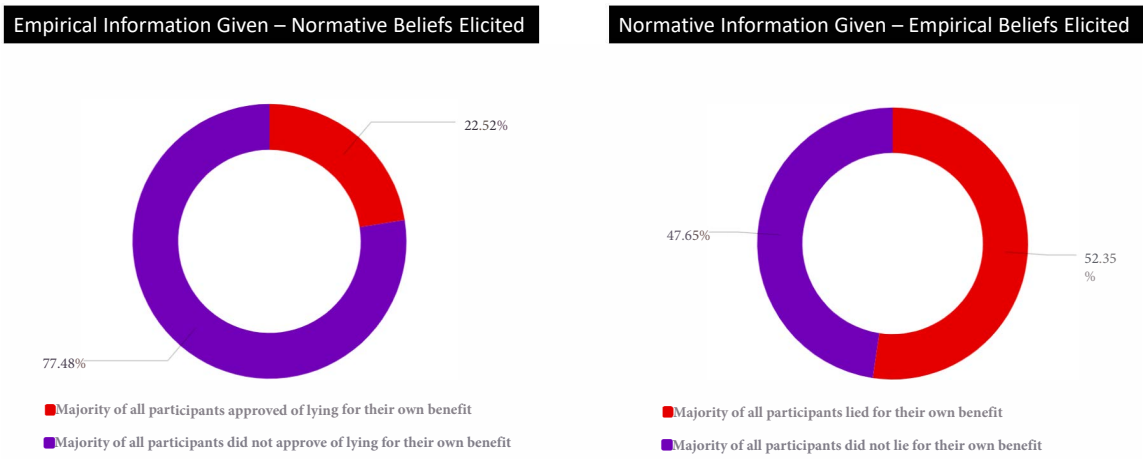


Figure 2: Follow up experiment. Left panel: participants were given empirical information (*Majority of subjects lied*) and were incentivized to guess the majoritarian normative conviction. Right panel: participants were given normative information (*Majority of subjects disapproved of lying*) and were incentivized to guess majoritarian behavior.

5.3. Main Experiment

Figure 3 summarizes our key results. First, in the CPK condition, we see that the share of reported winning numbers is the same in both the empirical and normative treatments (40.5% and 40.2%), and corresponds to what we observed in the baseline treatment (34.7%), which is comparable to the existing literature (Abeler et al., 2018; Gerlach et al., 2019). This confirms our Hypothesis 1. Intuitively, the theory predicts that, if needed, conditional

subjects will distort their beliefs in order to lie. As a result, in the CPK and the baseline treatments, they always lie. The only possible case where conditional subjects may end up *not* lying is in the CPU treatment, where beliefs are formed unbiasedly (since subjects are unaware of the lying task ahead). In that case, it is possible that the beliefs they have formed at $t = 0$ might inhibit cheating and, thus, conditional subjects might behave honestly in the dice task.

Second, in line with our motivation, we also investigate belief distortion. To infer whether self-serving belief distortion occurred, we compare beliefs in the CPK and CPU conditions (top part of Figure 3). We start from the empirical treatment. As can be seen in Figure 3, the belief that "the majority lies" increases considerably as we move from CPU to CPK (38.1% versus 62.7%, Equality of Proportions Test (EPT), $p<0.001$). This result indicates that subjects distorted their beliefs when they were alerted about the upcoming opportunity to lie. Moreover, in accordance with Hypothesis 2, belief distortion resulted in a greater incidence of winning reports in CPK compared to CPU (40.5% versus 21.4%, EPT, $p<0.001$), confirming that subjects distorted their beliefs in order to facilitate lying. For both findings, a post-hoc estimation reveals power in excess of 95%.

In the normative treatment, on the other hand, beliefs remained statistically indistinguishable in CPK and CPU (39.4% and 34.9%), from which we infer that belief distortion did not occur. Again in accordance with Hypothesis 2, winning reports also remained statistically indistinguishable, (40.2% in CPK versus 38.0% in CPU), indicating that the reason why beliefs were not distorted was that this was not needed for conditional subjects to cheat in the subsequent task, since believing that "the majority think lying is not OK" did not inhibit their own lying. We have confidence in these null-findings due to the tight confidence intervals that we observe yielding no economic or theoretically meaningful effect size (based on the previously discussed literature) to fall within its bounds. In contrast, the results obtained in the empirical treatment indicate that, when they believed that the "majority does not lie", conditional subjects chose to behave honestly, and this substantially reduced cheating in the empirical CPU treatment, where, by design, the subjects' elicited beliefs were not distorted.

Turning to Hypothesis 3, the previous discussion has highlighted that, where it occurred (namely, in the empirical treatment), belief distortion followed the pattern predicted by our norm model, namely people convincing themselves that lying is widespread. Further support for the model is provided by the fact that, given that belief distortion occurred

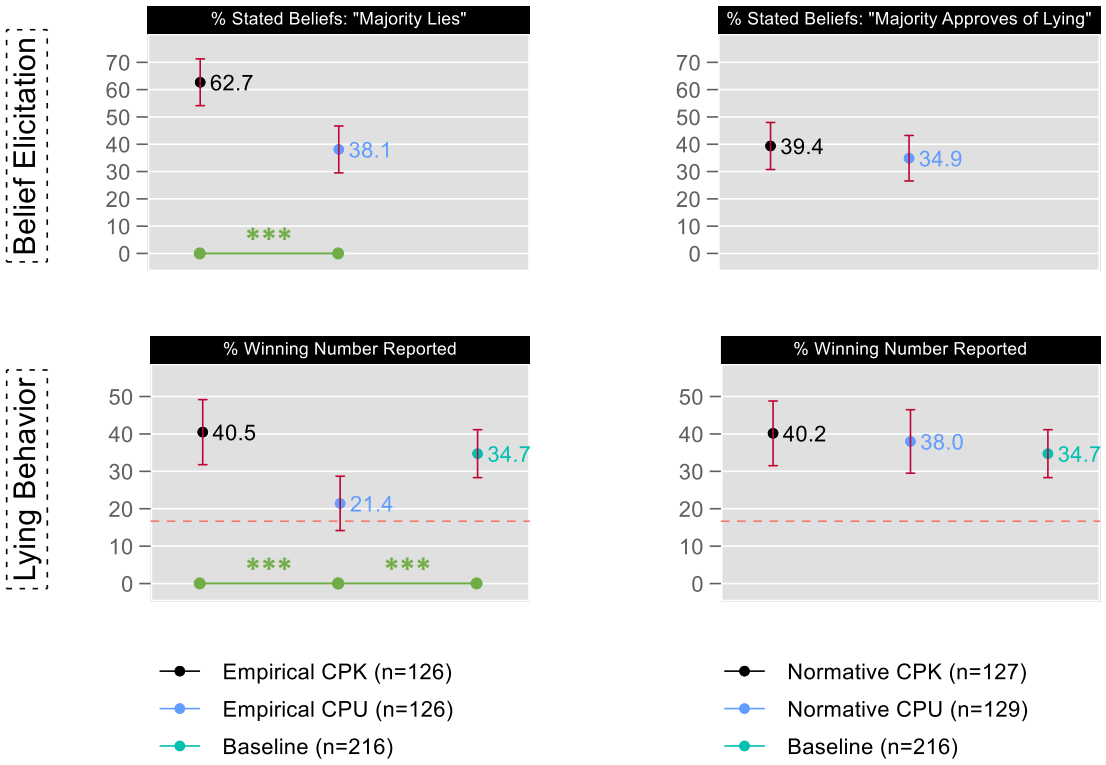


Figure 3: Results from belief elicitation (top panels) and lying frequency (bottom panels) broken down for the Empirical (left panels) and Normative (right panels) conditions and include the same Baseline. Dotted red line illustrates the expected value of 16.67%. Whiskers indicate 95% confidence intervals. 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 based on the method proposed by Benjamini and Hochberg (1995).

only in one treatment, this was the empirical treatment, as predicted by Hypothesis 4. Intuitively, this follows from the observation that, in our setup, inferences are much stronger in one direction (empirical → normative) than in the other (normative → empirical). In other words, if someone believes that the majority does not lie, then she also needs to acknowledge that a norm of truthtelling is very likely to apply; believing that the majority disapproves of lying is not inconsistent with believing that dishonesty is widespread.

We substantiate our previous results with a Logit regression analysis, examining both stated beliefs and lying frequency (Odds Ratios reported). We ensure the robustness by including a battery of controls (age, gender, risk (SOEP), and Cognitive Reflection Test (CRT) score). The regressions fully confirm the previous results: when comparing our empirical CPU and CPK conditions, both lying frequency and the reported beliefs are

significantly different. Conversely, and in line with our previous analysis and theoretical model, neither is true when comparing the normative CPU and CPK conditions. We also provide robustness checks for Hypotheses 1 and 2 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. This is what our results show and are displayed on the right-hand side of Panel B of Table 1.

Table 1: Logit Regression (Odds Ratios) Analysis of Reported Beliefs and Lying Frequency

Panel A *DV: Reported Beliefs*	(1)	(2)	(3)	(4)
Treatment *(Baseline: CPK Condition)*				
Empirical CPU *(1 = Majority Lies)*	0.366*** (0.095)	0.388*** (0.105)		
Normative CPU *(1 = Majority Does Not Approve of Lying)*			0.825 (0.214)	0.779 (0.214)
Controls	No	Yes	No	Yes
Observations	252	252	256	255

Panel B *DV: Lying Behavior*	(1)	(2)	(3)	(4)	Empirical CPK = Normative CPK *(Hypothesis 2)*		Empirical CPU < Normative CPU *(Hypothesis 4)*	
Treatment *(Baseline: CPK Condition)*								
Empirical CPU *(1 = Majority Lies)*	0.401*** (0.114)	0.425*** (0.122)			0.987 (0.253)	1.031 (0.266)	2.246*** (0.637)	2.089** (0.610)
Normative CPU *(1 = Majority Does Not Approve of Lying)*			0.913 (0.234)	0.837 (0.226)				
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	252	252	256	256	253	253	255	255

Pannel A and Panel B: Logit regressions (odds ratios reported) with robust standard errors clustered at the individual level displayed in parentheses. 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.[23] 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 empirical CPU (31.3% vs. 15.4%,

[23]Because we run pairwise comparisons, we take into account that such multiple comparisons cause an inflation of type-I-error. To counteract this inflation, we employ the false discovery rate correction by Benjamini and Hochberg (1995), which has superior features compared to the Bonferroni correction.

p=0.0351) but, in line with our model, the *difference* in the propensity to report the winning number for each belief is substantially larger in CPK (50.6%-23.4% = 27.2%) than in CPU (31.3%-15.4% = 15.9%). Intuitively, in the CPK treatment, subjects purposefully distort their beliefs in order to lie in the subsequent task. This strengthens the correlation between lying and holding the belief that the majority of people lie.

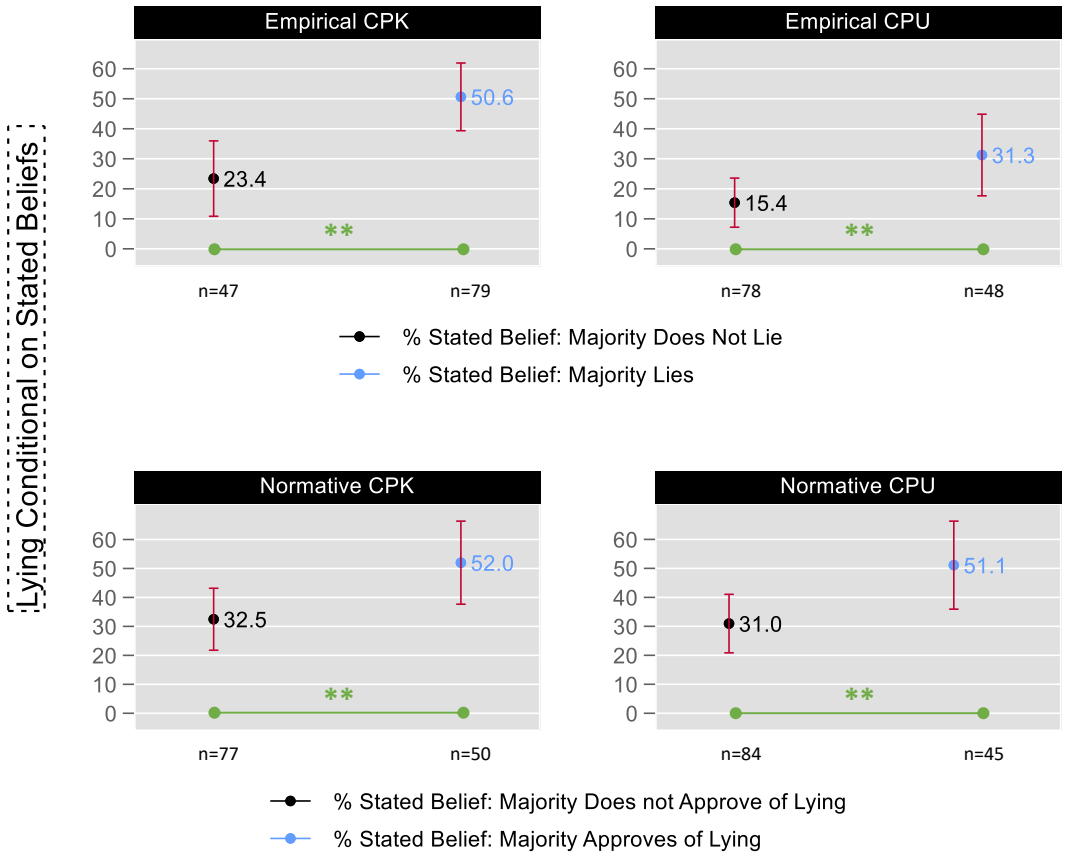


Figure 4: Results from conditional lying behavior for the self-serving conditions (individual is beneficiary of the lie) and are broken down for the Empirical (top panels) and Normative (bottom panels) conditions. Whiskers indicate 95% confidence intervals. Stars indicate significant differences at the conventional levels of *p<0.1, **p<0.05, and ***p<0.01 and are false discovery rate corrected based on the method proposed by Benjamini and Hochberg (1995).

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.[24] The key statistic for our purposes, however, is again the extent to which the predictive power of beliefs for lying frequency changes as we move from 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%-31.0% = 20%). This fits our theoretical account since, as we have seen, in the normative treatment, conditional subjects did not engage in belief distortion. As a result, there is no reason to expect different correlations between beliefs and lying frequency in CPK and CPU.[25]

6. Discussion

Following our theoretical and empirical examination, we deem it important to provide the reader with a comprehensive discussion of extensions of our model and alternative interpretations of our findings. We will use both empirical and theoretical arguments to strengthen the core of our findings and position our contribution within the existing economic research. We begin by discussing an extension of our setup: the possibility that the cost of distorting beliefs about the normative convictions of others (normative expectations) may differ from the cost of distorting beliefs about what others do (empirical expectations). Second, we discuss the possibility that belief distortion may be motivated by the desire to enhance the subject's psychological (rather than material) well-being. Lastly, we address alternative theoretical models of conditional subjects' motivation.

Differing Costs of Belief Distortion

Our stylized model assumes that the cost of distorting beliefs is constant (and equal to zero for simplicity). However, it is possible to envision a more general model in which the cost of distortion depends on the nature of the belief to be distorted. This opens up the possibility that our experimental finding of belief distortion not occurring in the normative treatment may be due to normative expectations being harder to distort than empirical expectations. Intuitively, this might be the case if the ex-ante probability that the majority considers lying acceptable is smaller than the ex-ante probability that the majority might

[24] While 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.

[25] 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.

actually lie. This would happen in the presence of a well-established norm – e.g. fairness or reciprocity – where it is hard to believe that a majority would approve of unfairness or lack of reciprocity. Formally, suppose that the cost of self-servingly inducing belief b_i, $i = a, m$ by means of ad hoc information processing is a decreasing function of the prior probability that i may actually occur. In our setup, the ex-ante probability that $a = L$ is $1 - q$, while the ex-ante probability that $m = R$ is $q(1 - g) + (1 - q)(1 - r)$. If the latter is smaller than the former, then the cost of inducing belief b_R instead of b_W will be larger than the cost of inducing b_L instead of b_T. Intuitively, if it is ex-ante very unlikely that in any given situation the majority of individuals might consider lying acceptable, then it seems plausible that subjects might find it harder to find arguments to convince themselves that this is the case when it actually is not.

Although this is a potentially interesting extension of our model, we do not think that our experimental findings in the normative treatment should be taken to indicate that normative expectations are harder to distort in our setup. The key observation here is that the incidence of lying in the normative treatment (whether in CPK, CPU or the baseline condition) is *the same* as in empirical CPK (and higher than empirical CPU). In other words, in the normative treatment subjects lie as much as they do when engaging in belief distortion in the empirical treatment. As discussed, our interpretation is that distorting normative expectations is not needed to induce lying. Hence, in the normative treatment, belief distortion will not occur even if the cost of distorting normative beliefs is negligible (as in our main model). Our theoretical analysis explains why this is the case.

Belief Distortion is Motivated by Psychological Considerations

Our model assumes that belief distortion is entirely motivated by material payoff: at $t = 0$, a subject distorts his beliefs *only if* this is necessary to induce him to cheat at $t = 1$. This rules our an alternative hypothesis, namely that a subject who would lie *anyway* may engage in belief distortion in order to feel less guilty about lying.

In our experiment, this would generate a sensible difference between CPK and CPU in terms of beliefs, but no difference in terms of behavior. However, this is not the pattern we observe. In our data, CPK and CPU either deliver the same beliefs and the same behavior (normative treatment), or differ in both beliefs and behavior (empirical treatment). In fact, in the empirical treatment, the difference between the share of subjects who believe that the majority lie in CPK and CPU (62.7-38.1=24.6) is almost identical to the difference in reported winning numbers between the two conditions (40.5-21.4=19.1). This suggests

that the overwhelming majority of those who distort their beliefs in empirical CPK do so to change their behavior in the lying task.

Generic Cognitive Dissonance and Moral Dissonance per Rabin (1994)

We now discuss possible alternative motivations for Conditional Liars, and their implications for our experiment. In his classic monograph, Festinger (1962) defines cognitive dissonance as the mental discomfort (psychological stress) experienced by a person who holds contradictory beliefs. To reduce this dissonance, individuals may engage in self-serving belief distortion. Our model can be seen as a special case of this general principle, in the sense that CL subjects engage in belief distortion in order to avoid the psychological cost θ, incurred when subjects lie while simultaneously believing that a norm of honesty applies – and could arise from difficulty considering themselves "moral individuals."[26]

This raises the question whether our findings could be explained by a more general model where, for CL subjects, cognitive dissonance costs arise when their behavior conflicts with the (empirical or normative) expectations that they hold about other individuals. Our results in the normative treatment show that this is not the case: subjects do not feel the need to engage in belief distortion, and the belief that "the majority disapproves of lying" does not deter lying. This rules out that CL subjects experience a (non-negligible) psychological discomfort when lying while holding that belief.

Our findings in the normative treatment also rule out that the Rabin (1994) theory of moral dissonance might apply in this setup. In Rabin's model, people suffer a disutility (dissonance cost) when engaging in activities they believe to be immoral – a phenomenon he calls "moral dissonance."[27] The theory assumes that, if society considers a behavior (say, lying) immoral, this makes it harder for a liar to avoid incurring the dissonance cost. In our setup, this implies that, in the normative treatment, CL subjects will want to convince themselves that the majority approve of lying, in order to facilitate lying in the subsequent dice task. Belief distortion may also occur in the empirical treatment, but only to the extent to which behavior informs what the majority deems to be morally acceptable. Hence, if belief distortion occurs in only one treatment, the theory of moral dissonance predicts that it should be the normative treatment. This is in fact the opposite

[26] Shalvi et al. (2015); Buckenmaier et al. (2019); Turmunkh et al. (2019) discuss some of the techniques or justifications that people employ to alleviate the tension between engaging in selfish behavior and seeing themselves as moral individuals. See also Pittarello et al. (2015); Klein et al. (2017).

[27] Barkan et al. (2012) use the term "ethical dissonance".

of what we actually observe.

Relevant Dichotomy is not Majority/Minority

Recall that in our model, conditional liars prefer not lying if they believe that a) most people do not lie (empirical expectation) and b) most people disapprove of lying (normative expectation) (Bicchieri, 2006). Although the partition between majoritarian and minoritarian behavior and beliefs is natural, it is conceivable that the relevant threshold might in fact differ from 50% and, if true, challenge our empirical analysis. Suppose that conditional liars suffer a disutility from lying if they believe that at least 75% of people do not lie and disapprove of lying. In this case, belief distortion aimed at facilitating lying in the dice task could take the form of CL subjects convincing themselves in the CPK condition that the share of truthtellers is, say, 60% rather than 80%. Since both these shares are majoritarian, this distortion would not be picked up by our belief elicitation task (which focuses on the majority/minority dichotomy).

The resulting empirical pattern would then take the form of observing higher lying rates in CPK compared to CPU (reflecting the fact that, in CPK, belief distortion did take place), while at the same time observing no difference in elicited beliefs. This is however not what we see in our data. In the empirical treatment, higher lying rates in CPK correspond to a higher share of subjects believing that the majority lies. In the normative treatment, lying rates are the same in CPK and CPU, and this is matched by correspondingly similar elicited beliefs in both treatments.

Reputation for Honesty (With no Under-Reporting) and Other Image Models

Recent contributions to the lying literature – such as Abeler et al. (2018); Gneezy et al. (2018a) – have shown that reputation and image concerns play an important role in explaining lying behavior. The desire to appear honest (or to signal high lying costs) rationalizes a number of findings that cannot be explained by other models, such as that, when lies come in different "sizes," a share of those who choose to lie typically avoids the maximal report – choosing "smaller" lies instead.[28]

Thus, it is important to understand the extent to which our empirical findings are consistent with reputation models proposed in the literature. It should be noted that our

[28] Consider for example the experiment by Gneezy et al. (2018a), in which participants draw a number from 1 to 10, and payment is equal to the number reported. Although the report that generates the largest monetary return ("maximal report") is 10, a significant share of subjects choose to lie by reporting 8 or 9. See also Abeler et al. (2014).

experiment was conducted online with no direct human interaction, an environment where image concerns have been shown to be minimal (see, e.g., Cohn et al., 2018; Bolton et al., 2019; Dimant et al., 2019).[29] In that sense, we expect that, in our setup, these types of concerns played a smaller role than in other environments. We start by presenting a setup that is inspired by the model of social image by Gneezy et al. (2018a) – see also Khalmetski and Sliwka (2017); Dufwenberg and Dufwenberg (2018). To make the key effects easier to grasp, we will supplement the discussion with some formal analysis.

Suppose that conditional subjects derive utility from their *social image*, which is defined (as in Abeler et al. 2018; Gneezy et al. 2018a and Khalmetski and Sliwka 2017) as the probability that their report is interpreted as being honest by an outside observer. The underlying idea is that, for image conscious agents, being viewed as honest is an intrinsically valued part of their social identity. The utility of a conditional type is given by,

$$\begin{cases} \delta \text{ if he reports the losing number} \\ \mu + \delta\rho \text{ if he reports the winning number} \end{cases} \tag{5}$$

where $\delta > 0$ is the weight given to social image and $\rho \equiv \Pr(\text{honest report} \mid \text{report} = \text{winning \#})$ is the social image that derives from reporting the winning number, and corresponds to the assessment by an external observer of the likelihood that the report is true. The social image from reporting a losing number is normalized to 1. This rules out under-reporting since, even if an observer believes that a subject who reports the losing number had in fact drawn the winning number, it is not penalized, in the sense that it does not lower the subject's social image.[30] In what follows we assume $\delta > \mu$, implying that a conditional agent will not report the winning number if the social image associated with it is sufficiently low. A subject who draws the losing number will lie and dishonestly report the winning number if,

$$\rho > 1 - \frac{\mu}{\delta}$$

[29] See also Arechar et al. (2018); Coppock et al. (2018); Snowberg and Yariv (2019) for a discussion of the robustness of behavior on MTurk versus the laboratory.

[30] In contrast, if a subject is thought to be reporting the winning number in spite of drawing the losing number in order to obtain a financial gain, this negatively affects his social image. The absence of under-reporting is in line with available evidence. For instance, in the experiment by Gneezy et al. (2018a) (where, in one of the treatments, the experimenter could ex-post verify the number observed by each subject) only 1 participant out of 602 chose to under-report.

and will report truthfully otherwise. We follow Bénabou and Tirole (2011) and assume that, when assessing the likelihood that a report may or may not be truthful, the external observer knows the underlying state. This implies that the social image from reporting the winning number is derived from Bayesian updating from the true state (as well as from the equilibrium behavior of CL types). Denoting as ρ^S the social image from reporting the winning number in state S, in the Appendix we prove that, in equilibrium,

Lemma 1: $\min\left\{\rho^{H1}, \rho^{H2}\right\} > \max\left\{\rho^{L1}, \rho^{L2}\right\}$.

Proof: in Appendix.

Intuitively, when there are many liars misreporting the winning number (state L1 or L2), the social image from reporting the winning number is low, since it is likely to be a lie. Conversely, when it is known that majority of people tell the truth (state H1 or H2), a winning number report is likely to be truthful, and thus yields a higher social image. Image conscious agents are thus more inclined to lie in the latter case than in the former.

Consider now belief distortion. We start by analyzing the CPK condition. Recall that belief distortion will occur if it induces the subject to lie in a situation where, in the absence of distortion, she wouldn't lie. Consider the empirical treatment. Clearly enough, Pr(H1 or H2 | b_T) = 1 and, similarly, Pr(L1 or L2 | b_L) = 1. Given Lemma 1, this implies that, at t=1, the returns from lying for a conditional type when her belief is $b = b_T$ are higher than when $b = b_L$. Hence, in contrast with the norm-based model, in the social image model the only type of belief distortion 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 will interpret a winning report as likely to be honest. Instead, in situations where most agents lie, a winning number report is interpreted as likely to be dishonest, and thus results in a low social image. The incentive for an image-conscious individual to report the winning number when he believes that most individuals lie is therefore lower than when he believes that most agents tell the truth.

We now turn to the normative treatment. We prove in the Appendix that, in equilibrium, the following holds:

Lemma 2: *If $g > r$, the only form of belief distortion that may occur in the normative treatment takes the form of inducing b_W instead of b_R .*

Proof: in Appendix.

Although the relationship between g and r is, in principle, an empirical question, we

conjecture that $g > r$, and accordingly assume this to be the case in what follows.[31]

Proposition 5: *In equilibrium, belief distortion may only take the form of: (i) inducing belief b_T instead of b_L (empirical treatment), or (ii) (under the assumption that $g > r$) inducing belief b_W instead of b_R (normative treatment).*

We focus on the case where the underlying state is H1, i.e. the majority tell the truth and believe lying to be wrong (since this is the state that turned out to apply in our experiment). From Proposition 5, in this case image motivated conditional subjects do not need to engage in belief distortion in order to lie, implying that,

Proposition 6: *Suppose that the underlying state is* H1*. Then, (i) belief distortion does not occur, and (ii) CL subjects lie in all treatments: empirical CPK, normative CPK, empirical CPU and normative CPU.*

This shows that a model à la Gneezy et al. (2018a), where conditional subjects are motivated by the desire to appear honest (and where underreporting is ruled out), would deliver qualitatively different predictions than our norm-based model.[32] The predictions of Propositions 5 and 6 are are not borne out in our data, which suggests that our data cannot be explained by this type of image concerns.

There are other image based models, however, that could yield predictions in line with our findings. Consider a model along the lines of Bénabou and Tirole (2006b, 2011) where, rather than depending on the probability that a subject's report is honest/dishonest, social image depends on an outside observer's beliefs about the subject's *lying costs* (see also Adriani and Sonderegger, 2019). This implies, for instance, that if in equilibrium types with relatively low lying costs tell the truth by reporting a losing number (as may occur for instance if these inherently "dishonest" types are very concerned with their reputation), truthtelling will yield a lower social image than if it only involved high lying costs subjects. Under appropriate assumptions about the underlying type distributions, a model of this kind could explain our data by predicting that, in order to lie more easily, subjects should convince themselves that lying is widespread (or that the majority approve of lying). At the same time, under different assumptions, this model could make the opposite empirical

[31] Recall that g is the probability that Unconditionally Honest agents believe that lying is wrong, while r is the equivalent probability for Unconditional Liars.

[32] The above analysis considers a stylized model, but the results hold generally, as proved e.g. by Abeler et al. (2018).

predictions, which depend on features difficult to empirically verify. Thus, the "reputation for lying costs" model is harder to falsify than the norm-based model we proposed.[33]

Stated Beliefs as a Signaling Device

Our analysis assumes that the subjects' elicited beliefs reflect their true beliefs, distorted or not. However, our previous discussion about self-image concerns raises the possibility (at least in principle) that subjects may strategically select their stated beliefs, with the aim of producing a positive impression in an external observer.[34] For instance, it is conceivable that if a subject who reports the winning number (thus raising the suspicion that she may be cheating) also reports a belief that the majority lie or that the majority considers lying acceptable, this reported belief may be seen by an observer as a valid excuse for (possible) cheating behavior. Modeling this scenario requires a careful explanation of why the stated beliefs may help improve the subject's image.[35]

More to the point, it is unlikely that this might have occurred in our experiment. Again, the key observation here comes from the normative treatment, where the subjects' stated beliefs are the same in CPK and CPU conditions. This casts doubt on the possibility that the stated beliefs are used strategically by our subjects. Intuitively, if this were the case, then in the normative treatment we would expect the share of subjects reporting that they believe that the majority finds lying to be acceptable to increase in CPK (where subjects are aware of the lying opportunity ahead) compared to CPU (where they are not aware of it). However, this is not what we observe. To explain this evidence, a theory of stated beliefs as signaling devices should therefore explain why the subjects' incentives to strategically distort reported beliefs might differ between empirical and normative treatments.

Pure Conformity

Finally, we note that our experimental results are also consistent with a pure-conformity explanation of lying behavior, in which conditional subjects simply care about conforming with what they believe the majority does, without concerning themselves with normative considerations. In this light, lying or honest behavior are simply dictated by the dominant convention, similar to left- or right-hand driving (Young, 1996; Bicchieri, 2006).

[33] The same applies to a reputation for honesty model where under-reporting occurs with positive probability in equilibrium, see Abeler et al. (2018).

[34] This is documented by Andreoni and Sanchez (2019) within the context of a trust game.

[35] The notion that individuals may make use of excuses in a strategic manner, in order to order to influence the image they project is formally modeled by Bénabou et al. (2018b).

Although theoretically possible, we think that this explanation is implausible in our setting. In most societies, lying behavior is the object of normative prescriptions, with strong moral connotations, as exemplified by the "thou shalt not lie" biblical prescription or Kant's imperative to never lie. At the same time, there is also a shared understanding that, in some cases, lying may be justifiable – examples include lying to save lives or lying to confound an enemy; more generally, "white lies" or lies that do not harm anyone are often discounted (e.g., Erat and Gneezy, 2012). Indeed, in a number of environments, people are unsure about the extent to which others find lying reprehensible. In sum, the notion that people are concerned with what others approve of when choosing how to behave is intuitive and well documented in many domains (e.g., Bicchieri et al., 2019a; Dimant, 2019, for an overview of this vast literature in the domains of economics, psychology, and philosophy, see Bicchieri et al., 2017), and lying behavior certainly seems to belong to this category. The psychological foundation of such conditionality lies in the notion that humans naturally strive to obtain approval and avoid disapproval from others (e.g., Sugden, 1998; Bicchieri, 2006, 2016). This notion is also consistent with the tradition of conditionality in social preferences, such as cooperation in economic research (e.g., Fehr and Gächter, 2000, see also Fehr and Schurtenberger, 2018 for its relationship with norms).

7. Concluding Remarks

Norm violation often provides a material benefit to the transgressor. If we can convince ourselves that a norm does not apply to our situation, or is not presently followed, then we have reason to disregard it. Often it is not fully clear what other people do or have done in the same situation, or whether they approve of specific behaviors (for a recent overview, see Fehr and Schurtenberger, 2018). Uncertainty about what the norm is or whether it is presently followed can be solved in a self-serving way if we provide reasonable justifications for norm-violation (Bicchieri and Chavez, 2013). For example, evidence consistent with the desired behavior often receives preferential treatment (Kunda, 1990). 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 which 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, and as we show here, in situations where we either think a norm is not presently followed or we are unclear what the norm is, we will often find justifications for selfish behavior.

Our experimental design sheds light on these mechanisms and varies both the nature of belief elicitation (empirical: what others do, versus normative: 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 shortly be faced with the dice task, while in the "Cheating Possibility Unknown" (CPU) condition, participants learn about the dice task only after the belief elicitation has taken place. Our working assumption is that, if individuals are aware of the forthcoming lying opportunity at the time they form their beliefs, they may engage in self-serving belief manipulation to facilitate lying in the subsequent task. We augment the empirical exercise with a theoretical model of belief distortion in the spirit of Bénabou and Tirole (2006a, 2011, 2016) connected to the model of social norms in the spirit of Bicchieri (2006), from which we derive clear and testable predictions.

Our experiment yields a number of interesting results consistent with our theory. In the empirical treatment, we find convincing evidence for belief distortion; individuals choose to believe that most people lie because this facilitates their own lying. This, in turn, leads to higher cheating rates overall, supporting the notion that the reason for individuals to distort their beliefs is to enhance their material payoff by cheating more. Conversely, when belief elicitation concerns normative elements, we do not observe belief distortion, and beliefs as well as actual lying rates remain statistically invariant between CPK and CPU. 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 distort one's beliefs and lie.

From a policy perspective, our results raise awareness that environments in which norms are not clearly defined can facilitate individually and socially detrimental behavior. We recommend sending unambiguous information about common behavior in particular environments. If the common behavior is negative, the best option would be to identify subgroups that behave in a positive way and broadcast their behavior. As for 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 may be be a combination of congruent empirical and normative information, as positive and unambiguous

as possible (see Schultz et al., 2007). Finally, our analysis shows that once people form a belief, it tends to stick, even when it would be profitable to change it. In our experiment, subjects with previously formed unbiased empirical beliefs were less likely to lie in the subsequent dice task. This endorses early interventions aimed at shaping individual views and expectations on matters of common interest. Once crystallized, these views are harder to distort in a self-serving way, which plays an important role in inhibiting deviant behavior.

References

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, 113:96–104.

Abeler, J., Nosenzo, D., and Raymond, C. (2018). Preferences for truth-telling. *Econometrica*.

Adriani, F., Matheson, J., and Sonderegger, S. (2018). Why do parents socialize their children to behave pro-socially? an information-based theory. *Journal of Economic Behavior and Organization*, 145:511–529.

Adriani, F. and Sonderegger, S. (2009). Why do parents socialize their children to behave pro-socially? an information-based theory. *Journal of Public Economics*, 93:1119–1124.

Adriani, F. and Sonderegger, S. (2018). Signaling about norms: Socialization under strategic uncertainty. *Scandinavian Journal of Economics*, 120:685–716.

Adriani, F. and Sonderegger, S. (2019). A theory of esteem based peer pressure. *Games and Economic Behavior*, 115:314–335.

Andersen, D. J. and Lau, R. R. (2018). Pay rates and subject performance in social science experiments using crowdsourced online samples. *Journal of Experimental Political Science*, pages 1–13.

Andreoni, J. and Sanchez, A. (2019). Fooling myself or fooling observers? avoiding social pressures by manipulating perceptions of deservingness of others. *Economic Inquiry*.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1):99–131.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5):1337–1343.

Barkan, R., Ayal, S., Gino, F., and Ariely, D. (2012). The pot calling the kettle black: Distancing response to ethical dissonance. *Journal of Experimental Psychology: General*, 141(4):757.

Bénabou, R. (2015). Conférence jean-jacques laffont the economics of motivated beliefs. *Revue d'économie politique*, pages 665–685.

Bénabou, R., Falk, A., and Tirole, J. (2018a). Narratives, imperatives, and moral reasoning. Working Paper, National Bureau of Economic Research.

Bénabou, R., Falk, A., and Tirole, J. (2018b). Narratives, imperatives, and moral reasoning. Working Paper 24798, National Bureau of Economic Research.

Bénabou, R. and Tirole, J. (2006a). Belief in a just world and redistributive politics. *The Quarterly Journal of Economics*, 121(2):699–746.

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

Bénabou, R. and Tirole, J. (2011). Laws and norms. Working Paper, National Bureau of Economic Research.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological science*, 28(8):1041–1055.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C. and Chavez, A. K. (2013). Norm manipulation, norm evasion: experimental evidence. *Economics & Philosophy*, 29(2):175–198.

Bicchieri, C. and Dimant, E. (2019). Nudging with Care: The Risks and Benefits of Social Information. *Public Choice*.

Bicchieri, C., Dimant, E., Gaechter, S., and Nosenzo, D. (2019a). Observability, social proximity, and the erosion of norm compliance. Working Paper Available at SSRN: https://ssrn.com/abstract=3355028.

Bicchieri, C., Dimant, E., and Xiao, E. (2019b). Deviant or wrong? the effects of norm information on the efficacy of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294371.

Bicchieri, C., Muldoon, R., and Sontuoso, A. (2017). Social norms. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Bicchieri, C. and Sontuoso, A. (2018). Game-theoretic accounts of social norms: The role of normative expectations. In M. Capra, R. Croson, T. R. and Rigdon, M., editors, *The Handbook of Experimental Game Theory*. Edward Elgar Publishing.

Bolton, G., Dimant, E., and Schmidt, U. (2019). When a Nudge Backfires: Using Observation with Social and Economic Incentives to Promote Pro-Social Behavior. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294375.

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.

Buckenmaier, J., Dimant, E., Posten, A.-C., and Schmidt, U. (2019). Efficient institutions and effective deterrence: On timing and uncertainty of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3300563.

Chen, Z. C. and Gesche, T. (2017). Persistent bias in advice-giving. *Working Paper*.

Cialdini, R. B., Reno, R. R., and Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, 58(6):1015.

Cohn, A., Gesche, T., and Maréchal, M. A. (2018). Honesty in the digital age. *Working Paper*.

Cohn, A. and Maréchal, M. A. (2018). Laboratory measure of cheating predicts school misconduct. *The Economic Journal*, 128(615):2743–2754.

Cohn, A., Maréchal, M. A., and Noll, T. (2015). Bad boys: How criminal identity salience affects rule violation. *The Review of Economic Studies*, 82(4):1289–1308.

Cohn, A., Maréchal, M. A., Tannenbaum, D., and Zünd, C. L. (2019). Civic honesty around the globe. *Science*, page eaau8712.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446.

DellaVigna, S. and Pope, D. (2017). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–42.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Dimant, E., Gerben, A. v. K., and Shalvi, S. (2019). Requiem for a nudge: Framing effects in nudging honesty. Working Paper Available at SSRN: https://ssrn.com/abstract=3416399.

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

Erat, S. and Gneezy, U. (2012). White lies. *Management Science*, 58(4):723–733.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Exley, C. and Kessler, J. B. (2018). Motivated errors. *Working Paper*.

Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Falk, A. and Zimmermann, F. (2017). Information processing and commitment. *The Economic Journal*, 128(613):1983–2002.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives*, 14(3):159–181.

Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2:458–486.

Festinger, L. (1962). *A theory of cognitive dissonance*, volume 2. Stanford university press.

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

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496.

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

Ging-Jehli, N. R., Schneider, F., and Weber, R. A. (2019). On self-serving strategic beliefs. *University of Zurich, Department of Economics, Working Paper*, (315).

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018a). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and van Veldhuizen, R. (2018b). Bribing the self. Working Paper.

Gneezy, U., Saccardo, S., and van Veldhuizen, R. (2018c). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association*, 17(3):917–946.

Hanna, R. and Wang, S.-Y. (2017). Dishonesty and selection into public service: Evidence from india. *American Economic Journal: Economic Policy*, 9(3):262–90.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM.

Harrison, G. and Swarthout, J. T. (2019). Belief distribution, overconfidence, and bayes rule. Working Paper.

Isenberg, A. (1968). Deontology and the ethics of lying. In Thomson, J. and Dworkin, G., editors, *Ethics*, pages 163–185. Harper & Row, New York.

Khalmetski, K. and Sliwka, D. (2017). Disguising lies-image concerns and partial lying in cheating games. *Working Paper*.

Klein, S. A., Thielmann, I., Hilbig, B. E., and Zettler, I. (2017). Between me and we: The importance of self-profit versus social justifiability for ethical decision making. *Judgment and Decision Making*, 12(6):563.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3):480.

Kuziemko, I., Norton, M. I., Saez, E., and Stantcheva, S. (2015). How elastic are preferences for redistribution? evidence from randomized survey experiments. *American Economic Review*, 105(4):1478–1508.

Levine, E. E. and Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106.

Oster, E., Shoulson, I., and Dorsey, E. (2013). Optimal expectations and limited medical testing: evidence from Huntington disease. *American Economic Review*, 103(2):804–30.

Pittarello, A., Leib, M., Gordon-Hecker, T., and Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological Science*, 26(6):794–804.

Rabin, M. (1994). Cognitive dissonance and social change. *Journal of Economic Behavior and Organization*, 23:177–194.

Saucet, C. and Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Schwardmann, P. and van der Weele, J. J. (2019). Deception and self-deception. *Nature Human Behaviour*.

Schweitzer, M. E. and Hsee, C. K. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *Journal of Risk and Uncertainty*, 25(2):185–201.

Shalvi, S., Gino, F., Barkan, R., and Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, 24(2):125–130.

Shalvi, S., Handgraaf, M. J., and De Dreu, C. K. (2011). Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management*, 22(s1).

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4):967–1004.

Sliwka, D. (2007). Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes. *American Economic Review*, 97(3):999–1012.

Snowberg, E. and Yariv, L. (2019). Testing the waters: Behavior across participant pools. Forthcoming Journal of Political Economy.

Sobel, J. (2019). Lying and deception in games. *Journal of Political Economy*.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Sugden, R. (1998). Normative expectations: the simultaneous evolution of institutions and norms. *Economics, values, and organization*, 73.

Turmunkh, U., van den Assem, M. J., and van Dolder, D. (2019). Malleable lies: Communication and cooperation in a high stakes tv game show. *Management Science*.

Young, H. P. (1996). The economics of convention. *Journal of Economic Perspectives*, 10(2):105–122.

Zimmermann, F. (2019). The dynamics of motivated beliefs. Forthcoming American Economic Review.

Appendix

I. Theoretical Appendix

I.1 Predictions in States of the World Other Than H1

In the main text we focused on deriving empirical predictions for state H1, which, as discussed, corresponds to the true underlying state in our setup. Here, for completeness, we include the predictions of our theoretical model in the other possible states of the world. For brevity, we focus on the predictions for the norm model. The predictions for the image model can be obtained in an analogous fashion. The first observation is that Hypotheses 1, 2 and 3 are independent of of the underlying state of the world and thus continue to hold straightforwardly. Consider then Hypothesis 4.

State H2

When this is the underlying state, belief manipulation will never occur in the normative treatment, while in the empirical treatment it may occur. Hypothesis 4 changes as follows: Predictions a) (i) and (ii) now hold independently of the nature of the relationship between $p(b_T)$ and $p(b_W)$.

State L1

When this is the underlying state, belief manipulation will never occur in the empirical treatment, while in the normative treatment it may occur. Hypothesis 4 changes as follows: Predictions b) (i) and (ii) now hold independently of the nature of the relationship between $p(b_T)$ and $p(b_W)$.

State L2

When this is the underlying state, belief manipulation will never occur in either the empirical or the normative treatment. The share of subjects who report the winning number is the same in both treatments (empirical/normative) and in both conditions (CPK/CPU).

I.2 Robustness: Monetary incentives at the belief elicitation stage

We now show how, in a slightly modified version of our main setup, CN subjects may engage in belief manipulation also when offered explicit monetary incentives at the belief elicitation stage, and even when these incentives are equal the monetary return from reporting a winning number, namely μ.

Consider a setup in which, with small but positive probability, $b_{\text{no manip}}$ does not reflect the true underlying state of the world. For instance, it is possible that one's own past

experiences might be somewhat biased, and may therefore lead a subject to draw incorrect inferences, even in the absence of belief manipulation. Formally, this implies that, for some $\pi > 1/2$, the following holds.

- When $a = T$ (resp. L), in the empirical treatment: $b_{\text{no manip}} = b_T$ (resp., b_L) with probability π, and $b_{\text{no manip}} = b_L$ (resp., b_T) with probability $1 - \pi$.
- When $m = W$ (resp. R), in the normative treatment: $b_{\text{no manip}} = b_W$ (resp., b_R) with probability π, and $b_{\text{no manip}} = b_R$ (resp., b_W) with probability $1 - \pi$.

Note that the model examined in the main text is a special case of this more general model, in which $\pi = 1$. It is straightforward to see that Bayesian updating then implies $\Pr(a \mid b_{\text{no manip}} = b_a) < 1$ and $\Pr(m \mid b_{\text{no manip}} = b_m) < 1$. In other words, even if the subject does not manipulate beliefs, the subject is not certain of the underlying majoritarian action or majoritarian moral conviction.

Consider now the choice to manipulate beliefs at $t = 0$. Suppose that, if manipulation occurs, the subject will lie at $t = 1$, while if manipulation does not occur, the subject will not lie at $t = 1$. It is then clear that, at $t = 0$, manipulation is in the subject's material interests. This is because, if the subject does not manipulate, the subject will earn μ with probability less than 1, as explained above. Conversely, if the subject manipulates (and thus reports the winning number with certainty at t=1), the subject will earn μ for sure.

I.3 Reputation for Honesty: Proofs of Lemmata 1 and 2

Proof of lemma 1: Let $\eta \in [0, 1]$ denote the probability that a type CL who draws a losing number dishonestly reports the winning number. By Bayesian updating, $\rho = \frac{\frac{1}{6}}{\frac{1}{6}+\frac{5}{6}[\alpha_{UL}+\eta(1-h-l)]}$, decreasing in η. Substituting for $\eta = 0$ (resp., $\eta = 1$) in this expression, we obtain an upper (lower) bound for ρ for a given value of α_{UL}, obtaining $\rho \in \left[\frac{1}{1+5[\alpha_{UL}+1-h-l]}, \frac{1}{1+5\alpha_{UL}}\right]$. As a result, $\min\{\rho^{\text{H1}}, \rho^{\text{H2}}\} \geq \frac{1}{1+5(1-h)}$ (since $\alpha_{UL} = l$ in those states), while $\max\{\rho^{\text{L1}}, \rho^{\text{L2}}\} \leq \frac{1}{1+5h}$ (since $\alpha_{UL} = h$). Since $h > 1/2$, $\frac{1}{1+5(1-h)} > \frac{1}{1+5h}$ and hence, $\min\{\rho^{\text{H1}}, \rho^{\text{H2}}\} > \max\{\rho^{\text{L1}}, \rho^{\text{L2}}\}$. □

Proof of lemma 2. We prove the lemma by showing that the opposite form of belief manipulation cannot emerge in equilibrium, as it would generate a contradiction. To this purpose, first note that, if $g > r$, then $\Pr(\text{H1}|\ b_W) > \Pr(\text{H2}|\ b_R)$. Suppose now that $E(\rho \mid b_R) > 1 - \frac{\mu}{\delta} > E(\rho \mid b_W)$, implying that CL subjects lie at $t = 1$ when $b = b_R$ but

not when $b = b_W$. Then, $\rho^{\text{H1}} = \frac{1}{1+5l}$ and $\rho^{\text{H2}} = \frac{1}{1+5(1-h)} < \rho^{\text{H1}}$, and similarly, $\rho^{\text{L1}} = \frac{1}{1+5h}$ and $\rho^{\text{L2}} = \frac{1}{1+5(1-l)} < \rho^{\text{L1}}$. In turn, this implies that

$$\Pr(\text{H1} \mid b_W)\rho^{\text{H1}} + (1 - \Pr(\text{H1} \mid b_W))\,\rho^{\text{L1}} > \Pr(\text{H2} \mid b_W)\rho^{\text{H2}} + (1 - \Pr(\text{H2} \mid b_W))\,\rho^{\text{L2}}$$

and, hence, $E(\rho \mid b_W) > \text{E}(\rho \mid b_R)$, a contradiction. A similar argument rules out that $E(\rho \mid b_R) > 1 - \frac{\mu}{\delta} = E(\rho \mid b_W)$. The only possible remaining cases are: (i) both $E(\rho \mid b_R)$ and $E(\rho \mid b_W)$ are $>1 - \frac{\mu}{\delta}$, (ii) both $E(\rho \mid b_R)$ and $E(\rho \mid b_W)$ are $<1 - \frac{\mu}{\delta}$, (iii) $E(\rho \mid b_W) > 1 - \frac{\mu}{\delta} \geq E(\rho \mid b_R)$. In cases (i) and (ii), the subject does not benefit from belief manipulation in the normative treatment, and thus does not engage in it. In case (iii), CL subjects lie at $t = 1$ when $b = b_W$ but not when $b = b_R$. In this case, belief manipulation may emerge, taking the form of inducing belief b_W instead of b_R. □

I.4 Other-Regarding Condition: Toy Model

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 CL subject lies, he may incur a psychological cost but also experiences a warm glow, denoted by $\gamma\mu \geq 0$, where μ 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 $t = 0$.[36] Second, suppose that the underlying state is H1.[37] Optimal behavior at $t = 1$ is given by:

$$\begin{cases} a = L \text{ if } p(b) < \frac{\gamma\mu}{\theta} \\ a = T \text{ if } p(b) \geq \frac{\gamma\mu}{\theta} \end{cases} \tag{6}$$

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

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

[37] We focus on H1 since this is the state which turned out to apply in the other-regarding condition (as well as the main treatment). Note however that, differently from the main body, in the other regarding case it could also be possible that a norm for lying might emerge (Levine and Schweitzer, 2015). Our theoretical setup focuses on the existence (or not) of a norm for honesty, but it could be easily adapted to lying norms.

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 *If the underlying state is H1: 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 *If the underlying state is H1, within each treatment (empirical/normative), lying in the self-serving condition (weakly) exceeds lying in the other regarding condition.*

II. Other-Regarding (Charity) Condition: Empirical Analysis

We now turn to the empirical analysis analysis for the other-regarding condition, in which the beneficiary of the lie is a charity. For these treatments, a total of 724 have been collected and are analyzed below (a detailed breakdown is provided at the bottom of the figures below). Consistent with our theoretical predictions, the observed behavior is in stark contrast to our results from the main experiment, in which the beneficiary of the lie is the participant: beliefs, lying rates, and conditional lying rates and we do not observe any significant belief distortion, neither in the empirical treatment (45.7% vs. 51.6%, Equality of Proportions Test (EPT), p=0.3462) nor in the normative treatment (37.6% vs. 44.4%, EPT, p=0.2785). As expected, the absence of belief distortion also leads to an absence in significant changes of lying frequency in all treatments (all p-values above 0.3).[38]

[38] 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 and the induced uncertainty about the appropriateness of altruistic behavior, which corroborates existing literature (Di Tella et al., 2015; Exley, 2015).

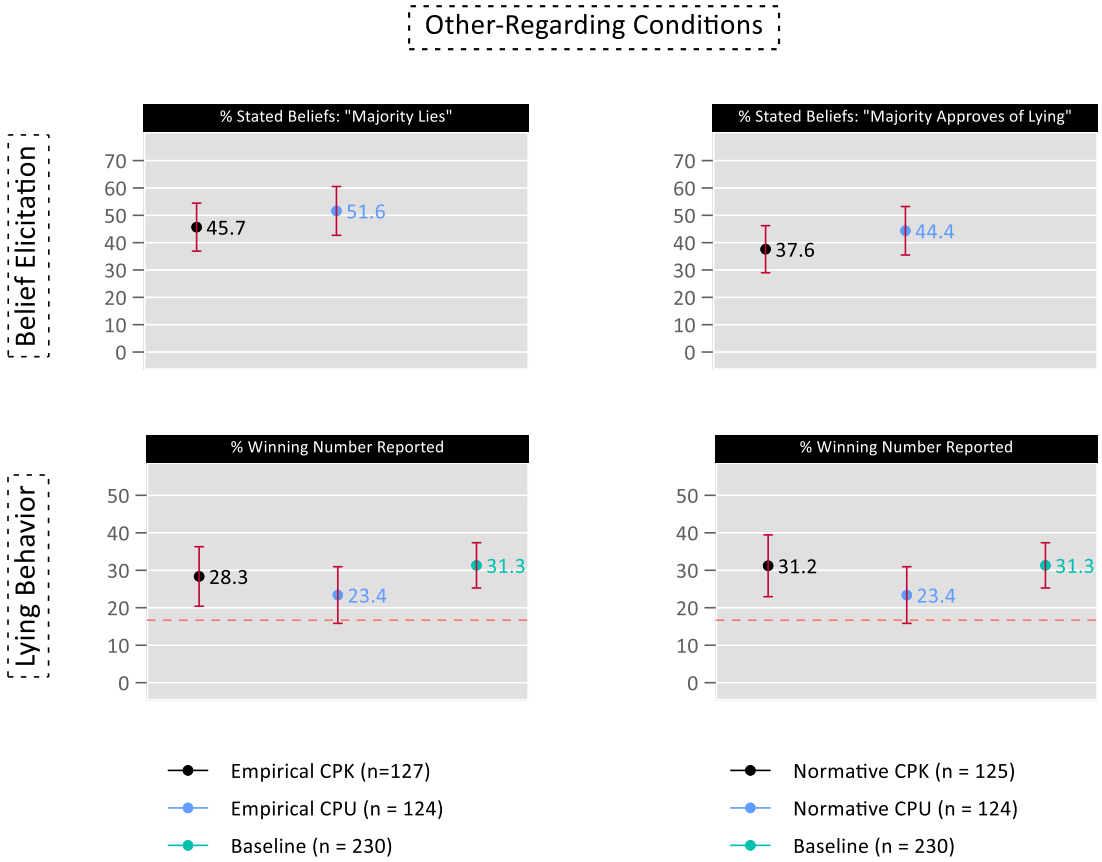


Figure A.1: Results from belief elicitation (top panels) and lying frequency (bottom panels) broken down for the Empirical (left panels) and Normative (right panels) conditions and include the same Baseline. Dotted red line illustrates the expected value of 16.67%. Whiskers indicate 95% confidence intervals. 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 based on the method proposed by Benjamini and Hochberg (1995).

We substantiate our previous results through the lens of a Logit regression analysis, examining both stated beliefs and lying frequency. We ensure the robustness by including a battery of controls (age, gender, risk (SOEP), and CRT score). The regressions fully confirm the previous results and presented in Table A.1.

Table A.1: Logit Regression (Odds Ratios) Analysis of Reported Beliefs and Lying Frequency

Panel A *DV: Reported Beliefs*	(1)	(2)	(3)	(4)
Treatment *(Baseline: CPK Condition)*				
Empirical CPU *(1 = Majority Lies)*	1.269 (0.322)	1.190 (0.311)		
Normative CPU *(1 = Majority Does Not Approve of Lying)*			1.323 (0.343)	1.544 (0.424)
Controls	No	Yes	No	Yes
Observations	251	251	249	249

Panel B *DV: Lying Behavior*	(1)	(2)	(3)	(4)	Empirical CPK = Normative CPK *(Hypothesis 2)*		Empirical CPU < Normative CPU *(Hypothesis 4)*	
Treatment *(Baseline: CPK Condition)*								
Empirical CPU *(1 = Majority Lies)*	0.772 (0.224)	0.802 (0.249)			1.146 (0.317)	1.153 (0.328)	1.000 (0.301)	1.010 (0.305)
Normative CPU *(1 = Majority Does Not Approve of Lying)*			0.673 (0.193)	0.616 (0.188)				
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Observations	251	251	249	249	252	252	248	248

Pannel A and Panel B (Other-Regarding Conditions): Logit regressions (odds ratios reported) with robust standard errors clustered at the individual level displayed in parentheses. 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.

Moreover, in line with our theoretical predictions, we also do not observe any significant relationship between stated beliefs and subsequent lying frequency when a charity is the beneficiary of a lie, neither in the empirical nor in the normative treatment (see Figure A.2). Across all treatments, we do not observe any significant relationship between stated beliefs and subsequent lying behavior.

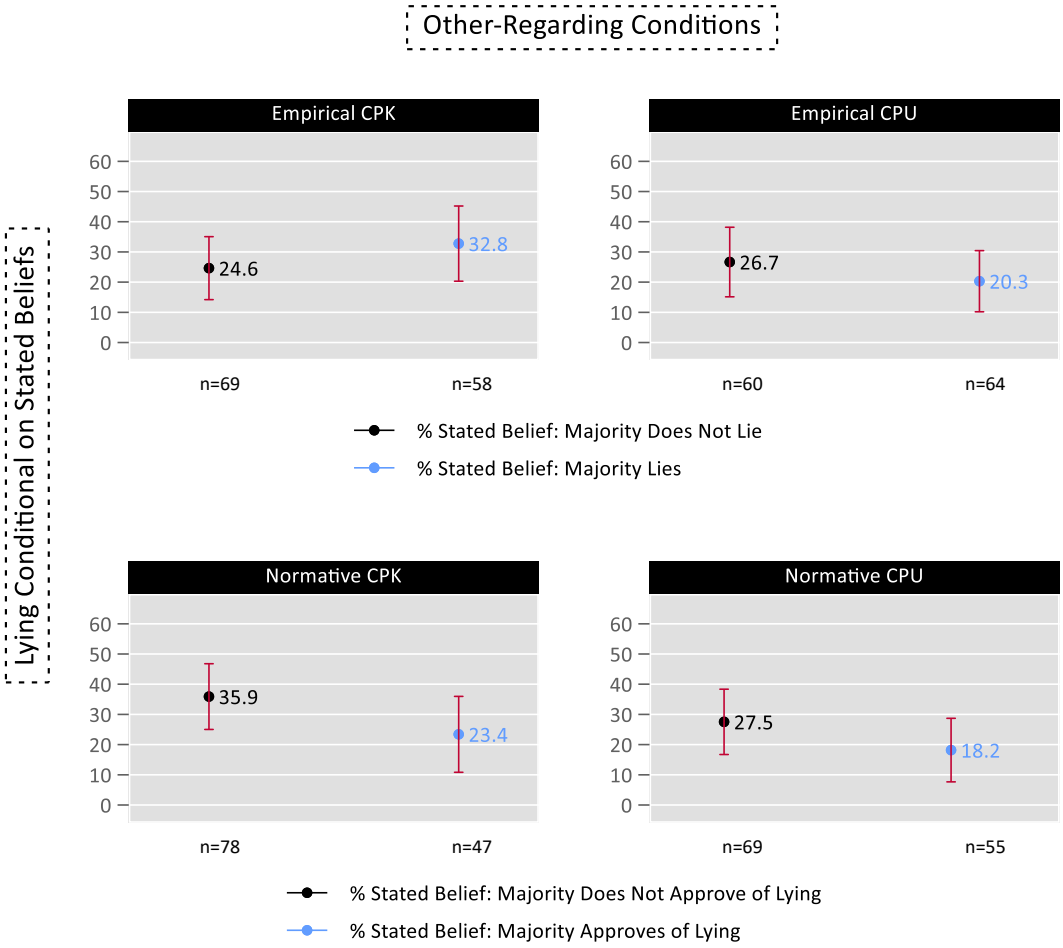


Figure A.2: Results from conditional lying frequency for the self-serving conditions (individual is beneficiary of the lie) and are broken down for the Empirical (top panel) and Normative (bottom panel) conditions. Whiskers indicate 95% confidence intervals. Stars indicate significant differences at the conventional levels of *p<0.1, **p<0.05, and ***p<0.01 and are false discovery rate corrected based on the method proposed by Benjamini and Hochberg (1995).

III. Instructions: Main Experiment

Below we present the instructions that participants were given in the respective treatments. Horizontal lines indicate that information was presented on separate screens. Color-coding emphasizes difference across treatments: Empirical Treatments — Normative Treatments — Only in CPK (and not in CPU) Treatments

University of Pennsylvania
Department of Philosophy, Politics, and Economics
Claudia Cohen Hall, Room 311
Philadelphia, PA 19104
Phone: (215)-898-3023
Fax: (215) 573-2231

Informed Consent/ Assent Form for Non-Pool Participants Earning Money

You are invited to take part in a study named Dice Roll. The purpose of this research study is to explore human decision-making. You will complete a series of computer tasks, each involving semantic as well as visual stimuli materials. If you agree to be in this study, you will need to make decisions and answer questions regarding the study materials. We will also ask you to provide demographic information. We will not ask for your name or any information that will make you identifiable. Overall, this study will take approximately 10 minutes.

For your participation in this study, you will receive a fixed payment of $0.50. Additionally, you may receive a monetary bonus depending on your answers. The exact amount depends on your results in the experiment. The risks to participating are no greater than those encountered in everyday life. Your participation in this study is completely voluntary, and you may refuse to participate or withdraw from the study without penalty or loss of benefits to which you may otherwise be entitled. Compensation will be awarded upon completion of the entire study.

Results may include summary data, but you will never be identified. If you have any questions about this study, you may contact Hannah Harney (Email ppebelab@gmail.com ; Phone: (215) 898-3023.)

For any questions, concerns, suggestions, or complaints that are not being addressed by the researcher, please contact the Institutional Review Board at the University of Pennsylvania, 3624 Market Street, Suite 301 South Philadelphia, PA 19104-6006. Phone: (215) 898-2614.

Please feel free to print or save a copy of this consent form.
By continuing from this page, you are indicating that you have read and understand this consent form and wish to continue your participation in this study.

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.

Belief Elicitation - Self-Serving Condition

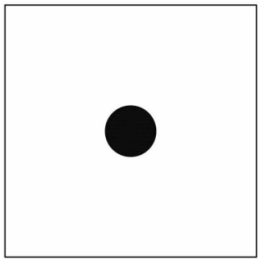
Previous participants rolled an electronic 6-sided dice. This roll generator was **fair** and **unrecorded**. The experimenters were not able to observe what number they rolled and the rolls were generated at random. Those participants then reported the roll that determined what, if any, additional earnings they have received.

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 their own benefit." / "In a similar study most people said it is OK to lie for your own benefit."
- "In a similar study most people did not lie for their own benefit." / "In a similar study most people said it is not OK to lie for your own benefit."

Dice Roll - Self-Serving Condition



Click button to role dice Click Here

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

Belief Elicitation – Other-Regarding Condition

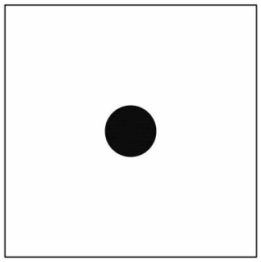
Previous participants rolled an electronic 6-sided dice. This roll generator was **fair** and **unrecorded**. The experimenters were not able to observe what number they rolled and the rolls were generated at random. Those participants then reported the roll that determined what, if any, additional earnings they have received.

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 said it is OK to lie for the benefit of a charity."
- "In a similar study most people did not lie for the benefit of a charity." / "In a similar study most people said it is not OK to lie for the benefit of a charity."

Dice Roll – Other-Regarding Condition



Click button to role dice Click Here

A roll of 1, 2, 3, 4, or 6 does not result in additional 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

Unincentivized questionnaire concluded the experiment (e.g., demographic questions, elicitation of risk). Exact screens available upon request

IV. Instructions: Follow-Up Experiment

Below we present the instructions that participants were given in the respective treatments (same post-experimental questionnaire as in main experiment). Horizontal lines indicate that information was presented on separate screens. Color-coding emphasizes difference across treatments:

Normative Information Provided → Empirical Information Requested —
Empirical Information Provided → Normative Information Requested

---

University of Pennsylvania
Department of Philosophy, Politics, and Economics
Claudia Cohen Hall, Room 311
Philadelphia, PA 19104
Phone: (215)-898-3023
Fax: (215) 573-2231

Informed Consent/ Assent Form for Non-Pool Participants Earning Money

You are invited to take part in a study named Dice Roll Scenario. The purpose of this research study is to explore human decision-making. If you agree to be in this study, you will need to make decisions and answer questions regarding the study materials. We will also ask you to provide demographic information. We will not ask for your name or any information that will make you identifiable. Overall, this study will take approximately 5-10 minutes.

For your participation in this study, you will receive a fixed payment of $0.50. Additionally, you may receive a monetary bonus depending on your answers. The exact amount depends on your results in the experiment. The risks to participating are no greater than those encountered in everyday life. Your participation in this study is completely voluntary, and you may refuse to participate or withdraw from the study without penalty or loss of benefits to which you may otherwise be entitled. Compensation will be awarded upon completion of the entire study.

Results may include summary data, but you will never be identified. If you have any questions about this study, you may contact Hannah Harney (Email ppebelab@gmail.com ; Phone: (215) 898-3023.)

For any questions, concerns, suggestions, or complaints that are not being addressed by the researcher, please contact the Institutional Review Board at the University of Pennsylvania, 3624 Market Street, Suite 301 South Philadelphia, PA 19104-6006. Phone: (215) 898-2614.

Please feel free to print or save a copy of this consent form.
By continuing from this page, you are indicating that you have read and understand this consent form and wish to continue your participation in this study.

Thank you for choosing to participate in this study! It is important that you **read all of the instructions carefully to get your earnings**. This is a survey-based study where you will be asked to answer questions. At the end of the study, you will be asked to fill out a short questionnaire and demographic questions. All presented information is truthful and there is **no deception** in this study.

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 to continue the survey.

In what follows, we describe a truthful scenario from a previous experiment to you:

Scenario: in a previous survey, participants played a game in which they rolled a die. The die was **fair** and the **roll anonymous**, i.e., the experimenter was not able to observe what number was actually obtained. Participants rolled the die and reported a number. The reported number (which could be different from the actual outcome of the roll) determined if the participant received additional earnings. Participants were told that:

- A reported outcome of **1, 2, 3, 4, or 6** would not result in additional earnings for the participant.
- A reported outcome of **5** would earn the participant an **additional $0.25**.

Please read the subsequent statements carefully and answer the questions accordingly, which can yield a bonus payment of $0.25.

As explained above, participants had the option of lying for their own self-interest without incurring in any sanctions in a previous experiment. People were also asked to express their opinions about the acceptability of lying.

Truthful statement: The majority of participants in the experiment did not approve lying for their own benefit

Truthful statement: The majority of participants in the experiment did not lie for their own benefit

Please tell us: what do you think all of the participants in the experiment actually did in the experiment?

Important: guessing correctly will yield an additional bonus payment of $0.25. The correct answer will be revealed at the end of the experiment.

I believe that the...

- Majority of all participants **lied** for their own benefit
- Majority of all participants **did not lie** for their own benefit

I believe that the...

- Majority of all participants **approved of lying** for their own benefit
- Majority of all participants **did not approve of lying** for their own benefit

Please explain your reasoning in detail (text box):

You were just told that the "**majority** of participants in the experiment did not approve lying for their own benefit".

You were just told that the "**majority** of participants in the experiment did not lie for their own benefit".

Please tell us what you believe this "majority" corresponds to in % between 50.1% and 100%.

Input: ___

Post-experimental questionnaire followed (same as in main experiment)