Lying opportunities and incentives to lie: Reference dependence versus reputation *

Eberhard Feess and Florian Kerzenmacher†

November 23, 2017

Abstract

Recent experiments on lying behavior show that the lying frequency in case of low outcomes increases in the ex ante probability of high outcomes. This finding is in line with models consisting of internal lying costs and external reputation costs, but also with models combining internal lying costs and loss aversion. To compare the explanatory power of these two approaches, we manipulate the ex-ante probability that lying is possible at all. We show theoretically that the reputation model predicts that the lying frequency decreases in the probability that lying is possible, while the loss aversion model suggests the opposite. Our experimental results strongly support the reputation model. From an applied perspective, our results suggest that safeguards for reducing the probability that lying is possible may (partially) backfire.

Keywords: dishonesty, truth-telling, lying, reputation models, loss aversion, experimental economics

JEL Classification: D03, D82, H26

*We are grateful to Simon Dato, Tim Friehe, Werner Güth, Jenny Kagl, Kai Konrad, Gerd Mühlheußer, Petra Nieken, Daniele Nosenzo, Jack Robles, Hannes Rusch, Elisabeth Schulte, Christoph Schumacher for very valuable comments. We also benefited from comments at the EEA annual congress (Lisbon, 2017), Verein für Socialpolitik Jahrestagung (Vienna, 2017), Economic Science Association European Meeting (Vienna, 2017), GEABA Symposium (Hohenheim, 2017) and during seminar presentations in Auckland, Frankfurt, Karlsruhe, Munich, Marburg and Wellington.

†Frankfurt School of Finance & Management, Sonnemannstr. 9-11, 60314 Frankfurt, Germany; Email: e.feess@fs.de and f.kerzenmacher@fs.de.

1 Introduction

In recent years, many laboratory experiments have investigated the degree of misreporting of private information (lying) in cases where such behavior increases the own payoff and can neither be observed nor punished. There is a large heterogeneity in behavior ranging from truth-telling over partial lying to payoff maximization, but subjects often leave about 70% of what they could get on the table (Abeler et al., 2017). Many theories including (internal) lying costs, (external) reputation costs, social conformity and loss aversion are suggested as explanations, but only a few papers aim at discriminating the explanatory power of those theories. We contribute to this growing literature by designing an experiment for which two prominent theories yield strictly contradictory predictions: The first theory is based on reputation costs and assumes that subjects face a disutility that increases in the lying probability assumed by (outside) observers.[1] The second theory assumes that subjects are loss averse where losses are defined with respect to reference points. Our experimental results support the reputation model and reject the loss aversion model.

Our analysis draws on papers by Abeler et al. (2017) (henceforth ANR, 2017), Garbarino et al. (2016) (henceforth GSV, 2016) and Gneezy et al. (2016) who all vary the success probability in lotteries. All three papers find that the lying frequency in case of low outcome increases significantly in the success probability. ANR (2017) and Gneezy et al. (2016) show that this is in line with reputation models, while models based only on internal lying costs predict no impact of the success probability. GSV (2016), however, argue that the same prediction as in reputation models arises in a model with loss aversion when the expected outcome of the lottery is taken as reference point. Both theories are thus compatible with the experimental findings in the three papers. The main value added of our setting is thus that the two theories make opposite predictions.

Specifically, we consider two treatments which differ only in the probability that lying is possible at all (q). We first show theoretically that the reputation model predicts

[1] To avoid misunderstandings it is worth being noted that "reputation" in the literature we draw on refers to preference costs arising from the assumed perception of outsiders, and has thus nothing to do with reputation effects of lying in dynamic games (for the latter string of literature see e.g. Charness et al. (2011) where reputation on being trustworthy positively affects other parties' behavior).

that the lying frequency in case of low outcome decreases in q, while the loss aversion model predicts that lying increases in q. A somewhat simplified intuition for the fact that reputation and loss aversion models yield similar predictions for variations in the success probability of lotteries (p) but contradictory predictions for variations in q proceeds as follows: In models with loss aversion, the behavior depends on the *own* expectations over outcomes. A higher value of p always increases the expected payoff, and the same holds for higher q whenever an individual assumes to lie with positive probability. In both cases, this upwards shift in the reference point increases the perceived monetary loss in case of the low outcome when individuals are loss averse. This leads to a higher incentive to lie. In contrast to the loss aversion model where only the own perceptions matter, incentives in the reputation model depend on the beliefs of *outsiders*. As a consequence, the impacts of p and q go in opposite directions: If the success probability is high, then outsiders infer a lower lying probability when the high outcome is reported.[2] Conversely, they tend to suspect a lie if the probability that lying is possible at all is large. Therefore, the disutility from lying in reputation models decreases in p but increases in q.

We run two treatments, one with a high ex-ante probability of 90% that lying is possible in case of low outcome (treatment H) and one where this probability is only 50% (treatment L). As lying is only an issue in those case where the outcome is low *and* lying is possible, many observations are needed for the statistical analysis. Therefore, we decided to perform an online- instead of a laboratory experiment and used Amazon Mechanical Turk (Mturk) to recruit and pay participants. 320 subjects participated in treatment H and 576 subjects in treatment L, so that the expected number of subjects who can lie, given by $(1-p)\,q$, is identical and equal to 216 in both treatments. We framed the decision situation in a neutral way and ensured that participants understood that their actual outcome is unobservable to the experimenter (see section 3 for details). Our results are clearly in favor of the reputation model: The lying frequency in case lying is feasible is 38% in treatment H compared to 50.5% in treatment L. This amounts to a decrease of about 25%; significant at the 1%-level in a χ^2-test, and is the opposite

[2]The effect is more subtle as the observer's belief depends on the equilibrium reports of all subjects, but the basic intuition prevails.

of what the loss aversion model predicts.

Our results underline the view that reputation issues play a crucial rule in lying behavior. ANR (2017), GSV (2016) and Gneezy et al. (2016) find that a higher success probability increases the lying frequency. We complement this finding by showing that a higher possibility that lying is possible reduces lying. One may hence conclude that theories on lying behavior should be consistent with evidence that altering these two probabilities influences the lying behavior in opposite directions. To the best of our knowledge, this is only predicted by reputation models: Models based on lying costs only (e.g. Ellingsen and Johannesson, 2004; Kartik, 2009; DellaVigna et al., 2017) predict invariance with respect to both manipulations; simply because a lie is a lie anyway. Next, in models with social conformity (e.g. Weibull and Villa, 2005; Charness and Dufwenberg, 2006; Gibson et al., 2013; Diekmann et al., 2015) where individuals feel less guilty when norms are violated by others we well, lying costs decrease in the expected frequency of lying by others. While an increase in p leads ceteris paribus to less lying and establishes truth-telling as a norm, the opposite holds for an increase in q. This model type thus predicts the opposite of what is observed. Note that our experimental design is conservative with respect to the reputation model: experiments on Mturk are anonymous, demand effects are likely to be lower compared to laboratory experiments and, most importantly, our design ensured that we get no information on whether the actual outcome was high or low.

The main purpose of our design is to get opposing predictions for the two theories in our horse race. Nevertheless, our setting also seems relevant from an applied perspective. Assume that a principal delegates the evaluation of two mutually exclusive projects A and B to an agent who derives a private benefit from project A. With probability q, the evaluation yields only soft information that is difficult to understand, so that the agent can use e.g. earnings manipulation techniques to display project A as superior (even though it is not). The principal knows ex-ante that this may be possible with probability q, but as the agent has superior knowledge, the principal cannot find out ex-post whether the agent could actually have misguided her, and hence follows the agent's advice (i.e. she delegates real authority in the terminology of Aghion and Tirole,

1997). Our experimental results then suggest that reducing the ex-ante probability that manipulations are possible partially backfires as it increases the frequency in case misreporting is possible.[3]

Concerning our theoretical models, two points should be noted; one for the loss aversion model and one for the reputation model: For the loss aversion model, we apply a full-fledged Kőszegi and Rabin (2006)-type model (henceforth KR-type model) which accounts for rational beliefs on strategies in the formation of reference points. This implies that subjects who are going to lie in case the outcome is low and lying is possible form different reference points than subjects who will tell the truth irrespective of the outcome. This approach seems to be the natural application of KR-type models to lying as we see no good reason why subjects who are always honest should expect to get the same monetary payoff as those who make use of their lying opportunity. If we followed GSV (2016) instead who assume that the lottery's expected outcome serves as reference point irrespective of the anticipated strategy, then q would have no impact at all. The reason is that, in the KR-type model, q influences the reference point in the lying strategy, but not in the truth-telling strategy.[4]

For the reputation model, the following point should be noted: A sufficient condition for an unambiguous prediction of the impact of q is that reporting the low outcome instead of the high outcome (usually referred to as downwards lying) is impossible. By contrast to all other lying experiments with unobservable outcomes we are aware of, we solve that problem by excluding downwards lying in the experimental design (see section 3 for details). Thus, we can also safely consider a reputation model without downwards lying in the theory part. Excluding downwards lying has the additional advantages that it allows for a more straightforward quantitative interpretation of the experimental results.

[3]In the conclusion, we discuss why our setting extends to the less extreme case in which the principal gets an additional signal on lying from observing the agent's report. We postpone this discussion to the conclusion as we first need to present the mechanics of the two theories.

[4]We consider the simple case where the lottery's expected outcome serves as reference point in the Appendix. Note that the prediction of no impact is also not supported by the data as the lying frequency decreases significantly in q. The same irrelevance result emerges in a Sugden (2003)-type model where reference points are only formed with respect to scenarios resulting from choices in the current opportunity set. In our setting, this would mean that reference points are formed only after participants have learned whether they can lie. Then, it is obvious that q has no impact.

Our study is most closely related to papers on lying costs that argue with either reputation for honesty or loss aversion. Besides ANR (2017), recent papers on reputation models include Mazar et al. (2008), Gneezy et al. (2016), Khalmetski and Sliwka (2017), Frankel and Kartik (2016) and Dufwenberg Jr. and Dufwenberg Sr. (2016), who refer to reputation costs as "perceived cheating aversion" , and who show that their model yields predictions in line with data in dice experiments based on Fischbacher and Föllmi-Heusi (2013).[5] Several papers find that the misreporting frequency is higher when outcomes are framed as losses rather than as gains (see Cameron and Miller, 2009; Grolleau et al., 2016; Schindler and Pfattheicher, 2017), which is in line with loss aversion models. While our results are far from claiming that loss aversion is meaningless in the context of reporting private information, they reinforce findings by ANR (2017) that reputation models fit the data best.

Similar to the papers by ANR (2017), GSV (2016) and Gneezy et al. (2016), our setting is particular simple in the sense that decisions do not influence the payoff of other subjects in the experiment. Important research taking those impacts into account include papers on guilt aversion, the impact of promises and the role of (monetary) incentives. Guilt aversion differs from pure lying costs as it depends on how other players are harmed by a lie. Specifically, Battigalli and Dufwenberg (2009) introduce guilt aversion that depends on i's belief on j's belief on how much j's payoff is reduced due to player i's lie (see also Battigalli et al., 2013). As lying in our experiment influences only the experimenter's payoff, guilt aversion is likely to be of lower importance compared to experiments where other subjects are affected. Papers that include promises investigate to which degree cheating and lying decreases if participants can promise to behave cooperatively or tell the truth (Charness and Dufwenberg, 2006; Sutter, 2009; Vanberg, 2008). Charness and Dufwenberg (2010) find that promises have no impact on trust, but do increase trustworthiness at least to some extent. Kajackaite and Gneezy (2017) show that higher benefits from lying lead to more lying in cases where subjects cannot be exposed as liars, which suggests that, by contrast to reputation costs, intrinsic lying

[5] We follow the general approach of Fischbacher and Föllmi-Heusi (2013) as the outcome is private information and as no individual report can be identified as a lie, but we apply a setting with only two outcomes as this is the easiest way for testing the hypotheses based on the two theories we are interested in. See the overview in ANR (2017) for other experiments with two outcomes only.

costs do not (strongly) increase in stakes. Testing this finding in our model would require re-running the experiment with different amounts.

Section 2 presents the model. The experimental design is introduced in section 3. Section 4 shows results and section 5 concludes.

2 The model

At $t = 0$, an agent enters a lottery which yields an entitlement to payoff $H > 0$ with probability p, and zero with $1 - p$. The lottery's outcome $e \in \{0, H\}$ is private information to the agent. In case of low outcome, the agent can report the high outcome with probability $q \in (0, 1)$. Both p and q are common knowledge at $t = 0$. After observing outcome e at $t = 1$, the agent reports an entitlement $m \in \{0, H\}$ at $t = 2$ and receives payoff $\pi = m$ at $t = 3$. Lying on the lottery's outcome is never detected but, depending on her type, the agent faces lying costs of $\theta_\ell \in [0, \theta_\ell^{\max}]$. θ_ℓ is distributed with continuous density $f(\theta_\ell)$ with full support on $[0, \theta_\ell^{\max}]$. $\theta_\ell^{\max} > H$ excludes that all agent types prefer to lie in case of low outcome.

Figure 1 summarizes the time line.

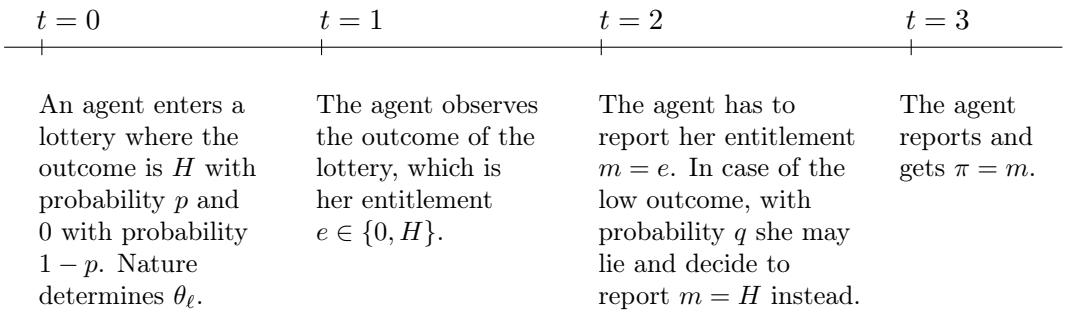

Figure 1: Time line

Note that the standard model of expected utility, amended by lying costs only, predicts that q has no impact, because the agent lies for $e = 0$ if and only if $H > \theta_\ell$.

2.1 Expectation based reference points and loss aversion

We start with the model that combines loss aversion with lying costs. Assume that the agent's utility depends on consumption c and on reference points (ref), which are determined by expectations. We follow Kőszegi and Rabin (2006) and Ericson and Fuster (2011) by distinguishing between a *consumption utility* $u_k(c_k)$ and a a *gain-loss utility* $u_k(ref_k)$.

Both utilities consist of two dimensions, the monetary dimension M and the lying dimension L; i.e. $k \in \{M, L\}$.[6] The monetary part of the consumption utility equals the payoff for all agent types, but lying costs depend on types; $u_L(c_L) \in \{-\theta_\ell, 0\}$. The agent's utility is $u(c|ref) = \sum_k u_k(c_k) + \mu(u_k(c_k) - u_k(ref_k))$, where $u_k(c_k) - u_k(ref_k)$ is the difference between the consumption utility in dimension k and the reference point determined by expectations. Thus, $u_k(c_k)$ is the consumption utility and $\mu(u_k(c_k) - u_k(ref_k))$ the gain-loss utility.

To capture loss aversion as simple as possible, assume $\mu\left(u_k(c_k) - u_k(ref_k)\right) = \eta_k(u_k(c_k) - u_k(ref_k))$ for $u_k(c_k) - u_k(ref_k) > 0$ and $\mu(u_k(c_k) - u_k(ref_k)) = \eta_k\lambda_k(u_k(c_k) - u_k(ref_k))$ for $u_k(c_k) - u_k(ref_k) \leq 0$, where $\eta_k \geq 0$ and $\lambda_k \geq 1$. For $\eta_k = 0$, the model boils down to expected utility theory with lying costs, and for $\eta_k > 0$ but $\lambda_k = 1$, there is no loss aversion. In the main part of the model, we assume that reference points are equally important for monetary payoffs and lying ($\eta := \eta_M = \eta_L$).[7] In the Appendix, we show that our results on the impact of q also hold for the (plausible) special case considered by GSV (2016), where gains and losses are only evaluated in the monetary dimension, while lying costs are independent of expectations ($\eta_L = 0$).

2.1.1 Lying strategy

By definition of a lying strategy, the agent reports $m = H$ when possible. To see how the gain-loss utility is calculated in our setting, let us consider each possible outcome in the lying strategy in turn.

[6] Referring to lying costs as part of the *consumption utility* may sound irritating; but we want to follow the literature's terminology in case of two utility-dimensions as closely as possible.

[7] However, we always allow for different degrees of loss aversion; $\lambda_M \lessgtr \lambda_L$.

- With probability pq, the outcome is high and lying is possible. In this case, consumption utilities are $c^{H,lp} = \{H, 0\}$[8] as the agent receives the high outcome H without lying;

- With $p(1-q)$, the outcome is high and lying is impossible, which again gives $c^{H,lnp} = \{H, 0\}$;[9]

- With $(1-p)q$, the outcome is low and lying is possible, which gives $c^{0,lp} = \{H, -\theta_\ell\}$;

- With $(1-p)(1-q)$, the outcome is low and lying is impossible, which gives $c^{0,lnp} = \{0, 0\}$.

As there is no rationale for downwards lying in the loss aversion-model, incentive compatibility of the lying strategy refers to the case where the outcome is low and lying is possible. Assume the outcome is low. Then, the agent's utility when sticking to her lying strategy is[10]

$$u_L^L(m = H|e = 0) = H - \theta_\ell - \eta\lambda_L \left(1 - (1-p)\, q\right) \theta_\ell + \eta(1-p)(1-q)H. \qquad (1)$$

Thereby, $H - \theta_\ell$ is the consumption utility, i.e. the monetary payoff minus type-specific lying costs. Next, with probability $(1-p)q$, the agent ex-ante expects that she lies, so that *expected* lying costs are $(1-p)q\theta_\ell$. As *actual* lying costs are θ_ℓ, the difference is $(1-(1-p)q)\theta_\ell$, and this utility loss beyond the expected loss needs to be weighted with $\eta\lambda_L$. Similarly, the agent expected to *get* the high output only with probability $p+(1-p)q$, so that $(1-(p+(1-p)q))H = (1-p)(1-q)H$ is the unexpected utility gain in the monetary dimension, which needs to be weighted by η.

If the agent deviates from her original lying-plan to truth-telling when lying is possible and after observing the low outcome, applying the same reasoning yields utility[11]

$$u_L^T(m = 0|e = 0) = (1-p)q\eta\theta_\ell - (p + (1-p)q)\lambda_M\eta H: \qquad (2)$$

[8] "lp" denotes the case where lying is possible, while "lnp" below denotes the case where lying is not possible.

[9] If the outcome is high, the consumption utility is independent of whether lying is possible or not.

[10] Subscripts express the original plan (L for Lying) and superscripts the actual behavior.

[11] Superscript "T" stands for truth-telling.

By deviating to truth-telling, the agent does not only save her lying costs θ_ℓ in the consumption utility, but has an additional gain from not lying as she expected to lie with probability $(1-p)q$. This gain is thus $(1-p)q\eta\theta_\ell$. In the monetary dimension, however, the agent does not only lose her monetary consumption utility H, but faces an additional utility cost from reference point violation and loss aversion. This additional cost amounts to $(p+(1-p)q)\lambda_M\eta H$ as the agent expected to follow her lying strategy, and hence to get the high output with probability $p+(1-p)q$.

A strategy is a personal equilibrium if and only if it is ex-post incentive compatible, i.e. if the agent has no incentive to deviate from the strategy *after* observing a specific situation. In our setting, the only case to be considered is where the outcome is low and lying is possible. Thus, lying is a personal equilibrium, if and only if $u_L^L(m=H|e=0) \geq u_L^T(m=0|e=0)$. After rearrangement, this condition can be written as

$$\theta_\ell \leq \overline{\theta_\ell} := \frac{1+\eta\lambda_M(p+q(1-p))+\eta(1-p)(1-q)}{1+\eta\lambda_L(1-(1-p)q)+q\eta(1-p)}H,$$

where $\overline{\theta_\ell}$ denotes the critical threshold level for lying costs which only just support the lying strategy as a personal equilibrium. Observe that $\overline{\theta_\ell}$ is strictly increasing in λ_M and strictly decreasing in λ_L due the fact that loss aversion in the lying dimension (λ_L) matters only if the agent sticks to the lying strategy, while loss aversion in the monetary dimension (λ_M) matters only when she deviates to truth-telling.[12] Furthermore, $\overline{\theta_\ell}$ is strictly increasing in q as the agent expects the high monetary payoff with high probability, so that the losses from reference point violation based on expectations in the monetary dimension when deviating to the truth-telling strategy are higher. Thus, the lying strategy is a personal equilibrium for more types if the ex-ante probability that lying is possible is high.

Let us now consider types $\theta_\ell \leq \overline{\theta_\ell}$ for which the lying strategy is incentive compatible, i.e. for which lying is a personal equilibrium. The *ex-ante* expected utility from the

[12] To see this, recall that, for incentive compatibility, we only need to consider the case where lying is possible.

lying strategy is then

$$\begin{aligned} u_L =& pH + (1-p)q(H-\theta_\ell) \\ &+ p\left((1-p)q\eta\theta_\ell + (1-p)(1-q)\eta H\right) \\ &+ (1-p)q\left(-p\eta\lambda_L\theta_\ell + (1-p)(1-q)\eta(H-\lambda_L\theta_\ell)\right) \\ &+ (1-p)(1-q)\left(-p\eta\lambda_M H + (1-p)q\eta(\theta_\ell - \lambda_M H)\right) \end{aligned} \tag{3}$$

The first line captures the consumption utility: With $p+(1-p)q$, the agent receives the high payoff, and with $(1-p)q$, she lies and incurs lying costs. All other parts summarize the additional gain-loss utilities from comparing the consumption utility to reference points. For instance, with probability p, the agent gets the high outcome without lying, and if so, this yields a utility gain of $(1-p)q\eta\theta_\ell$ in the lying dimension as the agent expected to lie with $(1-p)q$, and a utility gain of $(1-p)(1-q)\eta H$ in the monetary dimension as she expected not to get the high outcome with $(1-p)(1-q)$ (second line). The other terms can be interpreted accordingly.

2.1.2 Truth-telling strategy

Suppose next that the agent plans to report the truth. Then, the following situations can occur:

- With probability p, the outcome is high, which yields consumption levels $c^H = \{H, 0\}$;
- With probability $1-p$, the outcome is low, which yields consumption levels $c^0 = \{0, 0\}$.

The only interesting case for incentive compatibility arises again when the outcome is low and lying is possible. When the agent sticks to her plan, her utility is

$$u_T^T(m=0|e=0) = -p\eta\lambda_M H \tag{4}$$

as she expected to get the high outcome with probability p. When deviating to lying,

her utility is:

$$u_T^L(m = H|e = 0) = H - \theta_\ell - \eta\lambda_L\theta_\ell + (1-p)\eta H. \tag{5}$$

In addition to the consumption utility $H-\theta_\ell$, the agent faces costs of $\eta\lambda_L\theta_\ell$, because she expected not to lie with probability one, but also gets the additional gain of $(1-p)\eta H$. Truth-telling is a personal equilibrium if $u_T^T(m = 0|e = 0) \geq u_T^L(m = H|e = 0)$, which can be written as:

$$\theta_\ell \geq \underline{\theta_\ell} := \frac{1 + \eta(1 + p(\lambda_M - 1))}{1 + \eta\lambda_L} H, \tag{6}$$

where $\underline{\theta_\ell}$ denotes the critical threshold for lying costs which only just support incentive compatibility for truth-telling. $\underline{\theta_\ell}$ is strictly increasing in λ_M and strictly decreasing in λ_L for reasons similar to those explained for the lying-equilibrium. $\underline{\theta_\ell}$ is strictly increasing in p for $\eta > 0$. Most important in our context, $\underline{\theta_\ell}$ is independent of q for two reasons: First, if the agent tells the truth anyway, then the ex-ante probability that lying is possible has no impact on her utility. Second, incentive compatibility refers to the case where it has already turned out whether lying is feasible or not. By contrast, recall that the condition for lying to be a personal equilibrium depends on q as the agent expects to lie with probability $(1-p)q$.

The agent's expected utility from the truth-telling strategy is

$$\begin{aligned} u_T &= pH - (1-p)p\eta H\lambda_M + p(1-p)\eta H \\ &= pH - (1-p)p\eta H(\lambda_M - 1). \end{aligned}$$

Thereby, pH is the consumption utility, while the other two terms capture the gain-loss utilities: with probability $1-p$, the agent does not get the high outcome, and this yields losses of $p\eta H\lambda_M$ as she expected to get it with probability p. Analogously, she expected not to get the high outcome with $1-p$, so that her additional utility in case of high outcome (which happens with p) is $(1-p)\eta H$.

Considering the incentive compatibility constraints for the lying strategy and the truth-telling strategy, we find that $\overline{\theta_\ell} > \underline{\theta_\ell}$ for $\eta > 0$, i.e. whenever reference points matter at all. This ensures that at least one of the two pure strategies is a personal

equilibrium:

Proposition 1 (Expectation based reference points with loss aversion)**.** *There exists at least one personal equilibrium in pure strategies. For* $\underline{\theta_\ell} \leq \theta_\ell \leq \overline{\theta_\ell}$*, both the lying and the truth-telling strategies are personal equilibria. For* $\theta_\ell > \overline{\theta_\ell}$*, only truth-telling is a personal equilibrium. For* $\theta_\ell < \underline{\theta_\ell}$*, only lying is a personal equilibrium.* $\underline{\theta_\ell}$ *is independent of* q*, while* $\overline{\theta_\ell}$ *increases in* q*.*

Proof. See Appendix.

Proposition 1 implies that, all types with $\theta_\ell < \underline{\theta_\ell}(q, \cdot)$ will lie independently of q, while all types with $\theta_\ell > \overline{\theta_\ell}(q, \cdot)$ will tell the truth independently of q. As θ_ℓ is continuously distributed and because $\frac{\partial \overline{\theta_\ell}}{\partial q} > 0$, the fraction of agents for which only truth-telling is a personal equilibrium decreases in q. For all types with $\underline{\theta_\ell} \leq \theta \leq \overline{\theta_\ell}$, the behavior depends on which of the two strategies provides the higher expected utility. Thus, in order to derive the agent's preferred personal equilibrium, we next need to compare the payoffs of the lying and the truth-telling strategy under the assumption that $\underline{\theta_\ell} \leq \theta_\ell \leq \overline{\theta_\ell}$.

2.1.3 Preferred personal equilibrium

Assume that both personal equilibria exist, i.e. $\underline{\theta_\ell} \leq \theta_\ell \leq \overline{\theta_\ell}$ for all types θ_ℓ subsequently considered. Then, lying is the preferred personal equilibrium if and only if the expected utility of the lying strategy (weakly) exceeds the expected utility of the truth-telling strategy, which can be written as

$$\theta_\ell \leq \widehat{\theta}_\ell := \frac{1 - \eta(\lambda_M - 1)(1 - q - p(2 - q))}{1 + \eta(\lambda_L - 1)(1 - (1 - p)q)} H. \tag{7}$$

This yields the following Proposition:

Proposition 2 (Expectation based reference points with loss aversion)**.** *Suppose the low outcome is realized, lying is possible and both the lying and the truth-telling strategies are personal equilibria. Then, the critical value of lying costs* $\widehat{\theta}_\ell$ *such that lying is the preferred personal equilibrium increases in* q*.*

Proof. See Appendix.

Proposition 2 says that, for all combinations of loss aversion λ_L and λ_M, the lying strategy is more likely to offer higher utility than the truth-telling strategy when the ex-ante probability that lying is possible is high. While the expected utility in the truth-telling strategy is independent of q as the agent will not lie anyway, the expected consumption utility in the lying strategy increases in q as the agent gets H more often. Furthermore, when H is realized, then the perceived loss when deviating to truth-telling increases in q. There is a countervailing effect as the perceived monetary loss in case neither the outcome is high nor lying is possible is larger when q is high, but this effect cannot dominate.

In the proof of Proposition 2, we show in addition that the comparative statics for the probability of winning the lottery (p) is less clear than in the simplified loss aversion model adopted by GSV (2016), who take the expected outcome as reference point and assume no loss aversion in the lying dimension. The higher p, the higher is the expected monetary gain in both strategies. In the lying strategy, lying is less often required, which also leads to higher expected utility. The shift in reference points caused by larger p leads to larger perceived losses in the truth-telling strategy in the monetary dimension (weighted by λ_M) and to larger perceived losses in the lying strategy in the lying dimension (weighted by λ_L). If loss aversion from lying is weakly lower than loss aversion in monetary outcomes ($\lambda_L \leq \lambda_M$), then an increase in p is more beneficial in the lying strategy than in the truth-telling strategy. Thus, $\lambda_L \leq \lambda_M$ is a sufficient condition for $\frac{\partial \widehat{\theta_\ell}}{\partial p} > 0$. In this sense, the result in GSV (2016) emerges as the special case of our model. For $\lambda_L > \lambda_M$, however, the impact of p depends on all parameters of the model.

Recall from Proposition 1 that the lying strategy is a personal equilibrium for more types when q is high, i.e. $\frac{\partial \overline{\theta_\ell}}{\partial q} > 0$. Taking Propositions 1 and 2 together, we get the following testable Hypothesis:

Hypothesis 1 (expectation based reference points)**.** *The lying frequency in case the outcome is low and lying is possible increases in the ex-ante probability that lying is possible.*

2.2 Reputation model

In the reputation model, the agent's utility consists of monetary payoffs, lying costs and reputation concerns. Lying costs $\theta_\ell \in [0, \theta_\ell^{\max}]$ are the same as before. Reputation concerns depend on the agent's type θ_r and on the lying probability r inferred by others after observing report m. Reputation costs $\theta_r \in [0, \theta_r^{\max}]$ are distributed with continuous density $g(\theta_r)$ with full support on $[0, \theta_r^{\max}]$. $f(\theta_\ell)$ and $g(\theta_r)$ are common knowledge, and lying costs and reputation costs are independently distributed. $\theta_r^{\max} > H$ excludes that all agent types with lying costs $\theta_\ell = 0$ lie after observing the low outcome.

As we exclude downwards lying (by assumption in the model and by design in the experiment), $r(m = 0) = 0$. For the high report, r depends on the probability p for the high outcome in the lottery, on the probability q that lying is possible and on the fraction of types lying in equilibrium. Assuming separable reputation for honesty plus lying costs as in ANR (2017), the agent's utility is

$$u_R = m - \theta_\ell l(m, e) - \theta_r \mathbb{E}_{\theta_\ell \theta_r}[v(r)].$$

where l is an index variable that takes the value 1 if $m \neq e$ and zero otherwise. Reputation costs $v(r)$ are strictly increasing in the lying probability r inferred by an "outsider", who does not take any action but observes the agent's report. The outsider correctly updates the lying probability according to Bayes' rule for a given report. As r depends on the lying behavior of all agents, each agent needs to consider the equilibrium behavior of all types and needs to take expectations over types. We assume $v(r = 0) = 0$ and normalize $v(r = 1) = 1$ without loss of generality.

2.2.1 Equilibrium

The only interesting case is again that an agent has lost the lottery and has the opportunity to lie. We consider a psychological game as in Battigalli and Dufwenberg (2009), because beliefs on the beliefs of others are a direct component of utility. As each agent forms expectations over types and beliefs, we apply the Sequential Equilibrium as solution concept to the psychological game (see also ANR, 2017). Each agent type (and the

outsider) treat the probability distribution over r as given. Formally, the expectation of r over types is:

$$\begin{aligned}\mathbb{E}_{\theta_\ell\theta_r}[r] &= \frac{\Pr(m>e)q(1-p)}{\Pr(m>e)q(1-p)+p},\\ \Pr(m>e) &= \frac{\int_0^{\theta_\ell^{\max}}\int_0^{\theta_r^{\max}} f(\theta_\ell)g(\theta_r)\mathbb{1}_{u_R^L>u_R^T}d\theta_r d\theta_\ell}{\int_0^{\theta_\ell^{\max}}\int_0^{\theta_r^{\max}} f(\theta_\ell)g(\theta_r)d\theta_r d\theta_\ell},\end{aligned} \tag{8}$$

where $\mathbb{1}_{u_R^L>u_R^T}$ is an indicator function that is equal to 1, if the utility from lying $u_R^L := u_R(m=H|e=0)$ is higher than the utility from telling the truth $u_R^T := u_R(m=0)$, and 0 otherwise. Therefore, if and only if the utility from lying is higher than the utility from truth-telling, the agent is expected to lie.

For any r given, an agent's utility from lying decreases in her direct lying costs θ_ℓ and her reputation concerns θ_r. Our assumptions on the continuity of θ_ℓ and θ_r as well as $H < \theta_\ell^{\max}, \theta_r^{\max}$ ensure that there exist indifferent types such that lying and truth-telling yield the same utility.[13] As the utility from truth-telling in case of low outcome is zero, $u_R^L = u_R^T$ is

$$H - \theta_\ell - \theta_r \mathbb{E}_{\theta_\ell\theta_r}[v(r)] = 0. \tag{9}$$

As an agent's behavior depends on both θ_ℓ and θ_r, an equilibrium is characterized by the set of combined threshold types $\widehat{\theta_\ell}$ and $\widehat{\theta_r}$ that solve Equation (9). We can think of the threshold as a function $\widehat{\theta_\ell}(\theta_r) = \max\{H - \theta_r\mathbb{E}_{\theta_\ell\theta_r}[v(r)], 0\}$, such that all types with $\theta_\ell < \widehat{\theta_\ell}(\theta_r)$ lie, while all types with $\theta_\ell \geq \widehat{\theta_\ell}(\theta_r)$ tell the truth. As the utility function is linear in θ_ℓ and θ_r, the threshold function $\widehat{\theta_\ell}(\theta_r)$ will be linear as well (see Figure 2).

Due to linearity, the threshold function can be characterized by the intercepts $\overline{\theta_\ell} := \widehat{\theta_\ell}(0) = H$ and $\overline{\theta_r} := \widehat{\theta_\ell}^{-1}(0)$, i.e. where either lying costs or reputation costs are zero. As $\overline{\theta_\ell}$ depends only on the monetary payoff and all agents will thus have the same intercept, the equilibrium can be characterized by the intercept $\overline{\theta_r}$. An agent will choose this intercept as best response to the intercepts chosen by other agents. In equilibrium,

[13] Since types are continuously distributed, any particular type – and hence also the indifferent type – has measure zero. Thus, almost all agents play a pure strategy. Any single deviation from the equilibrium strategy will be interpreted as a "mistake" by all other players and does therefore not influence equilibrium beliefs.

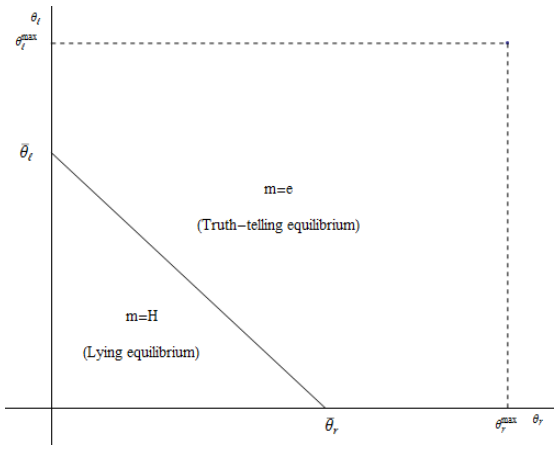

Figure 2: Equilibrium threshold

all agents' intercepts are identical as, with $\theta_\ell = 0$, everyone has the same threshold value for reputational concerns which only just triggers lying as the preferred strategy. Given the existence of indifferent types, an equilibrium exists where some types lie and some types tell the truth. In the Appendix, we prove hat the equilibrium is unique as there exists a unique intercept $\overline{\theta_r}$.

Lemma 1. *Suppose downwards lying is excluded. Then, there exists a unique* $\overline{\theta_r}$*. The lying probability inferred by outsiders is strictly increasing in* $\overline{\theta_r}$*.*

Proof. See Appendix.

From Lemma 1, we immediately get the result that the lying frequency decreases in the ex-ante probability q that lying is possible:

Proposition 3 (Reputation model)**.** *Suppose the low outcome is realized and lying is possible. Then, the critical value of lying costs* $\widehat{\theta}_\ell$ *decreases in* q*.*

Proof. See Appendix.

The basic intuition for Proposition 3 is as follows: Assume hypothetically that the threshold $\widehat{\theta}_\ell$ were independent of q. Then, an outside observer assigns a higher lying probability r after observing the high report as all types below $\widehat{\theta}_\ell$, and because lying is more often feasible. This cannot be an equilibrium as the utility decreases in r, so that less types have an incentive to lie. Thus, the critical type $\widehat{\theta}_\ell$ needs to decrease in q. In addition to our main result, we prove in the Appendix that, for all q, the reputation model predicts that the lying frequency increases in the success probability, i.e. $\frac{\partial \widehat{\theta}_\ell}{\partial p} > 0$.

Hypothesis 2 (Reputation for honesty)**.** *The lying frequency in case the outcome is low and lying is possible decreases in the ex-ante probability that lying is possible.*

Summing up, the analysis shows that the Kőszegi and Rabin (2006) type model and the reputation model, both amended by lying costs, yield strictly opposing predictions for the impact of q.

3 Experimental Design

For the reasons mentioned in the introduction, we decided to run an online – instead of a laboratory experiment. We created a Website written in PHP and used Amazon Mechanical Turk (Mturk) to recruit and to pay participants. The show-up fee was \$0.30 and a lottery determined whether participants were entitled to an additional \$0.30. The success probability was $p = 25\%$. The reason why we did not choose a higher probability is that this would have further reduced the number of participants who would be in the position to lie. Conditional on losing the lottery, there was a probability of $q_H = 90\%$ (treatment H) or $q_L = 50\%$ (treatment L) that participants would have the opportunity to lie. Again, we chose rather high numbers to increase the percentage of subjects that could lie.

The participants were informed that they receive one of two codes, a winning code or a losing code. In case of winning the lottery, the only thing participants had to do is to enter the winning code in order to receive \$0.60 in total. As they did not even learn the losing code, there was no possibility of downwards lying. Participants were clearly informed that, when they received the losing code, they are only entitled to the participation fee of \$0.30. However, they also learned that, with probability q (which was announced as 90% or 50%; depending on the treatment), they would nevertheless also receive the winning code before entering a code. Thus, they could enter the winning instead of the losing code even though they lost the lottery. In this case, they received the full payment of \$0.60.

We framed the instructions neutrally: Participants learned that they were only entitled to \$0.30 in case of losing the lottery, but we avoided any reference to lying or

other normatively loaded expressions (see the Appendix for the exact wording of the instructions). Furthermore, the experimental design and procedure ensured that the participants understood that they could not be tracked, i.e. that their actual outcome is unobservable. We conducted each session at roughly the same time of the day (between 8:30pm and 11:30pm Central European Time) on weekdays. Furthermore, we changed the winning and the losing codes for each session in order to prevent participants from communicating this information to each other.

The experiment was conducted in five sessions in January and February 2017. 320 subjects participated in treatment H and 576 subjects in treatment L, so that the expected number of subjects who can lie is identical ($n_i\,(1-p)\,q_i = 216;\ i = H, L.$) in both treatments. On average, it took a participant about a minute to complete the experiment, and average earnings where \$0.44.[14]

4 Results

As mentioned, there were 320 participants in treatment H and 576 participants in treatment L. In expectation, this yields 216 participants in each treatment who lost the lottery and had the opportunity to misreport the outcome. In treatment H, 162 individuals reported the high outcome and 158 individuals reported the low outcome. In treatment L, 253 individuals reported the high outcome, while 323 individuals reported the low outcome. Given that subjects lost the lottery and were able to lie, Figure 3 displays their decision.

As seen in Figure 3, the lying probability in case of low outcome when lying was possible was 50.5% in treatment L and 38.0% in treatment H. This difference is significant at the 1%-level in a χ^2-test with $p = 0.0089$)[15]. Thus, our findings support Hypothesis 2 and violate Hypothesis 1.

Model types based on expectation dependent reference points and loss aversion thus

[14] ANR (2017) find that the lying behavior in the setting where individual lies cannot be detected is robust with respect to payoffs, so that our low-stake experiment on Mturk does not seem to be problematic.

[15] The more conservative Fisher exact test yields a p-value of 0.0117, while a two-tailed population proportions tests yields a p-value of 0.0088.

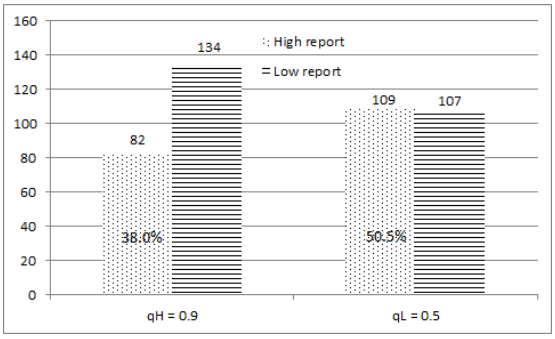

Figure 3: Reports given by individuals who were able to misreport.

predict the opposite of what is observed. Note that our experimental design is conservative with respect to the reputation model: Experiments on Mturk are anonymous, demand effects are likely to be lower compared to laboratory experiments and, most important, our design ensured that we get no information on whether the actual outcome was high or low.

While we are not able to determine how serious each individual took the task of playing the lottery and reporting the outcome, we can track that completing the task took 57 seconds on average. Assuming that the completion time is a good proxy for how careful participants read the instructions and understood the setup, Table 1 displays our results after excluding observations below several thresholds for completion time.

Minimum Time	10	20	30	40	50	60
$q_H = 0.9$	38.4%	38.5%	38.1%	40.3%	42.5%	42.3%
	N=318	N=308	N=270	N=226	N=175	N=140
$q_L = 0.5$	50.6%	49.8%	52.0%	55.5%	57.8%	62.5%
	N=571	N=545	N=454	N=345	N=255	N=192
p-value	0.0111	0.0209	0.0088	0.0108	0.0257	0.0096

Table 1: Lying frequencies excluding observations below several thresholds for the completion times in seconds (p-values according to two-tailed population proportions tests).

In both treatments, the fraction of liars increases slightly when we exclude lower completion times. However, the difference in lying frequencies between treatments remains significant at the 5%-level for all thresholds.

In order not to influence participants' reputational concerns when making lying decisions, we deliberately designed the experiment to be completely anonymous. Therefore, we did not collect any personal information such as gender or age.

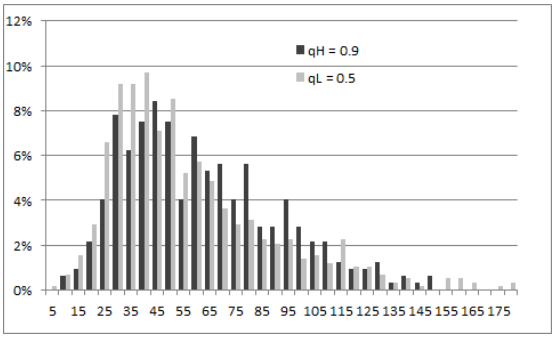

Figure 4: Decision times for each treatment in seconds.

5 Conclusion

Many laboratory and field experiments have shown that subjects do often not maximize their monetary payoffs in lying games; even without any risk of being observed or punished. It is less clear, however, which theories are best suited for understanding the underlying motives. Manipulating the success probability p in lotteries has shown that approaches based on (external) reputation costs and loss aversion-models, combined with direct (internal) lying costs are both in line with the data. Against this backdrop, we manipulate the probability q that lying is possible at all. By contrast to variations in p, reputation models and loss aversion models yield strictly contradictory predictions for variations in q. We first show theoretically that the reputation model without the possibility of downwards lying predicts that the lying frequency (in case lying is feasible and the outcome is low) decreases in q, while the loss aversion model predicts the opposite. We then perform an online experiment and find that the lying frequency is about 25% lower in the treatment with higher q, which strongly supports the reputation model and contradicts the loss aversion model.

Of course, we are aware that the performance of different theories for explaining lying behavior may be situation-specific. Thus, we do not claim that our findings indicate that reputation models generally outperform loss aversion models. Most naturally, reputation concerns and loss aversion may both play a role and may often (as in the case of manipulating the success probability) go in the same direction; thereby reinforcing the effects we observe. In our setting, however, predictions are strictly contradictory, and findings are only in line with reputation models. An important point to note is that our

experimental design is conservative in the sense that it is likely to under- rather than to overestimate reputation effects. First, the design ensured that participants understood that their outcome cannot be observed by other participants or the experimenter, so that lying can neither be observed nor punished. Thus, participants did know that there is no risk that lying may cause any real negative effects, i.e. reputation effects are as "indirect" as possible. Furthermore, experiments on Mturk are anonymous, demand effects are likely to be lower compared to laboratory experiments and, most importantly, our design ensured that we get no information on whether the actual outcome was high or low. Thus, one might tentatively suspect that the effect of q found in our online-experiment may be reinforced in laboratory experiments due to higher demand effects.

Except for reputation and loss aversion, one might think about other reasons why variations in q could influence the lying frequency in case lying is actually possible. As discussed in the introduction, herding models would predict that higher q leads to more lying, and models with only internal lying costs yield no impact of q. In addition, one could argue that subjects anticipate that higher q leads to higher overall payouts, so that participants with prosocial preferences lie less in order to save the experimenter's money. Such a reasoning, however, would also suggest that the lying frequency decreases in the success probability (p), which is the opposite of what is observed in the laboratory.

To motivate our setting also from an applied perspective, our introduction introduced an example in which principals delegate evaluations or decisions to partially selfish agents. Principals and agents have the same prior on the probability that agents can manipulate the outcome of their actual evaluation. The importance of our example might be questioned by arguing that principals often get an additional update about whether the agent effectively had the possibility to lie (or did actually lie) by observing her report (i.e. they might update q after observing the report). However, our setting carries over to this case: for reference dependence with loss aversion, the agent does not care about the principal's belief anyway, so that nothing changes.[16] Thus, the hypothesis that the lying frequency increases in q prevails. By contrast, the reputation model predicts that the lying frequency decreases even further if lying itself yields to

[16]This requires that there is no punishment potential. Punishment, however, reduces the incentive to lie in both theories and is excluded in most lying experiments in order to not confound the analysis with standard incentives from payoff maximization.

an update on the actual lying possibility from q to $\widetilde{q} > q$.[17]

This given, our experimental finding that subjects lie less often when the probability that they can do so increases appears to be interesting also from an applied perspective. In particular, it provides further evidence on hidden costs of control (see Falk and Kosfeld, 2006) as that reducing the probability that lying is possible through monitoring might reduce the direct benefits of control. While our experiment does not make any claims concerning the underlying motives of the observed behavior, reputation seems to be a natural candidate.

[17]The higher $\widetilde{q}$, the higher is r (see Proposition 3) and the lower is thus the threshold $\widehat{\theta}_\ell$.

References

Abeler, J., D. Nosenzo, and C. Raymond (2017, August). Preferences for Truth-Telling. SSRN Scholarly Paper ID 2866381, Social Science Research Network, Rochester, NY.

Aghion, P. and J. Tirole (1997, February). Formal and Real Authority in Organizations. *Journal of Political Economy 105*(1), 1–29.

Battigalli, P., G. Charness, and M. Dufwenberg (2013, September). Deception: The role of guilt. *Journal of Economic Behavior & Organization 93*, 227–232.

Battigalli, P. and M. Dufwenberg (2009, January). Dynamic psychological games. *Journal of Economic Theory 144*(1), 1–35.

Cameron, J. S. and D. T. Miller (2009). Ethical Standards in Gain versus Loss Frames. In *David de Cremer (Ed.): Psychological Perspectives on Ethical Behavior and Decision Making*, pp. 91 – 106. Charlotte, NC: Information Age Publishing.

Charness, G., N. Du, and C.-L. Yang (2011, June). Trust and trustworthiness reputations in an investment game. *Games and Economic Behavior 72*(2), 361–375.

Charness, G. and M. Dufwenberg (2006, November). Promises and Partnership. *Econometrica 74*(6), 1579–1601.

Charness, G. and M. Dufwenberg (2010, May). Bare promises: An experiment. *Economics Letters 107*(2), 281–283.

DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2017, January). Voting to Tell Others. *The Review of Economic Studies 84*(1), 143–181.

Diekmann, A., W. Przepiorka, and H. Rauhut (2015, August). Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. *Rationality and Society 27*(3), 309–333.

Dufwenberg Jr., M. and M. Dufwenberg Sr. (2016, December). Lies in Disguise - A Theoretical Analysis of Cheating. SSRN Scholarly Paper ID 2897478, Social Science Research Network, Rochester, NY.

Ellingsen, T. and M. Johannesson (2004, April). Promises, Threats and Fairness. *The Economic Journal 114*(495), 397–420.

Ericson, K. M. M. and A. Fuster (2011, November). Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments. *The Quarterly Journal of Economics 126*(4), 1879–1907.

Falk, A. and M. Kosfeld (2006, December). The Hidden Costs of Control. *The American Economic Review 96*(5), 1611–1630.

Fischbacher, U. and F. Föllmi-Heusi (2013, June). Lies in Disguise – an Experimental Study on Cheating. *Journal of the European Economic Association 11*(3), 525–547.

Frankel, A. and N. Kartik (2016). Muddled Information. mimeo.

Garbarino, E., R. Slonim, and M. C. Villeval (2016, December). Loss Aversion and Lying Behavior: Theory, Estimation and Empirical Evidence. SSRN Scholarly Paper ID 2875989, Social Science Research Network, Rochester, NY.

Gibson, R., C. Tanner, and A. F. Wagner (2013, February). Preferences for Truthfulness: Heterogeneity among and within Individuals. *The American Economic Review 103*(1), 532–548.

Gneezy, U., A. Kajackaite, and J. Sobel (2016, October). Lying Aversion and the Size of the Lie. SSRN Scholarly Paper ID 2852055, Social Science Research Network, Rochester, NY.

Grolleau, G., M. G. Kocher, and A. Sutan (2016, January). Cheating and Loss Aversion: Do People Cheat More to Avoid a Loss? *Management Science 62*(12), 3428–3438.

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

Kartik, N. (2009, October). Strategic Communication with Lying Costs. *The Review of Economic Studies 76*(4), 1359–1395.

Khalmetski, K. and D. Sliwka (2017). Disguising Lies – Image Concerns and Partial Lying in Cheating Games. mimeo.

Kőszegi, B. and M. Rabin (2006, November). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

Mazar, N., O. Amir, and D. Ariely (2008, December). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research 45*(6), 633–644.

Schindler, S. and S. Pfattheicher (2017, March). The frame of the game: Loss-framing increases dishonest behavior. *Journal of Experimental Social Psychology 69*, 172–177.

Sugden, R. (2003, August). Reference-dependent subjective expected utility. *Journal of Economic Theory 111*(2), 172–191.

Sutter, M. (2009, December). Individual Behavior and Group Membership: Comment. *American Economic Review 99*(5), 2247–2257.

Vanberg, C. (2008, November). Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica 76*(6), 1467–1480.

Weibull, J. W. and E. Villa (2005). Crime, punishment and social norms. Working Paper 610, SSE/EFI Working Paper Series in Economics and Finance.

Appendix

Proof of Proposition 1

Proof. From the main text, we know that lying is a personal equilibrium, if and only if $\theta_\ell \leq \overline{\theta_\ell}$. Similarly, telling the truth is a personal equilibrium, if and only if $\theta_\ell \geq \underline{\theta_\ell}$.

Furthermore, we have

$$\overline{\theta_\ell} = \frac{1+\eta\lambda_M(p+q(1-p))+\eta(1-p)(1-q)}{1+\eta\lambda_L(1-(1-p)q)+q\eta(1-p)}H > \frac{1+\eta(1+p(\lambda_M-1))}{1+\eta\lambda_L}H = \underline{\theta_\ell}, \quad (10)$$

i.e. the region where both lying and truth-telling a personal equilibria is non-empty. For all realizations of θ_ℓ below (above) $\underline{\theta_\ell}$ ($\overline{\theta_\ell}$) only lying (truth-telling) will be a personal equilibrium.

As the lower threshold is independent of q, we have $\frac{\partial\underline{\theta_\ell}}{\partial q} = 0$. The derivative of the upper threshold with respect to q is

$$\frac{\partial\overline{\theta_\ell}}{\partial q} = \frac{\eta H(1-p)\left(\lambda_L\left(\eta(p+1)\lambda_M-\eta p+1\right)+\lambda_M(1-\eta p)-\eta(1-p)-2\right)}{\left(1+\eta\lambda_L(1-(1-p)q)+q\eta(1-p)\right)^2}. \quad (11)$$

Obviously, the denominator is strictly positive. The numerator can be rewritten as:

$$\underbrace{\eta H(1-p)}_{>0}[\underbrace{\lambda_L-1}_{>0}+\underbrace{\lambda_M-1}_{>0}+\underbrace{(\lambda_L-1)p\eta\lambda_M}_{>0}+\underbrace{\eta(\lambda_L\lambda_M-1-p(\lambda_L-1))}_{>0}].$$

Therefore, $\frac{\partial\overline{\theta_\ell}}{\partial q} > 0$.

For now, we have assumed that there is loss aversion in both the monetary and the lying dimensions. If lying costs are independent of expectations ($\eta_L = 0$), the derivative of the upper threshold $\overline{\theta_{\ell m}}$ (loss aversion only in the monetary dimension) with respect to q becomes

$$\frac{\partial\overline{\theta_{\ell m}}}{\partial q} = \eta H(1-p)(\lambda_M-1) > 0. \quad (12)$$

Therefore, for all other parameters given, the lying strategy is a personal equilibrium for more types θ_ℓ, when q is increasing. ■

Proof of Proposition 2

Proof. The lying frequency in case of the low outcome and given the possibility to lie is increasing in q, if and only if the set of lying cost types for which lying constitutes a preferred personal equilibrium is increasing. Recall the threshold

$$\widehat{\theta_\ell} = \frac{1 - \eta(\lambda_M - 1)(1 - q - p(2 - q))}{1 + \eta(\lambda_L - 1)(1 - (1 - p)q)} H, \tag{13}$$

where

$$\frac{\partial \widehat{\theta_\ell}}{\partial q} = \frac{\eta H (1 - p) \left(\lambda_L \left(2\eta p \lambda_M - 2\eta p + 1\right) + \lambda_M (1 - 2\eta p) + 2\eta p - 2\right)}{\left(1 + \eta(\lambda_L - 1)(1 - (1 - p)q)\right)^2}. \tag{14}$$

Obviously, the denominator is strictly positive. The numerator can be rewritten as:

$$\underbrace{\eta H (1 - p)}_{>0}[\underbrace{\lambda_L - 1}_{>0} + \underbrace{\lambda_M - 1}_{>0} + \underbrace{2\eta p(\lambda_L \lambda_M + 1 - \lambda_L - \lambda_M)}_{>0}].$$

Therefore, $\frac{\partial \widehat{\theta_\ell}}{\partial q} > 0$.

For now, we have assumed that there is loss aversion in both the monetary and the lying dimensions. If lying costs are independent of expectations ($\eta_L = 0$), the derivative of the threshold $\widehat{\theta}_{\ell m}$ (loss aversion only in the monetary dimension) with respect to q becomes

$$\frac{\partial \widehat{\theta}_{\ell m}}{\partial q} = \eta H (1 - p)(\lambda_M - 1) > 0. \tag{15}$$

Therefore, for all other parameters given, the lying strategy is a preferred personal equilibrium for more types θ_ℓ, when q is increasing. ■

For comparing our findings on the impact of the success probability p to the literature, note that $\lambda_M \geq \lambda_L$ is a sufficient condition for $\frac{\partial \widehat{\theta_\ell}}{\partial p} > 0$. $\frac{\partial \widehat{\theta_\ell}}{\partial p} > 0$ also holds for the more restrictive case where $\eta_L = 0$, i.e. where expectations or loss aversion does not matter in the lying dimension.

For $\lambda_M \geq \lambda_L$, the threshold $\widehat{\theta_\ell}$ is strictly increasing in p, i.e. lying is a preferred personal equilibrium for more values of θ_ℓ. In this case, $\widehat{\theta_\ell}$ is increasing in p, if q is low or η is high.

Impact of q when the expected outcome serves as reference point

Suppose that, by contrast to our KR-type approach, we would have followed GSV (2016) by assuming that the expected outcome of the lottery serves as reference point; irrespective of whether the individual plans to adopt the lying or the truth-telling strategy. Given that the outcome is low and lying is possible, the utility in the lying strategy is then

$$u_L^L(m = H|e = 0) = H - \theta_\ell + \eta(1 - p)H$$

where $(1-p)H$ is the difference between the actual monetary payoff H and the lottery's expected outcome pH.[18]

If the individual deviates to truth-telling, her utility is

$$u_L^T(m = 0|e = 0) = -\eta\lambda_M pH$$

as she expected to get the high outcome with probability p. It follows that the lying strategy is a personal equilibrium if and only if

$$\theta_\ell \leq H + \eta[1 + (\lambda_M - 1)p]H.$$

As $\lambda_M > 1$, the RHS increases in p, which triggers the GSV (2016) result that high success probabilities make it more likely that lying is a personal equilibrium if p is large. More importantly in our context, however, the RHS is independent of q.

In the truth-telling strategy, the agent's utility from actually telling the truth is

$$u_T^T(m = 0|e = 0) = -\eta\lambda_M pH$$

as she gets zero payoff and had expected payoff of pH. If she deviates to lying, her utility is

$$u_T^L(m = H|e = 0) = H - \theta_\ell + \eta(1 - p)H$$

[18] Note two things: First, considering $(1-p)H$ instead of $(1-p)(1-q))H$ as "mark up" on the expected monetary gain ultimately implies that individuals do not anticipate their own behavior. Second, with such a reference point formation, there is no room for a gain-loss utility in the lying dimension (due to the fact that the reference point refers only to the monetary dimension).

Thus, incentive compatibility of the truth-telling strategy is just the mirror image of incentive compatibility of the lying strategy; i.e. truth-telling is a personal equilibrium if and only if

$$\theta_\ell \geq H + \eta[1 + (\lambda_M - 1)p]H.$$

It follows that exactly one of the two strategies will be a personal equilibrium, and the condition for this is independent of q.

Proof of Lemma 1

Proof. Given that $H < \theta_\ell^{\max}, \theta_r^{\max}$ and the continuity of types, we know that for each $\widehat{\theta_\ell} \leq \overline{\theta_\ell}$ ($\widehat{\theta_r} \leq \overline{\theta_r}$), there exists a unique $\widehat{\theta_r}$ ($\widehat{\theta_\ell}$) such that $\widehat{\theta_\ell}$ and $\widehat{\theta_r}$ satisfy Equation (9), i.e. there exists an equilibrium.

Now consider $\theta_\ell = 0$. Given the low outcome, the agent reports $m = H$, if $\theta_r < \overline{\theta_r}$, while she reports $m = 0$, if $\theta_r \geq \overline{\theta_r}$. At the threshold, she is indifferent, i.e. $H - \overline{\theta_r} v(r) = 0$. As the LHS is monotonically increasing in $\overline{\theta_r}$ and the LHS is strictly greater than 0 for $\overline{\theta_r} = 0$ and strictly lower than 0 for $\overline{\theta_r} = \theta_r^{\max}$, there exists a unique $\overline{\theta_r} \in (0, \theta_r^{\max})$. As $\overline{\theta_\ell}$ is independent of $\overline{\theta_r}$, it follows that $\frac{v(r)}{\overline{\theta_r}} > 0$ and there exists a unique threshold for reputational concerns in equilibrium. ■

Proof of Proposition 3

Proof. Define $P(0 \to H) := \Pr(\theta_\ell < -\frac{\overline{\theta}_\ell}{\overline{\theta}_r}\theta_r + \overline{\theta}_\ell)$ (recall the linearity of the threshold function) as the probability that an entitlement of 0 leads to a message of H, then

$$r = \frac{P(0 \to H)q(1-p)}{P(0 \to H)q(1-p) + p}, \tag{16}$$

where $\frac{\partial r}{\partial q} > 0$. From Lemma 1, it follows that $\frac{\overline{\theta_r}}{q} < 0$. Furthermore, as the threshold $\widehat{\theta_\ell}(\theta_r)$ is decreasing in θ_r, it follows that $\frac{\partial \widehat{\theta_\ell}(\theta_r)}{\partial q} < 0$ and the proportion of liars reporting $m = H$ decreases in q. ■

Similarly, $\frac{\partial r}{\partial p} < 0$. It follows that $\frac{\overline{\theta_r}}{q} > 0$ and the proportion of liars reporting $m = H$ increases in p.

Experimental Instructions

Participants in Mturk were shown a link to our website and a text box to enter a code. After clicking the link, the text on our website for treatment H read as follows:

> You take part in a lottery. In any case, you get \$0.30 for your participation. Furthermore, if you win the lottery, you receive an additional \$0.30, i.e. \$0.60 in total.
>
> One out of four participants wins the lottery.
>
> **If you win:** You get the code for \$0.60. Just enter it in Mechanical Turk and you will get \$0.60.
>
> **If you do not win:** You get the code for \$0.30 (i.e. the payment for your participation). Nine out of ten participants also see what the \$0.60-code would have been. If you see the \$0.60-code and enter it even though you lost the lottery, you will get \$0.60 and we cannot observe if you actually won or not.
>
> Your payment depends only on the code you enter in Mechanical Turk.
>
> Click the following button to see whether you win or not.

In treatment L, the sentence "Nine out of ten participants also see what the \$0.60-code would have been." was changed to "Five out of ten participants also see what the \$0.60-code would have been."

At the bottom of the website there was a button that read "Start the lottery". After clicking that button, two random numbers p and q between 1 and 100 were evaluated. If $p \leq 25$, a text appeared that read

> Congratulations! You have won \$0.60 and the code to enter in Amazon Mechanical Turk is **891**.

If $p > 25$ and $q > 90$ ($q > 50$), the text became

> Unfortunately, you have not won and the code to enter in Amazon Mechanical Turk to get \$0.30 is **275**.

Finally, if $p > 25$ and $q \leq 90$ ($q \leq 50$), the result was

> Unfortunately, you have not won and the code to enter in Amazon Mechanical Turk to get \$0.30 is **275**. The code to get \$0.60 would have been 891.

Loading the website multiple times was prevented by using cookies and blocking the participant's IP address for one hour.