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

Cristina Bicchieri[a,b,c], Eugen Dimant[a,c,d]

[a] *University of Pennsylvania*
[b] *Wharton School of Business*
[c] *Behavioral Ethics Lab*
[d] *CeDEx*

Abstract

We explore the relationship between norm-uncertainty and lying. Lies are ubiquitous, and people often lie for their own benefit or for the benefit of others. Research in environments in which social norms are clearly defined and communicated finds that social norms influence personal decisions, even when they are not in our own self-interest. We deviate from this approach and study lying under norm uncertainty with scope for opportunistic interpretation of the norm. We introduce variation along three dimensions: salience of different types of norm-uncertainty (normative/empirical), the beneficiary of the lie (self/other), and ex-ante knowledge about the opportunity to tell a lie in order to tease out potential belief-distortion mechanisms. We also find compelling evidence that individuals engage in self-serving belief distortion to increase lying overall. However, we observe this only when uncertainty about what others do (empirical uncertainty), but not when uncertainty about what others approve of (normative uncertainty) is made salient. We also observe conditional liars, but only when the lie is self-serving rather than to the benefit of a third party. We discuss policy implications to improve the effectiveness of norm-based interventions.

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

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

Email addresses: cb36@sas.upenn.edu (Cristina Bicchieri), edimant@sas.upenn.edu (Eugen Dimant)

This version: August 10, 2018

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

People often lie to benefit themselves or others. Recent theoretical advancements in economics and psychology have helped shed light on what motivates lying (Dufwenberg Jr and Dufwenberg Sr, 2016; Garbarino et al., 2017; Abeler et al., 2018; Gneezy et al., 2018a). Such motives include, but are not limited to, the magnitude of reward, the impact of the lie on others, the probability of being caught, concern for appearing (dis)honest, social comparisons, and the potential punishment either from rule-breaking or simple non-conformity. A large stream of recent experimental evidence has refined and confirmed the theoretical insights (e.g., Erat and Gneezy, 2012; Abeler et al., 2014; Kajackaite and Gneezy, 2017. For an overview, see Jacobsen et al., 2018 and Abeler et al., 2018).

A different stream of literature examines the role of social norms in disrupting old, informing current, and guiding future behavior. Research in environments in which social norms are clearly defined and communicated to the individuals finds that social norms motivate and affect both self and other-regarding decisions (see Cialdini et al., 1990; Bicchieri, 2006, 2016; Fehr and Schurtenberger, 2018). These situations involve a trade-off between what someone wants to do for one's own benefit (self-serving behavior) and what they want to do for the benefit of others (other-regarding behavior). They also often include information about what others approve of (normative information) and what others actually do in the same situation (empirical information). In real life, however, there may be uncertainty as to which norms apply to (or are followed in) a specific situation, or even uncertainty as to the interpretation of a specific norm (e.g. the many interpretations of fairness in Babcock et al., 1995). Other recent studies have examined the role of risk and uncertainty in guiding and, importantly, finding excuses for not wanting to engage in prosocial behavior (Di Tella et al., 2015; Exley, 2015). These circumstances may lead people to choose what to believe as part of their motivated reasoning. This choice, though not completely arbitrary, has significant effects on how people react to and process different types of social information, especially in uncertain environments.

The ubiquity of norm-uncertainty in real life warrants a closer examination and we extend the existing literature along multiple dimensions. In particular, we explore the relationship between norm-uncertainty and lying. We are interested in how individuals form beliefs about whether a social norm applies, and how such belief-formation is affected by the source of the uncertainty and the purpose of the lie. Our working assumption is that given the opportunity, individuals will attempt to distort their beliefs in a convenient and self-serving way. Theoretical work points at the existence of the mechanism of belief distortion (i.e., Bénabou and Tirole, 2006; Benabou and Tirole, 2006) and experimental evidence suggests that beliefs matter for the deception of others and oneself and that the timing of belief-priming and incentives can bias assessments of conflict-of-interest and normative value judgments (Babcock et al., 1995; Schwardmann and van der Weele, 2017; Gneezy et al., 2018b). Unresolved, however, are questions pertaining to whether such a belief distortion works differently with respect to different parts of a social norm, empirical or normative, and to what extent such mechanisms differ with respect to the beneficiary of the lie. At the same time, our paper helps to augment the discussion regarding the role of norm activation in guiding behavior (Cialdini et al., 1991; Kallgren et al., 2000). It remains a point of debate whether norm salience affects the empirical or normative beliefs of a social norm (Fehr and Schurtenberger, 2018). We attempt to answer these questions.

We adopt a simple modification of the 'die-under-the-cup' paradigm (Shalvi et al., 2011b; Fischbacher and Föllmi-Heusi, 2013) and answer the following questions: first, is lying sensitive to the salience of different types of norm uncertainty? In particular, does it matter if the salient uncertainty is about what others *think* is appropriate (normative expectations), or about what others actually *do* (empirical expectations)? Second, do individuals distort their own beliefs about which norm applies (or even *if* a norm applies) differently when a lie benefits themselves (self-serving condition) as opposed to when it benefits a third party that they care about (other-regarding condition)?[1]

[1]For the latter question, existing research suggests that individuals lie more often if it also benefits others (Erat and Gneezy, 2012; Gino et al., 2013). In these settings, increased lying required that the lie simultaneously benefited both oneself and another. We deviate from this approach by creating an environment in which we isolate the party benefiting from the lie. In other words, in one treatment variation we entirely remove any self-serving element, making any lie result solely in a monetary benefit to a third party. We do the opposite in another treatment variation, removing other-regarding concerns, so that any benefit is self-serving. This approach helps us to clearly distinguish whether and to what extent the mechanisms that we study in this paper are driven by self-serving versus other-regarding motives (see also Exley 2015).

Our paper also makes a methodological contribution by proposing an incentive-compatible approach for studying behavior under norm-uncertainty. In the main experiment, prior to rolling the die, participants are presented with two mutually exclusive statements about what other participants in a previous experiment either did or said they approved of in the same setting, and are incentivized to guess the correct response. The accuracy of the participant's answer is not revealed until after the experiment and no other information is provided. Our treatment variations consist of the following elements: a) varying the type of uncertain information (empirical vs. normative), b) whether participants knew about an upcoming opportunity to lie at the time of belief elicitation (known versus unknown), and c) the beneficiary of the lie (self versus other).

Our experiment yields a number of interesting results. While we find that individuals who form beliefs that justify lying tend to lie more often, this is only true for self-serving lies, and not when a third party is the beneficiary of the lie. In this respect, a positive relation between lying and aligned beliefs exists only when the monetary benefit is self-serving. Importantly, our results also indicate the existence of self-serving belief distortion in the presence of empirical norm-uncertainty, but not when normative information is uncertain. A follow-up study reveals the mechanism: uncertainty regarding what other people do (empirical information) is more malleable, providing individuals with an opportunity to distort their worldview. On the other hand, individuals have a hard time distorting their understanding of what is normatively appropriate.

From a policy perspective, our results highlight the importance of disseminating clear, unambiguous norm information, be it empirical or normative. In our experiment, the normative information is less ambiguous, as lying is rarely seen as appropriate, so that a message stating that most people approve of lying is less credible, and thus less manipulable. Yet there are many cases in which different norms might apply (Xiao and Bicchieri, 2010), or the same norm can have alternative interpretations (think of different interpretations of fairness). In these latter cases, self-serving norm manipulation can occur. Furthermore, social norms have two different components - empirical and normative expectations - which can influence each other. We often have information about what people do as well as what they approve of. When this is not the case, having only one type of information, either empirical or normative, may produce different effects. When we are told that "most people do x", we tend to infer that whoever does x approves of it. Doing x becomes normalized. When the message is "most people approve of x", we may infer that some do x and others do not (think of hypocrisy). After all, words and deeds often differ, so a normative message –

unaccompanied by empirical evidence – is less powerful. Our results help to improve norm-based interventions to reduce deviance and also explain why some of these interventions are more successful than others (Miller and Prentice, 2016; Hallsworth et al., 2017).

We discuss relevant literature and derive testable hypotheses in Section 2, present the experimental design in Section 3 and the results in Section 4, and conclude in Section 5.

2. Related Literature and Hypotheses

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

Norm violation often provides a benefit to the transgressor. If we can convince ourselves that a norm is not followed in the situation we are in, then we have reason to transgress it. Often, we face uncertainty about the norm because it is not fully clear what other people are doing or have done in the same situation. Uncertainty about what the norm is or whether it is presently followed can be solved in a self-serving way if we can provide reasonable justifications for norm-violation (Bicchieri and Chavez, 2013). For example, evidence consistent with the desired behavior often receives preferential treatment (Kunda, 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 what belief is true (Schweitzer and Hsee, 2002). Finally, it is sometimes possible to give a subjective interpretation of a norm (e.g. fairness). If so, people will choose the interpretation that lowers the difference between what is normatively required and what one does (Spiekermann and Weiss, 2016).

The "best" action is a function of an epistemic state, and one often chooses whatever information makes selfish actions appropriate. As a consequence, in situations where we either think a norm is not presently followed or it is not clear what the norm is, we often find justifications for self-serving behavior.

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

Self-serving justifications for lying can thus be motivated either by the desire to protect one's self-image as a moral being, or by a (conditional) preference for not following a social norm, provided we can exploit the uncertainty about the norm itself.

In our experiment, we allow for uncertainty about whether a norm against lying is followed. We separate empirical and normative messages and check (a) whether self-serving beliefs occur more often if empirical or normative information are made salient, and (b) whether changing the recipient of the benefit of lying from oneself to a third party (a charity) leads to a dampening of selfish belief manipulation. More specifically, we ask participants to decide which of two mutually exclusive statements about actual behavior or acceptable behavior is true, and offer a monetary incentive for accuracy.

One can reasonably assume that empirical and normative information differ in their signaling value (see e.g., Danilov and Sliwka, 2016; Danilov et al., 2018). This claim is also substantiated in our follow-up survey, indicating that individuals perceive empirical information to be more malleable than normative information (for details, see results section).

Empirical information pointing out that most people do not lie also allows one to infer that they likely disapprove of lying. On the contrary, believing that most people disapprove of lying does not necessarily lead us to conclude that they also refrain from lying (think of hypocrisy). Here, the norm against lying may be weakened by the concurring belief that people disobey it, providing a conditional follower with a reason to transgress.

In addition, one can argue that it is easier to justify lying by believing that "most people lie", whereas we usually presume that most people disapprove of lying, so the manipulation of the normative message is much more difficult (Bicchieri et al., 2017). In the case of empirical beliefs, on the contrary, one may find widespread lying credible, so it is easier to convince ourselves that others do lie. In this case, we may come to believe that the conditions for norm-following are not met, justifying our own lying. Ethical maneuvering should also lead to justifications for lying, but believing that most people lie should not be a good reason to engage in an immoral act. If one is morally motivated, we should not observe belief manipulation. Based on this reasoning we can derive the following hypotheses.[2] First, we predict that, when participants know that they have the possibility of profiting from dishonest behavior, there will be belief manipulation of the empirical message, much less so of the normative one. Therefore:

H_1: *In uncertain situations, conditional norm-followers who know they can profit from lying will engage in self-serving belief manipulation more with empirical messages than with normative messages.*

H_2: *Conditional norm-followers who engage in self-serving belief manipulation of empir-*

[2]In the Appendix, we derive a model of belief distortion that is motivated by existing work Bénabou and Tirole (2006); Benabou and Tirole (2006). In our setting, a necessary assumption is that the salience of being asked to form beliefs represents a signal in the sense of the original model, which is consistent with the Focus Theory (Cialdini et al., 1991). Then, the signal characterizes a belief-production process, in which one anticipates the effect of current actions on one's future well-being. "Because material actions are more easily codified, recalled, and documented than the exact mix of motives that caused them (evaluation of a hard trade-off, momentary urges, feelings of guilt and pride), our past conduct can be informative about our "deep" preferences and predictive of later behaviors; yet at the same time, our choosing these self-signals makes future beliefs malleable." (Bénabou and Tirole, 2016, p. 147). The latter is at the heart of our experimental design. Importantly, other intuitive models, such as the one introduced by Brunnermeier and Parker (2005), can also be applied to our setup. Since we have not set out testing a particular theory, we remain agnostic about the theory best applicable to our experiment. Instead, we simply present an exemplary formalization of the decision mechanisms and how it maps onto our observed results.

ical messages will lie more than agents who refrain from belief manipulation of empirical messages.

We also predict that engaging in belief manipulation will justify lying for oneself, but will not usually be needed when lying for others. For some people, lying for others provides enough of an independent (altruistic) justification that there should be no need for belief distortion. The norm-based hypothesis predicts that one will choose a justification for not following a norm ("most people lie") if it is in one's self-interest to violate it (Bicchieri and Chavez, 2013), but will need no belief distortion if violating it benefits others. Therefore:

H₃: *In uncertain situations, conditional norm-followers are less likely to engage in self-serving belief manipulation if lying only benefits third parties.*

3. Experimental Design

3.1. General Procedure

We collected a total of 1109 participants with at least 200 observations per cell (we present a detailed breakdown of observations per treatment variation at the bottom of the respective figures in the results section), 100 participants for the pre-experimental survey to derive the truthful empirical and normative information, and another 203 participants for post-experimental survey to investigate the channel of belief distortion on Amazon Mechanical Turk.[3] The mean age of the participants was 35.5 years and 49.8% of them were female. The average duration of the experiment from start to finish was 6 minutes and the average hourly payoff was \$9.50 (including a show-up fee equivalent to \$5 per hour), which is well above average mTurk pay (Arechar et al., 2017).[4] As a result of participants' decisions, a total of \$101.50 was donated to UNICEF.[5]

[3] We applied the following restrictions to the participant pool: participants had to be in the U.S., approval rate was greater than 95% on mTurk, and they had not taken this study before.

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

[5] During the experiment, participants were ensured that the donations towards UNICEF are credible and that official proof of payment will be posted onto our website following the experiment.

3.2. Treatments

Our treatment conditions vary along three dimensions:

I) Whether the salient uncertainty was normative or empirical

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

III) Whether the result of the dice roll benefited the participant or the charity

Variations in (I) occurred between subjects, while variations in (II) occurred within subjects in randomized order. We refer to the treatment combinations as depicted below in Figure 1 and will discuss the detailed experimental procedure in the next section.

3.3. Detailed Procedure

Following the treatment randomization, the remainder of the experiment consists of two parts: the belief-elicitation phase and the dice task.[6]

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

We introduce salient norm-uncertainty through the use of an incentivized belief-elicitation procedure. We explicitly never expose participant to any norm-information, hence leaving participants equally uncertain about norms in all of our treatments. Instead, we make uncertainty salient by eliciting their beliefs about – depending on the treatment – the empirical or normative side of the norm. In line with the Focus Theory by Cialdini et al. (1991), eliciting beliefs will increase the salience and, in turn, lead to individuals confronting themselves about the property of a norm. This induced salience combined with the prevailing uncertainty allows us to study the mechanism of purposeful belief distortion and how it translates to cheating.

Upon signing a consent form, reading the instructions (see Appendix for screenshots), and passing a number of comprehension questions, participants were first presented with an incentivized belief-elicitation task. The belief task asked participants to indicate which of two mutually exclusive statements presented to them was true. The truthfulness of

[6]From the perspective of the participant, the different parts of the experiment had no names to avoid potential priming.

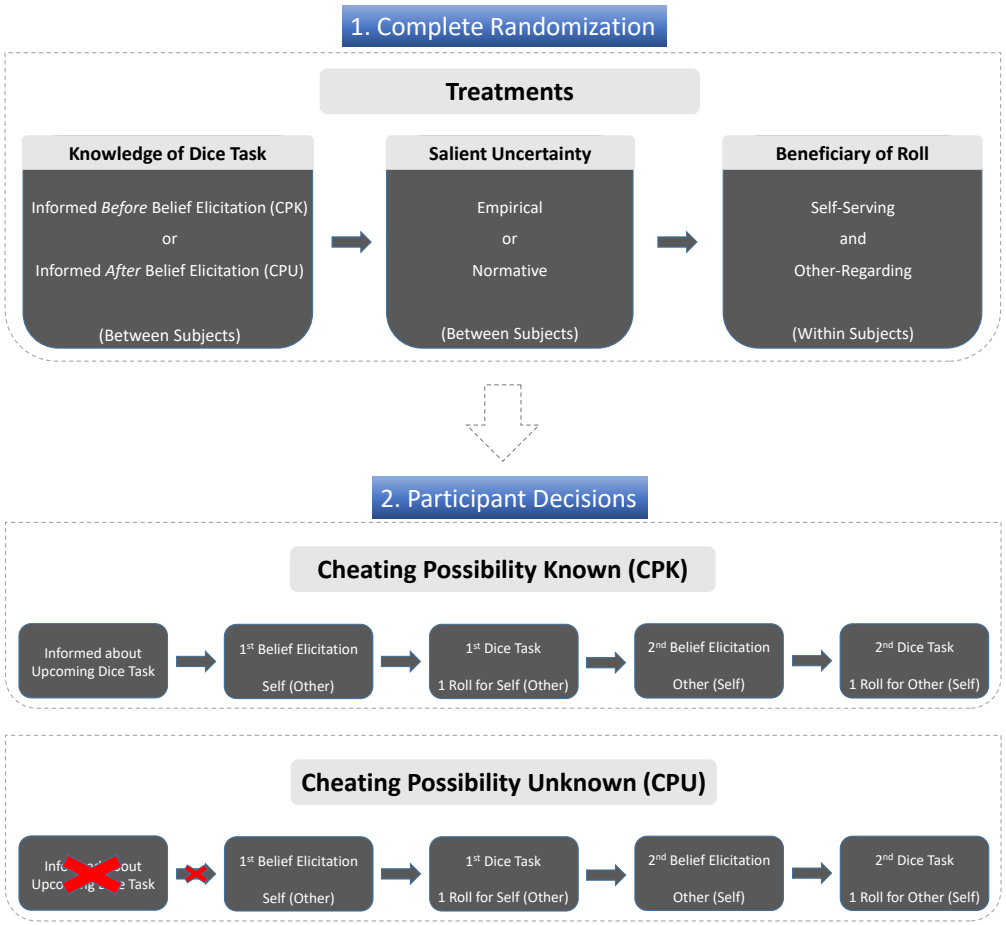


Figure 1: Experimental procedure.

the statements was based on the results from the pre-experimental survey. We use data from a trial session that included questions regarding the appropriateness of lying on the task and vary the beneficiary of the lie (self versus other) in the same way as in the main experiment. From this sample, we collect both empirical and normative beliefs about the frequency and appropriateness of lying and use them as part of the incentivized belief elicitation in the main experiment. At this point, depending on the exact treatment (details below), participants may or may not have already been aware that they will be engaging in the dice task with a cheating opportunity following the belief elicitation.

When presented with two mutually exclusive statements in the belief elicitation, participants had to choose one statement. A correct answer increased the participants' payoff

by \$0.25 (the equivalent of \$2.5 per hour). Importantly, the accuracy of the guesses was revealed only at the very end of the experiment and thus participants were not made aware of the actual truthfulness of the presented statements before the lying opportunity. This procedure was necessary in order to ensure that the norm remained uncertain (i.e. participants were not sure if their selected statement was correct) and did not directly affect participant behavior in the die task. The statements presented to the participants varied across treatments (see Table 1) and the exact wording is presented below (text in brackets indicate wording used in the other-regarding conditions):

NormInfo_Self (NormInfo_Other) Conditions:

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 (for the benefit of the charity)."

or

"In a similar study, most people said it is not OK to lie for your own benefit (for the benefit of the charity)."

EmpInfo_Self (EmpInfo_Other) Conditions:

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 (for the benefit of the charity)."

or

"In a similar study, most people did not lie for their own benefit (for the benefit of the charity)."

It is important to note that this paper does not examine the impact on behavior of providing information per se, i.e., we do not compare uncertain versus non-uncertain norm information. Instead, we are interested in changes in beliefs about a norm when information about it is uncertain, and the uncertainty has the potential to create sufficient wiggle-room that can be exploited to benefit oneself or a third party. Across all treatments, we always

make one type of uncertainty salient to participants and elicit their beliefs regarding the type of information (empirical or normative) they receive. Belief-elicitation alone is enough to produce our results.

Part II: Dice Task

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

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

Importantly, to make treatments comparable, participants were always explained the mechanics of the dice task at the beginning of all treatments, that is before the belief elicitation phase. This ensured that participants knew which task the presented empirical and normative information referred to. What varied between CPK and CPU conditions was the explicit mentioning of whether the participants themselves will engage in the task after the belief elicitation. Note that this ensures that any potentially observable belief-distortion mechanism cannot be explained by demand effects, since, by design, their existence would merely produce a level effect and be unable to explain differences within the same norm-treatment. Noteworthy, for the purpose of comparison, we also ran a baseline condition (Control) in which participants only played the dice task and were not prompted with uncertain norm information. In this sense, the baseline condition is closer to a simplified version of a CPU condition without uncertain norms where no belief distortion was possible. Since by design norms-uncertainty and belief elicitation are absent in this baseline condition, results are merely analyzed in the form of total cheating behavior in

conjunction with the corresponding norm-uncertainty treatments.

4. Results

Our analysis will vary by the beneficiary of the lie (self-serving vs. other-regarding), the extent of knowledge regarding the upcoming lying opportunity (CPK vs. CPU), as well as the type of the salient norm uncertainty (empirical vs. normative information).[7] Because the CPK and CPU conditions are the same except for the knowledge about the subsequent die task, any difference in belief distributions between these two treatments indicates active belief distortion.

We will unpack our findings in multiple steps. First, we will examine differences in stated beliefs. Second, we will examine the role of beliefs in affecting rates of lying. This two-step approach will help us understand whether and to what extent active belief distortion serves the purpose of facilitating lying. In the Appendix, we also report the conditional lying rates (i.e., lying conditional on stated beliefs). Since both the belief elicitation as well as the report of a winning number are binary, we use an equality of proportions test (hereafter: EPT) to examine statistical significance.

4.1. Self-Serving Condition: Beliefs and Lying Behavior

Figures 2 and 3 already indicate our main result: when the lie is self-serving, we observe significant belief distortion only when empirical information is uncertain, but not when normative information is uncertain. In particular, 60.6% of participants in the CPK condition said that most people lie, compared to only 42.4% of participants saying the same in the CPU condition (EPT, $p<0.001$). This alone already offers compelling evidence for active belief distortion because the only difference between both conditions is whether or not one's own lying opportunity could be anticipated at the time of the belief elicitation. The results are displayed in the top part of Figure 2.

Following our theoretical motivation, if belief distortion increases lying, then we should observe that belief distortion leads to significantly more lying. As expected, we find that the percentage of participants reporting the winning number was significantly higher in CPK compared to CPU (EPT, 37.4% vs. 28.8%, $p=0.050$), as displayed in the bottom part of

[7] As discussed above, we employ a randomized within-procedure with respect to the beneficiary of the roll. We do not observe order effects and results remain qualitatively the same when analysis is restricted to the first decision only. Results available upon request.

Figure 2. As we will see in subsequent regressions, this difference becomes even stronger once controlling for relevant covariates.[8] In sum, in combination with the difference in stated beliefs, this finding supports hypothesis $\mathbf{H_2}$.[9]

In a next step, we perform the same exercise for the conditions in which normative information was uncertain and present the results in Figure 3. In line with our predictions, we do not see any significant changes in the proportion of participants reporting that most people don't approve of lying between CPK and CPU (EPT, 67.3% vs 63.6%, p=0.41).

As expected, the absence of belief distortion also leads to the absence of an increase of conditional lying, as indicated in the bottom part of Figure 3. We find that the total percentage of participants reporting the winning number are statistically indistinguishable between CPK and CPU (EPT, 39.1% vs. 38.7%, p=0.94).[10] The absence of significant differences in both stated beliefs and total lying rates between CPK and CPU serves as a counterfactual for our previous main finding: the existence of belief distortion serves the purpose of increasing lying, which is seemingly only possible when empirical instead of normative information is uncertain. A follow-up study reveals the mechanism at play: uncertainty with respect to the empirical information of a norm (what other people do) is seemingly more malleable, providing individuals with an opportunity to distort their understanding of the social environment. On the other hand, individuals have a hard time distorting their understanding of what is normatively appropriate.[11]

[8] This result is even more surprising when considering that the monetary amount that the participant receives for reporting a correct belief is exactly the same as the additional money they gain from reporting a winning number. Thus, in monetary terms their active belief distortion results in forgoing the reward from stating the true distribution of other people's behavior in order to enable themselves to subsequently lie. Individuals seem to prefer to align their belief such that it corresponds to their deviant behavior, as opposed to misaligning beliefs and behavior, i.e., by reporting the correct belief but abstaining from lying. Evidently, people choose not to misalign beliefs and behavior by reporting the actual distribution and still lying (and hence be paid double the amount). Note that this result cannot be explained by our design choice to have participants play both versions of the beneficiary of the lie (self and other) in random order. If one assumes that in the CPU condition the appearance of a cheating opportunity in the second belief elicitation is not completely unknown anymore, the decision should then resemble a decision in the form of CPK and hence yield the exact belief distribution. Our empirical results strongly suggest that this is not the case and the randomization washes out order effects. Our results are robust to an analysis of the subset of only the first decisions in the CPU condition.

[9] We can also compare the frequency of lying to a baseline condition in which no norm-uncertainty was made salient. Here, lying in CPK is marginally significantly higher (EPT, 37.4% vs. 29.4%, p=0.085). Lying frequency in CPU and Baseline are statistically indistinguishable.

[10] Again, one can compare lying rates of these two conditions to the baseline, which are significantly larger in treatments cases (EPT, both p<0.04).

[11] This incentivized experiment was carried out on mTurk after the main sessions were conducted in

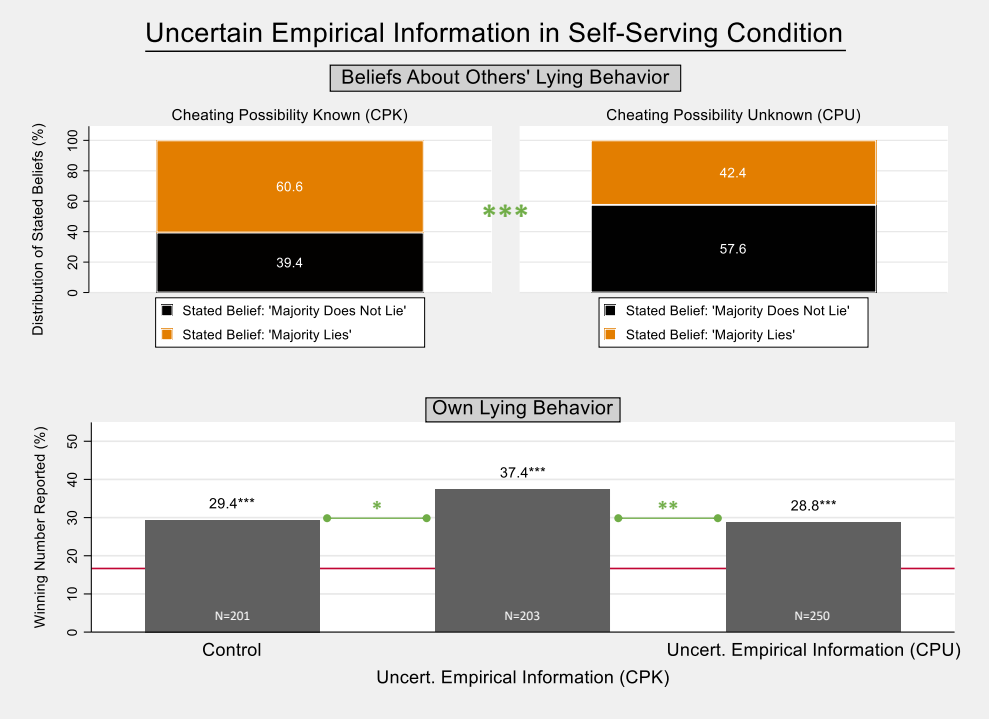


Figure 2: Distribution of stated beliefs (top part) and actual lying behavior (bottom part) when uncertain empirical information is salient in the 'Cheating Possibility Known' (CPK) and 'Cheating Possibility Unknown' (CPU) conditions in the self-serving condition (participant is beneficiary of the lie). In the baseline condition, participants only played the cheating task without any belief elicitation. Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6).

As a secondary result, we find a highly significant relationship between stated beliefs and lying behavior, both when empirical and normative norm-information is uncertain in the self-serving condition. In particular, conditional lying is substantially larger for those who said that most people lie or most people approve of lying, compared to those who said the opposite in the respective treatment. In the interest of brevity, we relegate the results to the Appendix Figures A.1 and A.2.

Overall, our results are strikingly consistent with our hypotheses in that active distortion of beliefs only occurs with salience of uncertain empirical information. Consistent with our subsequent test, normative beliefs are seemingly more difficult to manipulate because

order to investigate the mechanism of belief distortion in more detail. A total of 203 participants answered questions regarding malleability of behavior under varied norm-information and stated beliefs about the behavior of participants in the main experiment. Detailed results are available upon request.

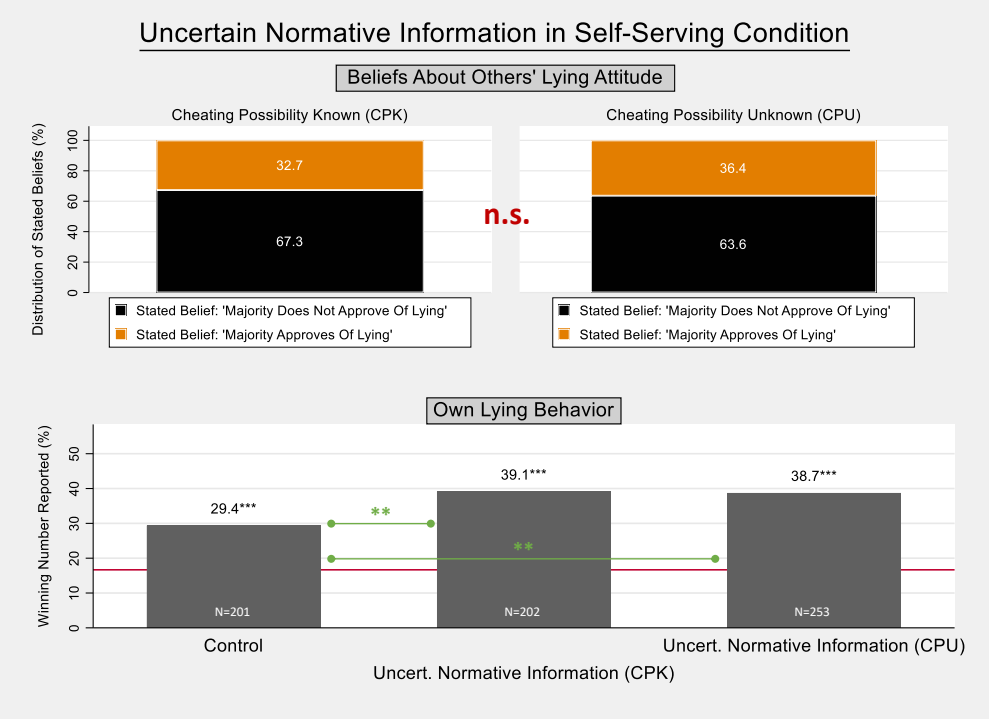


Figure 3: Distribution of stated beliefs (top part) and actual lying behavior (bottom part) when uncertain normative information is salient in the 'Cheating Possibility Known' (CPK) and 'Cheating Possibility Unknown' (CPU) conditions in the self-serving condition (participant is beneficiary of the lie). In the baseline condition, participants only played the cheating task without any belief elicitation. Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6).

our situation yields strong normative guidance. Our results are also in line with recent research suggesting that empirical information might be more uncertain than normative one and negatively affects cooperation rates in environments in which punishment is combined with such information (e.g., Miller and Prentice 2016; Bicchieri et al. 2017).

4.2. Other-Regarding Condition: Beliefs and Lying Behavior

When the beneficiary of the lie is a third party, our results look vastly different. In fact, we do not observe any belief distortion, even when empirical information is uncertain. From this, we conclude that when the beneficiary of a lie is not oneself but a third party in the form of a charity, belief distortion is absent and no relationship between stated belief and lying exists, supporting hypothesis $\mathbf{H_3}$.

The percentage stating that the majority does not lie is not statistically significantly

different between CPK and CPU (EPT, 55.2% vs 45.0%, p=0.13). In consequence, and as expected, the absence of belief distortions also leads to an absence of significant differences in total lying between CPK and CPU. Both proportions of lying are also not significantly different from the Baseline, as displayed at the bottom of Figure 4.

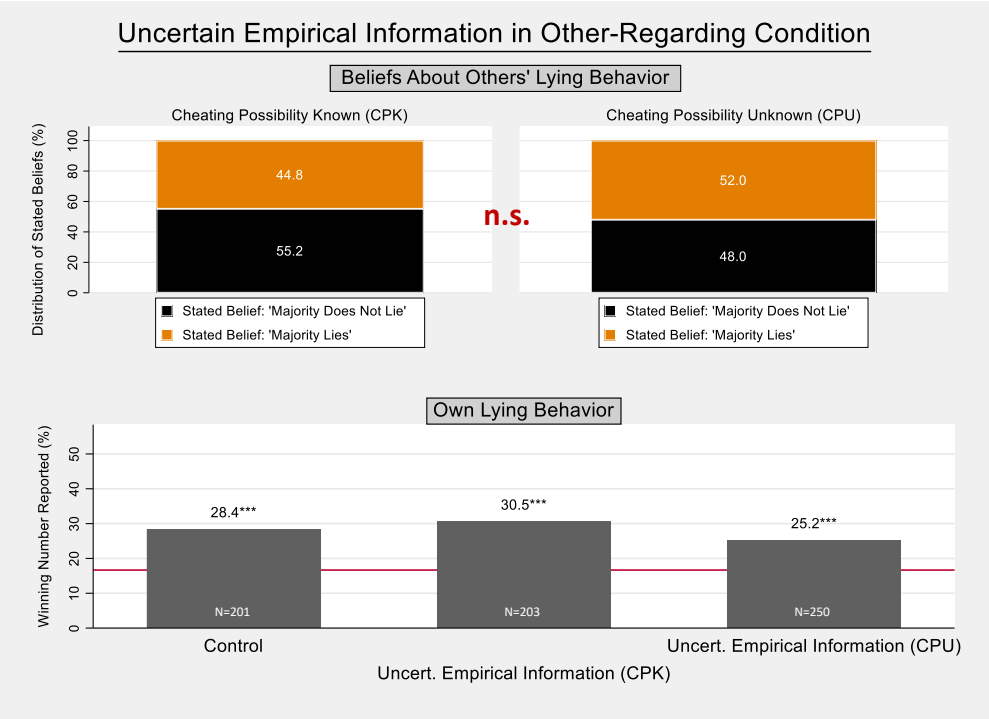


Figure 4: Distribution of stated beliefs (top part) and actual lying behavior (bottom part) when uncertain empirical information is salient in the 'Cheating Possibility Known' (CPK) and 'Cheating Possibility Unknown' (CPU) conditions in the other-regarding condition (charity is beneficiary of the lie). In the baseline condition, participants only played the cheating task without any belief elicitation. Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6).

We perform the same exercise for the treatments in which the salience of uncertainty was of the normative kind. As shown in Figure 5, in line with the results for uncertain empirical information in the other-regarding condition, belief distributions do not significantly change between the CPK and CPU conditions (EPT, p=0.10). Again, there is also no significant relationship between stated beliefs and lying frequency (see Appendix Figure A.2).

We conclude that for both types of uncertainty (empirical and normative), the absence of belief distortion between goes hand in hand with the absence of total lying between CPU

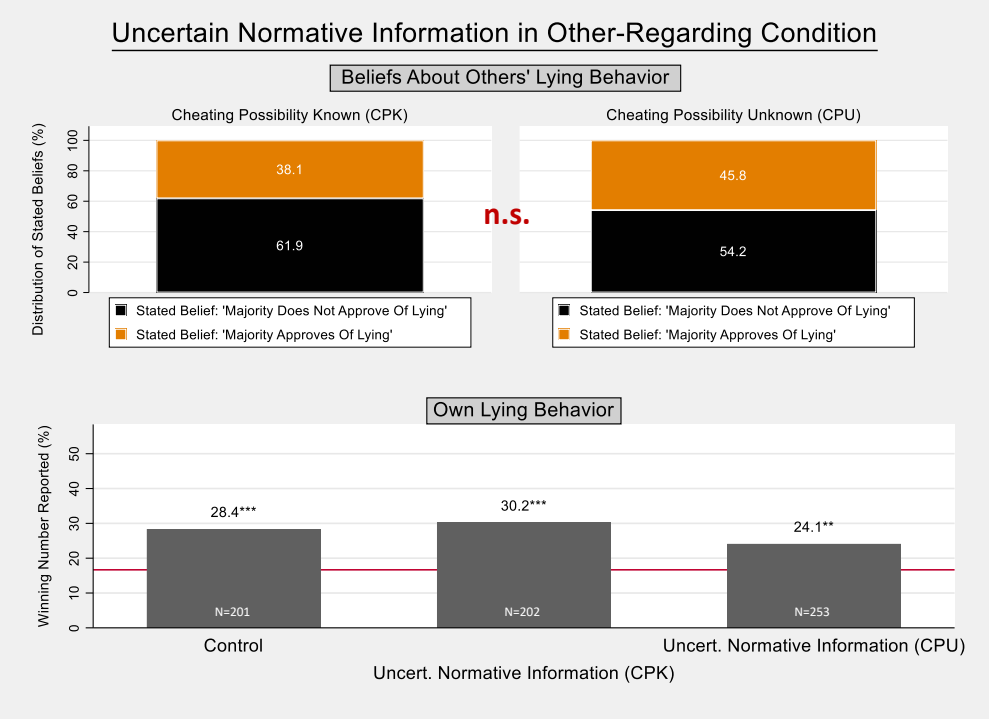


Figure 5: Distribution of stated beliefs (top part) and actual lying behavior (bottom part) when uncertain normative information is salient in the 'Cheating Possibility Known' (CPK) and 'Cheating Possibility Unknown' (CPU) conditions in the other-regarding condition (charity is beneficiary of the lie). In the baseline condition, participants only played the cheating task without any belief elicitation. Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6).

and CPK, as indicated at the bottom of Figures 4 and 5.[12] This finding is in line with our hypothesis $\boldsymbol{H_3}$ in that lying for others provides enough of an independent (altruistic) justification that there should be no need for belief distortion. The norm-based hypothesis predicts that one will choose a justification for not following a norm of *most people lie* if it is in one's self-interest to violate it (Bicchieri and Chavez, 2013), but will need no belief distortion if violating it benefits others. Finally, we also find no significant relationship between stated beliefs and likelihood of reporting a winning number in any of the other-regarding conditions (see Appendix Figure A.2).

[12]Noteworthy, we observe low overall levels of lying in the purely other-regarding conditions. We attribute this to both the absence of self-serving motives as well as the induced uncertainty about the appropriateness of altruistic behavior, which corroborates existing literature (Di Tella et al., 2015; Exley, 2015). Seemingly, this reasoning is even stronger when pure altruism results from a deviant act.

4.3. Regression Analysis

We examine the previously discussed results using Logit regressions (odds ratios indicated) and capitalize on a battery of covariates to examine the robustness of our previously discussed results. We focus on the likelihood of reporting a winning number conditional on the belief stated by the participants. More specifically, we examine behavior in the self-serving conditions in Table 2.

DV: Reported Winning Number	**Uncertain *Empirical* Information (Self-Serving)**		**Uncertain *Normative* Information (Self-Serving)**	
	(1)	(2)	(3)	(4)
Treatment				
Cheating Possibility Known - CPK	1.6072**	1.5015	1.1109	1.2164
(Baseline: Cheating Possibility Unknown - CPU)	(0.3533)	(0.5592)	(0.2290)	(0.3159)
Belief				
'Majority Lies'	1.8889***	1.7994**		
(Baseline: 'Majority Does Not Lie')	(0.4254)	(0.5378)		
'Majority Approves Of Lying'			2.0947***	2.3233***
(Baseline 'Majority Does Not Approve Of Lying')			(0.4396)	(0.6495)
Treatment × Belief		1.1088		0.7888
		(0.4906)		(0.3292)
Post-estimation test for conditional cheating within CPK		**p = 0.0403		**p = 0.0345
Controls	Yes	Yes	Yes	Yes
Observations	453	453	455	455

Table 2: Logit regressions (odds ratios reported) with robust standard errors. Control variables include Age (participant's age, larger value = older) Male (1 = male), Risk (standardized measure, higher score = more risk seeking), Sincerity (HEXACO, higher score = more sincere), CRT Score (1 = solved all questions correctly), Opinion of UNICEF (standardized measure, higher score = higher opinion of UNICEF). Significance levels: *p < 0.10, **p < 0.05, ***p < 0.01.

The results from the regressions mirror our previous findings: there is a clear and highly significant relationship between stated beliefs and report of a winning number as well as a significantly higher frequency of winning numbers reported in the CPK condition as a result of belief distortion.[13] As a result, the overall cheating increases by a factor of approximately 1.6 in CPK compared to cheating in CPU. The regressions for the other-regarding conditions mirror our previous non-parametric results in that there is neither a relationship between stated beliefs and lying behavior as well as no difference in overall

[13] Although not the focus of our paper, one can also estimate the local average treatment effect of the belief distortion under the assumption that beliefs are solely driven by the exogenous treatment variation. For this, we instrument the beliefs with the treatments and run a 2SLS regression, yielding that the highly significant effect is close to 55%, as opposed to the initial lower-bound estimate of the also highly significant average treatment effect of roughly 18% (see Figure 2, 60.6% vs. 42.4%).

cheating rates. In the interest of brevity, we relegate the results to the Appendix Table A1.

5. Discussion and Conclusion

Ample research suggests that individuals react to and are guided by norms. An important question is how individuals behave in environments in which norms are uncertain, in that relevant information is only partially available. The presence of uncertainty about empirical or normative information has scope for self-serving biases. Our main interest is whether, and to what extent, individuals distort their beliefs to facilitate lying. We employ a novel experimental design suited to studying behavior under norm uncertainty.

Our experimental design varies both the salient norm uncertainty (empirical uncertainty about what others do versus normative uncertainty about what others think should be done in this situation) and the beneficiary of a lie (self-serving versus other-regarding). We make norm uncertainty salient by eliciting incentive-compatible beliefs about behavior and beliefs of other participants in a previous survey. In order to establish causality, we vary the timing at which participants learn about their own opportunity to engage in lying: in the "Cheating Possibility Known" (CPK) conditions, participants know prior to the belief elicitation that they will have the opportunity to lie on the task. In the "Cheating Possibility Unknown" (CPU) conditions, participants learn about the opportunity to lie only after the belief elicitation took place. Our working assumption is that belief distortion takes place whenever individuals are aware of their own lying opportunity at the time of the belief elicitation. If we observe a change in the belief distribution, we would also assume that this translates into higher rates of lying.[14]

Our experiment yields a number of interesting results. We find a significant relationship between lying and expressing a belief aligned with it (stating either that most people lie or that most people approve of lying). We also find strong evidence that individuals engage in belief distortion in support of lying. However, this only takes place when two conditions apply: (a) the empirical information (what other people do) is uncertain and (b) one

[14] Arguably, players may honestly convince themselves about the truth of the "majority lies" statement in CPK. This would represent wishful thinking rather than active distortion. In that case, cheating would grant them the highest (subjective) expected value. In CPK, the desire to subsequently cheat may color the assessment of what others do, whereas in CPU no such desire could occur. It might be argued that even wishful thinking is a form of belief distortion, since, in the absence of knowledge of the cheating possibility, such thinking would not have occurred.

is the sole beneficiary of the lie. We make a causal inference by comparing behavior in a situation in which an upcoming opportunity to lie was known to one in which it was unknown to the participants at the time of the belief elicitation. In the former case, we find that most participants report that most people lie for their own benefit, whereas in the second case most participants state that most people do not lie for their own benefit. In this way, individuals choose a belief that is aligned with their own preferred subsequent action: lying. This, in turn, then also leads to higher cheating rates overall, which supports the notion that the reason for individuals to distort their beliefs is to increase cheating. Conversely, when normative uncertainty is made salient and one is the beneficiary of the lie, we do not observe belief distortion, although the relationship between lying and the aligned belief statement ("The majority of other people lie for their own benefit.") still holds. Interestingly, neither belief distortion nor the positive relationship between lying and aligned stated belief are present when the beneficiary of the lie is a third party. We observe that lying does not depend on the stated belief and that individuals do not engage in belief distortion. Lying for a good cause is enough of a justification for those so inclined.

In sum, our findings suggest that uncertain empirical information about a norm (whether it is followed or not) can be more damaging than uncertain normative information (whether the behavior is approved or not). As argued (and is conclusively showing up in a follow-up study), this is driven by the fact that these types of information vary in their signaling content: words are cheap, but actions are costly, which directly affects the decision to lie and distort one's beliefs (see Bénabou and Tirole (2006); Benabou and Tirole (2006); Bénabou and Tirole (2016) and also Appendix Section II). From a policy perspective, we recommend sending unambiguous information about what behavior is common in a particular environment. If the common behavior is negative, the best option would be identifying subgroups that behave in a positive way and broadcasting their behavior. As to normative messages, the results are mixed. Words and deeds often differ, and stressing what people approve of may not necessarily induce good responses. The best option seems to be a combination of congruent empirical and normative information, as positive and unambiguous as possible (for a discussion see Schultz et al., 2007, 2018).

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*.

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 Payne, A. A. (2003). Do government grants to private charities crowd out giving or fund-raising? *American Economic Review*, 93(3):792–812.

Arechar, A. A., Kraft-Todd, G. T., and Rand, D. G. (2017). Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, 3(1):1–11.

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.

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

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

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

Benson, P. G., Curley, S. P., and Smith, G. F. (1995). Belief assessment: An underdeveloped phase of probability elicitation. *Management Science*, 41(10):1639–1653.

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., Dimant, E., and Xiao, E. (2017). Deviant or wrong? the effects of norm information on the efficacy of punishment. Technical report, The Centre for Decision Research and Experimental Economics (CeDEx) No. 2017-14.

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

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

Cialdini, R. B., Kallgren, C. A., and Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*, volume 24, pages 201–234. Elsevier.

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.

Danilov, A., Khalmetski, K., Sliwka, D., et al. (2018). Norms and guilt. Technical report, CESifo Group Munich.

Danilov, A. and Sliwka, D. (2016). Can contracts signal social norms? experimental evidence. *Management Science*, 63(2):459–476.

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.

Dufwenberg Jr, M. and Dufwenberg Sr, M. (2016). Lies in disguise – A theoretical analysis of cheating. university of arizona. Technical report, mimeo.

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

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

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

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.

Garbarino, E., Slonim, R., and Villeval, M. C. (2017). Loss aversion and lying behavior: Theory, estimation and empirical evidence. Technical report, GATE Working Paper No. 1631.

Gino, F., Ayal, S., and Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of economic behavior & organization*, 93:285–292.

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., and van Veldhuizen, R. (2018b). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association*.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

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

Jacobsen, C., Fosgaard, T. R., and Pascual-Ezama, D. (2018). Why do we lie? A practical guide to the dishonesty literature. *Journal of Economic Surveys*, 32(2):357–387.

Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102:433–444.

Kallgren, C. A., Reno, R. R., and Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and social psychology bulletin*, 26(8):1002–1012.

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.

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

Miller, D. T. and Prentice, D. A. (2016). Changing norms to change behavior. *Annual review of psychology*, 67:339–361.

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.

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.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2018). The constructive, destructive, and reconstructive power of social norms: Reprise. *Perspectives on psychological science*, 13(2):249–254.

Schwardmann, P. and van der Weele, J. J. (2017). Deception and self-deception. Technical report, Discussion Paper.

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., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011a). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

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. (2011b). Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management*, 22(s1).

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.

Xiao, E. and Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology*, 31(3):456–470.

Appendix

I. Additional Results

First, we assess the relationship between stated beliefs and conditional lying for the self-serving treatments for uncertain empirical and normative separately. All results are indicated in the top and bottom parts of Figure A.1

I.1: Self-Serving Conditions

Uncertain Empirical Information: When the cheating possibility was known (CPK), 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 (EPT, 44.7% vs. 26.3%, $p<0.01$). The same holds in the CPU condition (EPT, 34.0% vs. 21.7%, $p=0.033$). Interestingly, when participants stated that most people people lie, lying was lower in the CPU as compared to the CPK condition at the 10% significance level (EPT, 34.0% vs. 44.7%, $p=0.074$). This result suggests that, conditional on stating that the majority lies, belief distortion leads to a higher rate of lying when the lying opportunity can be anticipated.

Uncertain Normative Information: As before, lying is associated with beliefs that lying is approved of, yielding highly significant differences in both the CPK condition (EPT, 50.0% vs. 33.8, $p<0.01$) and the CPU condition (EPT, 53.3% vs. 30.4%, $p<0.01$).

Note that the significant difference in lying behavior in CPU between those who believe that the majority does not lie and those who believe that the majority disapproves of lying can be explained by the asymmetry between empirical and normative information. When we believe that most people do not lie, we also infer that they disapprove of lying, a powerful inducement to be honest. On the contrary, believing that most people disapprove of lying does not necessarily lead us to conclude that they also refrain from lying. The norm against lying in this case may be weakened by the concurring belief that people disobey it, providing a conditional follower with a reason to transgress.

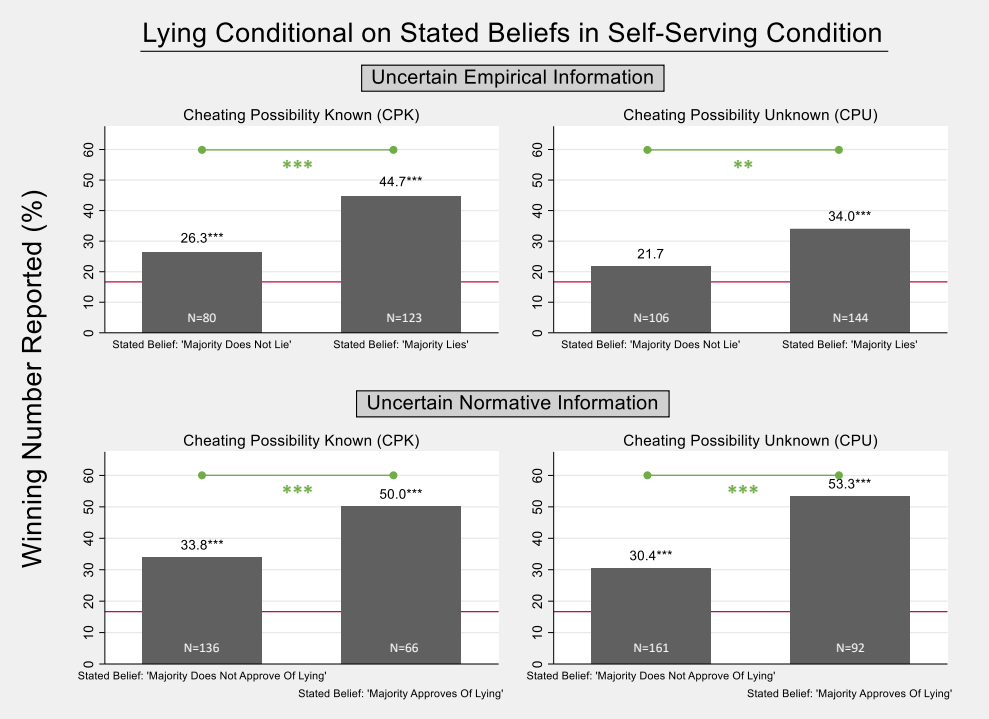


Figure A.1: Lying rates conditional on stated beliefs when uncertain empirical (top part) and normative (bottom part) information is made salient both in the "Cheating Possibility Known" (CPK) and "Cheating Possibility Unknown" (CPU) conditions and when the participant is the beneficiary of the lie (self-serving). Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6).

I.2: Other-Regarding Conditions

In contrast to the previous findings, we do not observe any significant relationship between stated beliefs and subsequent lying behavior when a charity is the beneficiary of a lie both when empirical and normative information is uncertain. The results are presented in Figure A.2.

Uncertain Empirical Information: When empirical information is uncertain, this is true for both the CPK condition (EPT, 36.2% vs. 26.8%, p=0.2) and the CPU condition (EPT, 20.8% vs. 30.0%, p=0.10).

Uncertain Normative Information: Again, when the charity is the beneficiary, we do not observe any significant relationship between stated beliefs and lying behavior, both

when the upcoming lying opportunity is known (EPT, 26.0% vs. 32.8%, p=0.30) and when it is unknown (EPT, 19.8% vs. 27.7%, p=0.14)

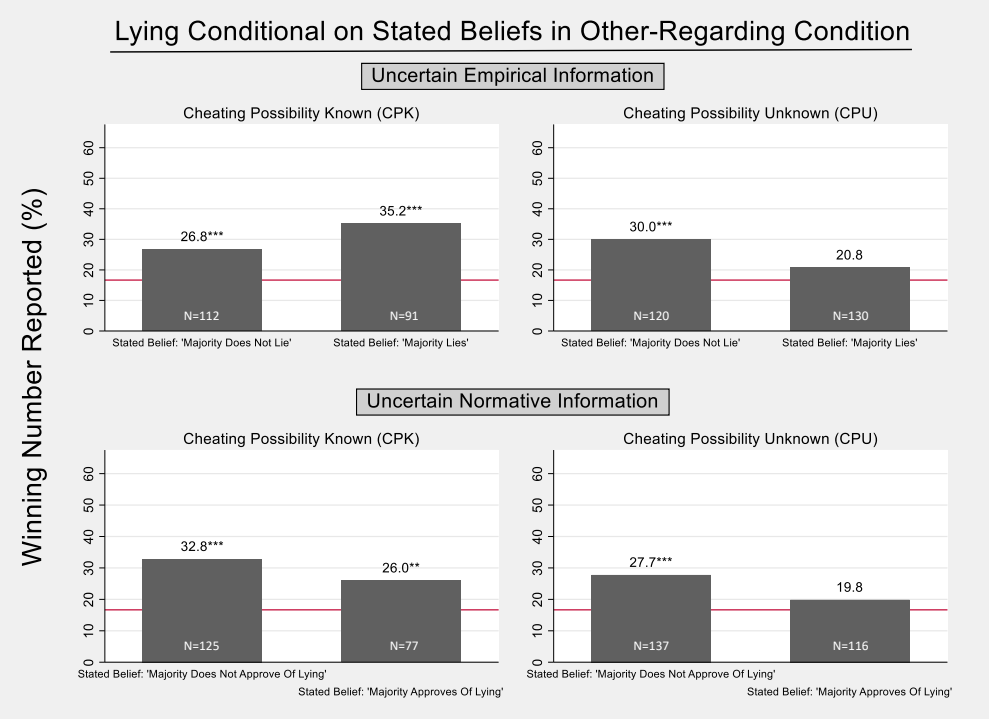


Figure A.2: Lying rates conditional on stated beliefs when uncertain empirical (top part) and normative (bottom part) information is made salient both in the "Cheating Possibility Known" (CPK) and "Cheating Possibility Unknown" (CPU) conditions and when the charity is the beneficiary of the lie (other-regarding). Stars next to the numbers indicate significant differences compared to the expected value of 16.67% (1 out of 6). No significant differences between treatments.

DV: Reported Winning Number	**Uncertain *Empirical* Information (Other-Regarding)**		**Uncertain *Normative* Information (Other-Regarding)**	
	(1)	(2)	(3)	(4)
Treatment				
Cheating Possibility Known - CPK	1.3497	0.9418	1.4015	1.3426
(Baseline: Cheating Possibility Unknown - CPU)	(0.2991)	(0.2879)	(0.3205)	(0.3747)
Belief				
'Majority Lies'	0.9182	0.6480	0.7419	0.7025
(Baseline: 'Majority Does Not Lie')	(0.1978)	(0.1947)	(0.1687)	(0.2189)
'Majority Approves Of Lying'				
(Baseline 'Majority Does Not Approve Of Lying')				
Treatment × Belief		2.0896*		1.1256
		(0.9005)		(0.5024)
Post-estimation test for conditional cheating within CPK		p = 0.1486		p = 0.2572
Controls	Yes	Yes	Yes	Yes
Observations	453	453	455	455

Table A1: Logit regressions (odds ratios reported) with robust standard errors. Control variables include Age (participant's age, larger value = older) Male (1 = male), Risk (standardized measure, higher score = more risk seeking), Sincerity (HEXACO, higher score = more sincere), CRT Score (1 = solved all questions correctly), Opinion of UNICEF (standardized measure, higher score = higher opinion of UNICEF). Significance levels: *p < 0.10, **p < 0.05, ***p < 0.01.

II. Theoretical Examination[1]

In this section, we capitalize on the theoretical ideas put forward in the existing literature (Bénabou and Tirole, 2006; Benabou and Tirole, 2006). The model solidifies the mechanism in which the anticipation of a lying opportunity triggers self-serving belief distortions.[2] At the heart of our argumentation is the idea that individuals may strategically distort their beliefs such that it provides them with an alibi for transgressive behavior. As will be shown, in the environment in which our experiment operates this is expected to occur when three conditions are met: (1) a lying opportunity can be anticipated prior to belief elicitation, (2) norm-uncertainty exists on the side of empirical information, that is what other people do, and (3) the lie is self-serving and monetary benefits oneself.

Consider a setup where individuals belong to one of three types: (i) unconditional Liars (UL) – incur no psychological cost from lying; (ii) unconditional Honest (UH) – incur a prohibitively high cost from lying and, as a result, never lie; (iii) conditional Liars (CL) – they incur a lying cost equal to θk, where θ is a positive constant (the same for all CL) and k denotes the (environment-specific) empirical norm, corresponding to the average share of truth-tellers in a given environment. We denote the share of UL in the population as α, the share of UH as γ and the share of CL in the population as $1 - \alpha - \gamma > 0$.

Agents do not know the exact value of k – this could occur if the exact values of α and γ in the population are not known, which in turn would generate uncertainty about the share of people who would lie in any given environment. For simplicity, we take a reduced form approach and assume that k can take one of two values, $\overline{k}$ and $\underline{k}$, where $\overline{k} > 1/2 > \underline{k}$.

Lying decision

We first consider the self-serving condition. Let the action lie/don't lie be indicated by $a \in \{0, 1\}$; lying corresponds to $a = 1$ and telling the truth corresponds to $a = 0$. Since UL agents always lie and UH never do, the decision to lie or not only concerns CL. If he lies, a CL agent incurs an immediate psychological cost given by $\theta E(k)$ (where $E(k)$ represents his beliefs about k), but obtains a monetary payoff that will allow him to enjoy greater consumption utility next period, indicated as $m > 0$.

[1]We thank Silvia Sonderegger for her input.

[2]Our working assumption is that lying costs depend on the fraction of liars (empirical information) in one's reference group, rather than what is being considered appropriate behavior (normative information). While this assumption is not crucial to what we have set out to do, it is in line with Bicchieri and Xiao (2009); Danilov et al. (2018); Abeler et al. (2018), among others.

We assume that individuals exhibit a present bias, formally captured by quasi-hyperbolic discounting (for simplicity we set the exponential discounting component equal to 1). Formally, a CL individual who lies obtains the following instantaneous utility

$$-\beta\theta E(k) + m$$

where $\beta > 1$ is the present bias. One way of thinking about this is that, when choosing whether to lie or not, the psychological cost of lying is more salient, since it is incurred immediately, while the benefit from the lie is somewhat delayed and thus less salient. We assume that $\overline{k} > \frac{m}{\beta\theta} > \underline{k}$, i.e. a CL agent who knows for sure that the state is high ($k = \overline{k}$) will choose not to lie, and a CL agent who knows for sure that the state is low ($k = \underline{k}$) will choose to lie.

Beliefs

The belief formation phase happens *before* the agent takes the lie/don't lie decision. The agent observes the realization of a signal.[3] $s \in \{L, H\}$ that is informative about k, as follows:

$$\begin{aligned} \text{if } k &= \overline{k} : s = H \text{ with probability } p,\ s = L \text{ with probability } 1-p \\ \text{if } k &= \underline{k} : s = H \text{ with probability } 1-q,\ s = L \text{ with probability } q \end{aligned}$$

where $p > 1 - q$. Let k^H denote $E(k \mid s = H)$, k^L denote $E(k \mid s = L)$ and $\widetilde{k}$ denote $E(k)$ – the unconditional prior common to all agents, where $k^H > \widetilde{k} > k^L$.

CPU Condition

We start of by considering the CPU condition; the agent has no interest in distorting his beliefs since he does not anticipate the future lying opportunity.

Proposition 1 *In the CPU condition:*

[3] In our experimental setup, the signal occurs in the form of a prompted belief elicitation, which is decomposed in two mutually exclusive statements (δ and $1 - \delta$) that are simultaneously presented to the individual in an incentive compatible way. Arguably, such a prompt can represent a salient signal for beliefs either to form new beliefs or retrieve already formed beliefs through introspection (see Benson et al. 1995 and Cialdini et al. 1991).

If $k^H < \frac{m}{\beta\theta}$*: a CL agent always lies.*

If $k^H > \frac{m}{\beta\theta}$*: a CL agent lies if he has observed* $s = L$*, but not if he has observed* $s = H$*.*

Intuitively, $k^H < \frac{m}{\beta\theta}$ means that the high signal is not informative enough to induce the agent to tell the truth. Even though he has observed the high signal, the agent still believes that the low state is sufficiently likely to find it optimal to lie.

CPK Condition

We now consider the CPK condition. This introduces the possbility that, upon observing $s = H$, the agent may choose to distort his signal – if he does, then when the time comes to choose to lie or not, he will recall $s = L$. Distorting involves a cost $c > 0$. Formally, the game has three stages.

- $t = 0$ *Belief formation* The agent observes s and (if $s = H$) decides whether to distort or not.
- $t = 1$ *Lying decision* The agent recalls H or L from $t = 0$ and decides to lie or not.
- $t = 2$ *Consumption* If the agent lied at $t = 1$, he enjoys utility m from consumption.

Benabou and Tirole assume that agents cannot systematically fool themselves, which here would imply that an agent who recalls L at $t = 1$ is aware that this might be the result of signal distortion at $t = 0$. In that case, the agent realizes that recalling $s = L$ may be uninformative, and his beliefs in that case are equal to $\widetilde{k}$, his prior beliefs. Accordingly, a necessary (but not sufficient) condition for distortion to arise in equilibrium is that

$$\frac{m}{\beta\theta} > \widetilde{k} \tag{1}$$

In what follows we will assume that (1) holds. This condition says that, under prior beliefs, a CL agent would always find it optimal to lie.[4]

[4] Alternatively, we could have assumed that agents are naive at $t = 1$ and do not realize that recalling $s = L$ may be the result of belief distortion. In that case, distortion would require a weaker condition, namely $\frac{m}{\beta\theta} > k^L$.

Consider now the decision at $t = 0$ of an agent who has observed $s = H$. Suppose that, to induce his future self to lie, the agent must engage in belief distortion. The agent chooses to distort if and only if

$$-\beta c - \theta k^H + m > 0. \tag{2}$$

The distortion cost c is multiplied by β because it is immediate, while the psychological cost of lying and the utility from higher consumption are both in the future. Note that, in (2), the agent's expectation of k is k^H since, at $t = 0$, the agent knows that the true realization is $s = H$. In what follows, we will restrict attention to environments where

$$m > \theta k^H. \tag{3}$$

From (2), this ensures that, if c is sufficiently low, a CL agent will find it optimal to distort his beliefs in order to induce his future self to lie.

Proposition 2 *In the CPK condition:*

If $k^H < \frac{m}{\beta\theta}$*: a CL agent never distorts, but always lies.*

If $k^H > \frac{m}{\beta\theta}$ *and* $\frac{m-\theta k^H}{\beta} > c$*: a CL agent always distorts and always lies.*

If $k^H > \frac{m}{\beta\theta}$ *and* $\frac{m-\theta k^H}{\beta} < c$*: a CL agent never distorts and lies only if he has observed s=L.*

Note that distortion may arise only if

$$\frac{m-\beta c}{k^H} > \frac{m}{\beta k^H} \;\rightarrow \frac{\beta - 1}{\beta^2} > \frac{c}{m}$$

which, fixing c and m, requires $\beta - 1$ sufficiently large. In words, this implies that the present bias must be sufficiently strong. If the present bias was entirely absent ($\beta = 1$) then belief distortion would never happen in equilibrium. Intuitively, the present bias introduces a conflict of interest between the $t = 0$-self and the $t = 1$-self, that takes the form of the latter wanting to lie less often than the former would like him to. That's because, due to present bias, the $t = 1$-self overweights the (immediate) psychological cost from lying relative to the delayed reward. By contrast, from the $t = 0$-self's pespective, both the psychological cost of lying and the reward will occur in the future and are thus unaffected by the present bias.

Propositions 1 and 2 have immediate implications in terms of lying shares. In what follows I will for concreteness assume that the underlying true state is $\overline{k}$.[5]

- In CPU, the share of liars is $\alpha + (1-p)(1-\alpha-\gamma)$ — namely, all UL agents + all CL agents who observe $s = L$.
- In CPK, when distortion occurs the share of liars is $\alpha + (1-\alpha-\gamma) = 1-\gamma$ — namely, all UL + all CL agents, and is thus higher than in CPU.

Corollary *When distortion occurs, the share of liars in CPK is strictly higher than in CPU.*

We can also make additional predictions. Suppose that distortion occurs in CPK, and that the agents whose beliefs have been distorted believe that "the majority lies."

- In CPK, the share of liars among those who believe that "the majority lies" is $\frac{1-\gamma-p\alpha}{1-p(\alpha+\gamma)}$. That's because the mass of those who believe that "the majority lies" is equal to the mass of UL and UH agents who have observed $s = L$ and *all* CL agents, and is thus equal to $(1-p)(\alpha+\gamma)+1-\alpha-\gamma$, while the mass of liars who also believe that "the majority lies" is $\alpha(1-p)+(1-\alpha-\gamma)$. The share of liars among those who believe that "the majority lies" is thus equal to $\frac{\alpha(1-p)+(1-\alpha-\gamma)}{(1-p)(\alpha+\gamma)+1-\alpha-\gamma} = \frac{1-\gamma-p\alpha}{1-p(\alpha+\gamma)}$.
- By contrast, in CPU, the share of liars among those who believe that "the majority lies" is $1-\gamma$ (and is thus lower than $\frac{1-\gamma-p\alpha}{1-p(\alpha+\gamma)}$, the corresponding share in CPK). That's because the mass of those who believe that "the majority lies" is equal to the mass of all agents who have observed $s = L$, and is thus equal to $1-p$, while the mass of liars who also believe that "the majority lies" is $\alpha(1-p)+(1-\alpha-\gamma)(1-p) = (1-\gamma)(1-p)$.

Proposition 3 *Suppose that the agents whose beliefs have been distorted believe that "the majority lies."Then, when distortion occurs in CPK, the share of liars among those who believe that "the majority lies" is higher than CPU.*

[5] Although this is in line with our empirical observation, this assumption is not particularly crucial and the results also hold for $\underline{k}$.

Discussion: Normative vs. Empirical Information

The normative and the empirical information conditions can be modelled as environments that differ in the nature of the signal available to the agent. Intuitively, it is easy to see that empirical information is more informative about underlying lying rates than normative information. In particular, due to the "hypocrisy effect" you describe, I would expect k^H, the agent's posterior beliefs about the share of truthtellers after observing a high signal, to be higher in the empirical than the normative treatment. In the normative treatment the high signal would correspond to "majority does not approve of lying,"which is not a strong signal that the majority won't lie since many people don't walk the talk, while in the empirical treatment it would correspond to "majority does not lie,"which is a stronger signal that the majority might actually not lie.

Suppose then that, in the normative treatment, $k^H < \frac{m}{\beta\theta}$, while in the empirical treatment, $k^H > \frac{m}{\beta\theta}$ *and* $\frac{m-\theta k^H}{\beta} > c$. Then, following proposition 2, the model would predict that, in the normative treatment, a CL agent never distorts, but always lies (in both CPU and CPK). While in the empirical treatment, a CL agent always distorts and always lies in CPK, and only lies when observing a low signal in CPU. This would deliver the following implication:

- The share of liars in both CPU and CPK of the normative treatment is the same as that of the CPK empirical treament and above that of the CPU empirical treatment.

Which is exactly what we observe.

Other-Regarding Condition

Suppose now that the beneficiary of the lie is a charity. If he lies, a CL agent incurs an immediate psychological cost given by $\theta E(k)$ but also experiences an immediate warm glow denoted by $w > 0$. Formally, a CL individual who lies obtains the following instantaneous utility

$$-\beta\theta E(k) + \beta w. \tag{4}$$

Note that this case differs from the self-serving condition in that *both* the psychological cost of lying and the psychological reward (warm glow) are incurred immediately. As we

will see, this implies that the present bias has no bite.[6] Consider now the decision at $t = 0$ of an agent who has observed $s = H$. Belief distortion may occur only if the agent knows that his future self would tell the truth if confronted with $s = H$, i.e., that

$$k^H > \frac{w}{\theta}. \tag{5}$$

If (5) doesn't hold, the agent would *always* lie, even upon observing the high signal, and thus there would be no need to distort. Suppose then that (5) holds. The agent chooses to distort if and only if

$$-\beta c - \theta k^H + w > 0 \rightarrow \frac{w - \beta c}{\theta} > k^H \tag{6}$$

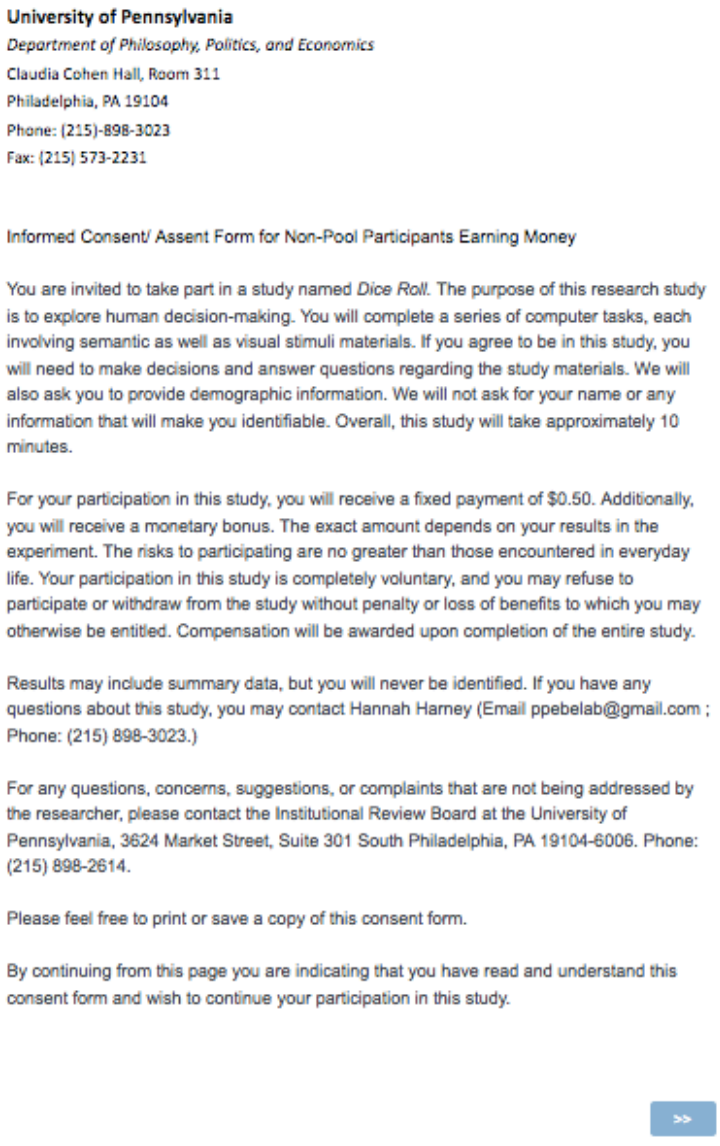
It is straightforward to see that conditions (5) and (6) are mutually incompatible. This implies that

Proposition 4 *In the other regarding condition, belief distortion never occurs.*

[6]The fact that warm-glow will be experienced immediately at the time of committing to giving follows Andreoni and Payne (2003). However, this assumption is not crucial to our formalization. We obtain the same results if the warm-glow boost instead occurs later as long as an individual likes oneself more than the charity, which is a reasonable assumption in most instances.

III. Experimental Screenshots

Exemplarily, we present below the screenshots for the empirical conditions. Normative treatment is identical, except for the statements phrased normatively. Screenshots from all treatment variations are available upon request.



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 will receive a monetary bonus. 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.

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

Is the dice you will roll recorded by the experimenter?

- ○ Yes
- ○ No

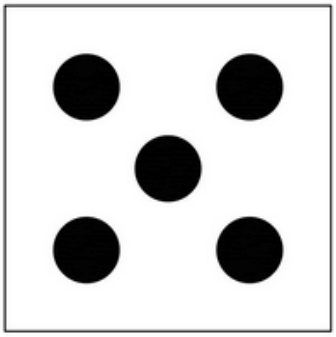
>>

Please read the following statements and determine whether you believe them to be true or false.
If you answer correctly you will earn $0.25 more in bonus payment.

Which statement is **TRUE**?

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

>>



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

What number did you roll?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6

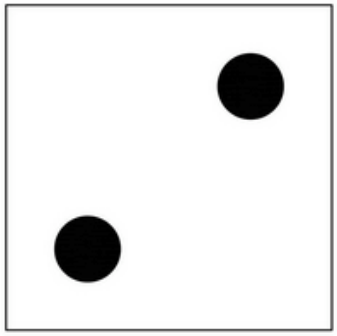
>>

Please read the following statements and determine whether you believe them to be true or false.
If you answer correctly you will earn $0.25 more in bonus payment.

Which statement is **TRUE**?

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

>>



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

What number did you roll?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6

>>

What is your gender?

- ○ Male
- ○ Female

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

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

How old are you in years?

>>

How do you see yourself:

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

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

0 1 2 3 4 5 6 7 8 9 10

Risk Willingness

What is your general opinion of charitable organizations?

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

0 1 2 3 4 5 6 7 8 9 10

General Opinion of Charities

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

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

0 1 2 3 4 5 6 7 8 9 10

General Opinion of UNICEF

>>