Games and Economic Behavior 147 (2024) 338–376

Contents lists available at ScienceDirect

Games and Economic Behavior

journal homepage: www.elsevier.com/locate/geb

Signaling motives in lying games

Tilman Fries [1]

Department of Economics, LMU Munich, Ludwigstr. 28, 80539, Munich, Germany

ARTICLE INFO

JEL classification:
D82
D91

Keywords:
Honesty
Image concerns
Lying
Psychological game theory

ABSTRACT

This paper studies the implications of agents signaling their moral type in a lying game. In the theoretical analysis, a signaling motive emerges where agents dislike being suspected of lying and where some lies are more stigmatized than others. The equilibrium prediction of the model can explain experimental data from previous studies, particularly on partial lying, where individuals lie to gain a non-payoff maximizing amount. I discuss the relationship with theoretical models of lying that conceptualize the image concern as an aversion to being suspected of lying and provide applications to narratives, learning, the disclosure of lies, and the selection into lying opportunities.

1. Introduction

The virtue ethics of the ancient Greeks recognize honesty among the desirable moral characteristics which can lead individuals to flourish and to live a "good life".[2] Religious texts and popular myths often stress the value of honesty.[3] Honesty also plays a role in economic situations; if Alice is a buyer and Bob is a seller in a credence goods market, it will be relevant for Alice to wonder not only if Bob was honest with her in the exchange they just had, but also whether Bob will be honest again in future exchanges. To form this latter expectation, Alice needs to have an idea about Bob's moral character, in particular about his honesty. This paper is concerned with the strategic implications that arise when individuals want to appear honest.

In strategic situations where different agents have different objectives and where some agents are better informed than others, truthful communication can be difficult or impossible. This impedes information transmission and can lead to market failures (Akerlof, 1970; Crawford and Sobel, 1982). Some of these inefficiencies can be overcome if lying is costly for agents (Kartik, 2009), but the size and form of lying costs is mainly an empirical question.

More recently, a literature has emerged that empirically investigates lying costs in laboratory experiments. In an experiment, Fischbacher and Föllmi-Heusi (2013)–or F&FH–gave participants a six-sided die. Participants were instructed to roll the die in private and report the number they rolled to the experimenter. Upon reporting, participants received a payoff in Swiss Franks corresponding

E-mail address: tilman.fries@econ.lmu.de.

[1] I thank the editor, an anonymous associate editor, and two anonymous referees for clear guidance and constructive comments. I am further grateful to Johannes Abeler, Kai Barron, Christian Basteck, Daniele Caliari, Martin Dufwenberg, Dirk Engelmann, Hoa Ho, Agne Kajackaite, and Daniel Parra for comments and discussions. I also thank participants at the ESA World Meetings 2020 and participants at the seventh CRC 190 Retreat. Financial support by WZB Berlin and Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged.

[2] See e.g. the Stanford Encyclopedia article on Virtue Ethics (Hursthouse and Pettigrove, 2018).

[3] Consider for example the cherry tree myth about a young George Washington who cuts down his father's tree with a hatchet. After finding the cut-down tree, the father confronts his son. Young George confesses and the father promptly embraces him because *"Such an act of heroism in my son is more worth than a thousand trees"* (Weems, 1918). The implied moral seems clear—George Washington did not only become a historical figure but did so honestly. His example should serve to inspire others also to be honest.

https://doi.org/10.1016/j.geb.2024.08.006
Received 29 March 2023

Available online 22 August 2024

to the reported realization of their die roll, except for number six, which paid nothing. Since the objective distribution of the die roll is known, lying behavior can be inferred from the aggregate report distribution. F&FH find that the empirical distribution of reports is consistent with some participants reporting honestly and other participants lying. In various follow-up experiments—that sometimes let participants flip coins instead of rolling a die—similar patterns emerge (Abeler et al., 2019).

One robust feature in experiments that use the F&FH die-roll task is that some individuals lie and dishonestly report 4 when they could have earned more money by lying and reporting 5. One reason for the observed behavior could be that individuals dislike being suspected of lying; since fewer individuals lie to report a number that does not maximize their monetary payoff, reporting a lower number evokes less suspicion. Papers by Dufwenberg and Dufwenberg (2018); Gneezy et al. (2018), and Khalmetski and Sliwka (2019) provide theoretical models that formalize this intuition.[4] In doing so, they all have to come to terms with the fact that lying decisions depend on perceived suspicion, which in turn depends on lying decisions. Suspicion therefore is an equilibrium outcome of a game between an agent and an audience, in which an agent observes a state (a number on a die, a coin flip) and makes a report to an audience. The report serves as a signal to the audience, who in turn forms a belief about the likelihood that the agent lied; a measure of suspicion. Anticipating this, the agent will take their belief over the audience's belief into account when deciding what to report. The agent's utility is *belief-dependent*, as it depends on the perceived *image* that the audience attaches to the agent after hearing the report. In their meta-study, Abeler et al. (2019)–from now on AN&R–conclude that such image concerns are key to explaining the stylized empirical facts observed in experiments on lying.

While image concerns are deemed to be important, there are different ways to conceptualize them. AN&R find that two kinds of image concerns can explain the observed empirical regularities in lying games. The first is an image concern that (in various forms) is used in models by D&D, GK&S, and K&S, where individuals want to signal that they did not lie.[5] The second is a lying model where the signaling motive is similar to the honor-stigma model of Bénabou and Tirole (2006)–hereafter B&T. In this model, individuals want to appear as someone who has a large intrinsic concern for honesty. The main difference between these two approaches is that in the former individuals want to signal a good deed (they did not lie), whereas in the latter model, individuals want to signal a moral character (the extent of their intrinsic honesty). In this paper, I ask if this second approach to image concerns can provide useful insights and extend our understanding of lying behavior. I derive a lying model based on B&T, which so far has only received cursory attention in the literature.[6]

I study the strategic implications of individuals signaling their moral character in a lying game. Agents see the realization of a random number (by rolling a die, flipping a coin, etc.) and make a report to an audience. They are morally concerned and incur a cost if their report does not equal the realization. Agents differ in the extent to which they are morally concerned; some suffer high and others low costs from lying. Individual types are private, but in equilibrium, the agents' reports are informative about their type. This happens because worse moral types are more likely to dishonestly report a high number than better types. In the model, the *credibility* of the report and the *honor-stigma gap* between those who do and do not lie influence an agent's image. A report is more *credible* the more likely it is that it was made truthfully. Moreover, the reputation attached to a report depends on the moral type of the liars reporting it.

To illustrate how reputations form in the character-based model, consider the following example of a professor who, on the day of a final exam, receives messages from some of her students that they are sick and cannot participate in the exam. By university guidelines, sickness is the only acceptable excuse for not writing the exam. Students also find it sufficiently unpleasant to write an exam when they are sick so every sick student will send a message to the professor. There might, however, also be reasons that induce a healthy student to send a message that they are sick. Suppose that some of the students who are not sick are in an *emergency*. Students who are neither sick nor in an emergency and excuse themselves from the exam are *shirking*. Professing to be sick when one is not constitutes a lie. Students dislike lying to different degrees, with some students being more moral (having a higher lying cost) than others. A healthy student will lie and claim to be sick if the benefits from not writing the exam are higher than their lying cost. Since writing the exam is arguably worse when in an emergency, more students will lie with than without an emergency. We can observe that this type of behavior implies sorting of moral types into falsely claiming sickness or not. Those in the left tail of the moral type distribution will lie about their health status while those in the right tail of the distribution will not. The threshold that divides the moral type distribution into a left and a right tail depends on the reasons that students have to lie about their health status. It will be higher for students with than without an emergency, which implies that, for students with an emergency, the left tail is comparatively larger and the right tail is smaller. Fig. 1 sketches out the sorting process from possible states of the world into student actions.

The professor does not observe the real reason for a student who claims to be sick. Therefore, upon receiving a message from a student, the professor forms a posterior expectation about the student's moral character by weighing all different potential motives behind sending the message with their empirical frequency. The posterior expectation after receiving a message will always be lower than the professor's prior expectation about the student, before receiving the message. This is because the professor cannot distinguish

[4] From now on in the text, I will refer to them as D&D, GK&S, and K&S respectively.

[5] GK&S and K&S introduce the image concern as either the probability to have told the truth, conditional on the report or as the probability to have lied, conditional on the report. D&D further interact the conditional probability of having lied with the perceived size of the lie. For example, in D&D the agent gets a lower image if they are suspected of reporting a 5 instead of a 1 than if they are suspected of reporting a 4 instead of a 3.

[6] Proposition 7 in AN&R, Appendix B, provides some general properties of such a model. Their analysis however remains too general to complement the insights derived from the deed-based image model. Indeed, the result that concludes AN&R's meta-study (Finding 10) cannot distinguish between a model that employs a deed-based image concern and a model that uses a character-based image concern as both account for exactly the same empirical facts (*"Only the Reputation for Honesty + LC [deed-based image] and the LC-Reputation [character-based image] models cannot be falsified by our data"* (AN&R, p. 1144)).

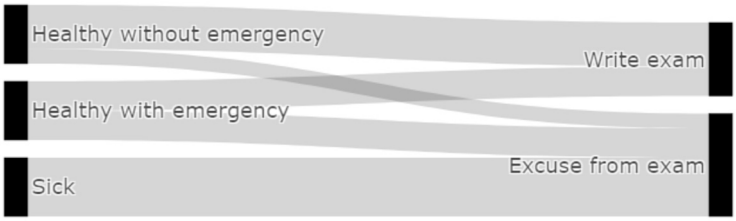


Fig. 1. Sorting from states of the world into actions.

between truthful and dishonest messages—while actual sickness is not correlated with moral types, the students who send a dishonest message pool with students who send a truthful message, and those who send the dishonest message come from the left tail of the type distribution, i.e., they are of a low expected moral type. In line with the idea that individuals want to be perceived of high moral character, a student's reputation is equal to (her beliefs about) the professor's posterior expectation. Now suppose that there is a (potentially pandemic-induced) increase in the probability that a student is sick at the exam date. All things equal, such an increase will increase the professor's posterior expectation. This reflects the credibility effect—if more students are actually sick, it is more likely that any student claiming to be so is telling the truth. Alternatively, consider an increase in the probability that any student faces an emergency at the exam date (which might also be pandemic-induced as they have to care for sick family members). Such an increase will also increase the professor's posterior expectation, as, conditional on not being sick, it is less likely that the student is simply shirking. This reflects the honor-stigma effect—even though they may still lie, students in an emergency who claim to be sick on average are of a higher moral type than students who shirk.

In the die roll game, the character-based model predicts an equilibrium that can include partial lying. Recall that agents have a financial incentive to overstate their number. Therefore, if some agents lie to report the highest paying number, this number will on average be reported by worse moral types. Because agents are image-concerned, they might then have an incentive to leave some money on the table in exchange for a higher image by reporting the second-highest or even lower payoff when they lie. This dynamic generates an equilibrium with characteristics that are similar to the deed-based image models of GK&S and K&S; agents lie only if the realized random number is smaller than or equal to some threshold and report a number that is above the threshold. Under an equilibrium refinement that restricts liars to symmetric strategies, this is the unique outcome of the game.

I compare the results of the character-based model with those of the deed-based model throughout the paper, mainly to sharpen the intuition for the character-based model and to highlight the differences between the two approaches. Both types of image concerns must not be mutually exclusive, and it is plausible that there are situations for which one model is better suited than the other. For example, the character-based model might especially apply to situations such as the professor-student example, where the student anticipates a repeated interaction with the professor and wants to build up trust, thus caring about the professor's belief about their moral type. Other examples include situations where agents want to appear to others as leading virtuous lives. The deed-based model might apply more to situations where an agent's primary objective is to not raise the audience's suspicion that they might have acted mischievously. For example, when reporting on their taxes, an agent might want to avoid making any statement that could also have been made by someone who cheats on their taxes to not even face the prospect of being investigated by the tax authority. In these situations, the deed-based model might be more appropriate.

When comparing the character- to the deed-based model, I identify three comparative statics where their predictions disagree. First, I show that increasing the probability of observing the lowest state *crowds out* lying in the deed-based model while it can *crowd in* lying in the character-based model. Second, I show how introducing types with exceptionally high lying costs into the population *crowds in* lying in the deed-based model but *crowds out* lying in the character-based model. Third, I trace how reducing the audience's uncertainty about the agents' moral types increases lying in the character-based model but has an ambiguous effect in the deed-based model. These differences are due to the honor-stigma effect being present in the character-based model but not in the deed-based model. The paper gives directions on how future experiments could tease the honor-stigma effect and the credibility effect apart. It demonstrates that such a disentanglement would be possible even if the researcher does not know the state observed by the agent.

The applications I consider study belief-based interventions, the disclosure of lies, and the selection into lying opportunities. A recurring theme will be that the effects of most interventions depend on the interplay between the credibility and the honor-stigma effect. As a first effect, an intervention can decrease the likelihood that agents who report a certain state are telling the truth.[7] This makes reporting this state less credible. It is the effect that leads to the kind of disguised behavior that much of the literature has focused on. It always leads to strategic substitutability of actions, where agents become less likely to lie as other agents become more likely to lie. As a second effect, an intervention can also affect the gap in image awarded to those who lie and those who remain honest after observing the realization of a particular state. Through this honor-stigma effect, situations with strategic complementarities can be created where agents lie because "everyone is doing it" or where they may be excessively honest because lying just "is not done". The character-based model thus provides a parsimonious framework for the disguised behavior that deed-based models focus on and the social norm aspect of the honor-stigma model.

[7] In the professor-student example, think of the university introducing a policy that automatically excuses students from exams if they send a doctor's statement of their sickness to a central university office. After the intervention is introduced, only students who could not obtain such a statement would contact the professor directly, with their credibility being consequently decreased.

The character-based model can account for several experimental findings in the literature. First, various experimental tests show evidence for the credibility effect. In one of their experimental treatments, GK&S reduce the probability with which participants observe, and therefore truthfully can report the highest state. The theoretical prediction is that, after reducing the probability, a wider range of non-payoff-maximizing states is reported because it is less credible that participants truthfully report the highest state. The experimental results are in line with this prediction. In a similar spirit, AN&R find that participants who observe the lower state in a two-state lying game become less likely to lie when the probability of observing the high state decreases. Feess and Kerzenmacher (2018) test a related mechanism. In their experiment, they exogenously vary the probability with which participants who toss the lower-paying side of a virtual coin can lie and report the higher-paying side. That is, some participants who toss low can lie while others can not. They find that a smaller proportion of participants lies if more participants can lie. This is also consistent with the notion that individuals care about how credible their report is.

Second, Bašić and Quercia (2022) show that participants who report higher payoffs in an experimental die roll game are considered less trustworthy, which is reflected on multiple dimensions. For example, observers indicate that they would be less likely to lend money or employ someone who reports a higher payoff. This is consistent with the idea that reports in the lying game are diagnostic about moral types.

Third, a strand of lying experiments exists that shifts beliefs about the lying of others and subsequently measures behavior. In one experiment of that sort, AN&R exogenously increase participants' beliefs that others will lie using an anchoring technique and find that this slightly decreases lying. This effect, though insignificant, goes in the direction predicted by the deed-based model. Results from related experiments typically provide less direct evidence for deed-based models. Experiments reported by Rauhut (2013); Diekmann et al. (2015), and Akın (2019) provide participants with information about how others lied to induce participants to update their beliefs. These experiments usually find zero average treatment effects that mask heterogeneous responses, where, after being provided with information, underestimators become more likely to lie and overestimators less likely to lie. These observations are inconsistent with the deed-based model but can be rationalized through the character-based model. As Section 3.2 will line out, individuals with a character-based image concern will react differently to information about the empirical reporting frequency depending on how they interpret it. Results from Le Maux et al. (2021) show that participants respond to information even when their lies are perfectly observed and there thus is no credibility effect. They can be taken as further evidence that the credibility effect is not always the only belief-based motive individuals hold.[8]

Fourth, the signaling motives implied by the character-based model can also account for findings from Bicchieri et al. (2023) who study motivated reasoning and lying. This paper argues and provides consistent experimental evidence, that individuals choose to believe that a higher fraction of other individuals are lying to justify their own lies. Thus, participants in their experiment choose to give up belief in the credibility of their report because the composition-based motive that "everybody is doing it" or that "nobody is perfect" provides a better excuse for dishonesty. Therefore, the credibility and honor-stigma effects of the character-based model provide a framework that we can use to organize the experimental evidence on how beliefs affect lying behavior.

The following Section 2 presents the model. I apply the model to investigate the determinants of reputation in Section 3.1. Section 3.2 provides comparative statics concerning the type distribution and provides an application to belief-based interventions. Section 3.3 extends the model to investigate interventions that detect liars and that study cheating in the context of selection. Throughout Section 3, I contrast predictions of the character-based model with predictions from a deed-based model. The paper concludes in Section 4. Proofs of all formal results appear in Appendix A.

2. Model

2.1. Setup

2.1.1. Game form

Consider a game between a continuum of agents and an audience. Nature initially chooses a state $j \in \mathcal{K} \equiv \{1, \dots, K\}$, where the probability of nature choosing state j is given by $p(j) \in (0,1)$ and where $\sum_{j=1}^{K} p(j) = 1$. Agents privately observe the realization of the state and each makes a report $a \in \mathcal{K}$ to the audience. They then receive a material payoff according to a function $y(a)$ which strictly increases in a. For notational convenience, we define as $\Delta(a, a') \equiv y(a) - y(a')$ the material payoff difference between reporting a and a'. The audience (she) is a passive player with no action whose payoff we do not further specify. She observes a but not j. The agents can be thought to be participants in an economic experiment who are asked by the experimenter (the audience) to roll a die. In this case, the state would be the outcome of a die roll, p would be uniform on $\mathcal{K}$, and $y(a)$ would reflect the experimenter's choice of rewards for reporting certain numbers of the die. An alternative interpretation of the setup could see agents as students who, on the day of an exam, are either sick or healthy. The agent-as-student would then always earn a higher material payoff by claiming to be sick than by claiming to be healthy.

[8] Information provision experiments without an active control group provide little experimental control over how treated participants update to information, relative to control (Haaland et al., 2023). It is, therefore, difficult to imagine treatments in this framework that could falsify the character-based model. For example, one problem of this research design is that underestimators might be different from overestimators in unobserved ways. In this case, the treatment assignment (whether participants update their beliefs downward or upward) is not exogenous. This is not necessarily a problem if the goal of the treatment is to measure the average effect of information provision. However, it renders these experiments less informative about potential theoretical mechanisms.

2.1.2. Psychological utility

We consider two psychological utility components—lying costs and belief-dependent image utility—that together with the material payoff add up to total utility.

Direct lying costs. Reporting $a \neq j$, agents incur cost t which is heterogeneous across agents. This cost arises through a purely intrinsic, moral preference for honesty. That individuals are heterogeneous in their preferences for honesty is documented in experiments such as Gibson et al. (2013), Gneezy et al. (2013), and Kajackaite and Gneezy (2017). Gibson et al. (2013) in particular show that the lying cost distribution function consists of many intermediate types, who begin to lie if the returns to lying are high enough. The intrinsic preference for honesty reflects that agents feel bad for lying. Modeling lying costs as fixed seems appropriate as a first approximation based on the evidence from observed lying games reported by AN&R and GK&S, where the experimenter observes the individual state realizations and reports. The data from these experiments shows a "missing middle" pattern, where individuals either tell the truth or lie to report the highest number, with only a minority of liars reporting a number in between. This suggests that cost functions that increase in the size of the lie, and which therefore could rationalize partial lying for intrinsic reasons, are not necessarily needed to describe lying behavior in these experiments.[9] The lying cost is unknown to the audience, who however knows that it is drawn from a distribution $F(t)$ with full support on $(\underline{t}, \bar{t}]$ and which is independent of j. We assume that $\underline{t} \in [0, \Delta(K,1))$, which ensures that lying is costly for all agents and that, for some, the material payoff gain from lying is higher than the intrinsic cost. While in most cases setting $\underline{t} = 0$ seems natural, we solve a more general version of the model as we will consider $\underline{t} > 0$ in parts of the comparative statics section. Upper bound $\bar{t}$ is a large number, to be specified in detail below.

I will use "lying cost" and "moral type" interchangeably when discussing t, as this section considers honesty as the only relevant moral dimension. This is due to the setup of the game, which reflects laboratory lying games and elements of verbal communication. In these settings, lying comes at no expense to a third party, which allows us to exclusively focus on honesty.[10] Further morality dimensions, such as altruism, might become relevant and interact with honesty in settings where agents cheat someone else, for example, stealing (Footnote 21 in Section 3 provides further discussion of this point).

Image utility. In addition to being intrinsically honest, agents also value a reputation for honesty. There can be instrumental reasons to value such a reputation. An expert might like to appear honest to build an enduring relationship with an advisee. A student who hopes to receive a good letter of support from their professor wants to appear sincere to them. There are also noninstrumental reasons why an agent might prefer to look honest; many individuals want to appear moral and one indicator of morality is honesty. This type of image concern follows B&T and other approaches in psychological game theory that formalize the idea that individuals want to signal "good traits" (Battigalli and Dufwenberg, 2022): Through their actions, agents tell others something about their intrinsic preferences, and agents want to look as if they have preferences which are valued by an audience. To make an inference, the audience forms a belief about the expected moral type of an agent reporting a. I call this type of image concern *character-based*.

Definition 1. The character-based image concern is equal to $\mathcal{R}_a^C \equiv \mathbb{E}(t|a)$.

Parts of the paper will compare predictions of the character-based image concern model to those of a model with a deed-based image concern. When making this comparison, I will follow the formal assumptions of GK&S:

Definition 2. The deed-based image concern is equal to $\mathcal{R}_a^D \equiv \mathrm{P}(\text{honest}|a)$.

The remainder of this section will be concerned with the model with character-based image concerns. The image utility equals the image concern weighted by a scalar $\mu > 0$,

$$\mu \mathcal{R}_a^C,$$

where μ is not too large so that agents are not disproportionately sensitive to changes in image utility.[11]

Total utility. Material payoffs and psychological utility add up to total utility. An agent of type (j,t) who reports a, and is observed by an audience whose beliefs are such that $\mathbb{E}(t|a) = \mathcal{R}_a^C$, earns utility

$$u(j,t,a,\mathcal{R}_a^C) = y(a) - 1_{a \neq j} t + \mu \mathcal{R}_a^C.$$

I now assume that the maximum lying cost is a number $\bar{t} > \Delta(K,1) + \mu \mathbb{E}(t)$. The assumption ensures, in line with the empirical evidence provided by AN&R, that there are agents who never lie, independent of the observed state. One immediate consequence of the assumption is that the audience always puts a positive probability on any state being reported. This property is helpful when solving for the equilibrium, as described next.

[9] I discuss how the model predictions would change in extensions of the model to more complex cost functions in Section 4.

[10] The setup might further reflect tax reporting, where individual contributions are a negligible part of total tax earnings.

[11] If μ is large multiple equilibria can obtain. An explicit upper bound will depend on the preference distribution function. The Online Appendix shows that $\mu \leq 1$ is sufficient if $F(t)$ is log-concave.

2.2. Equilibrium

The structure of the game makes it a *psychological game* (Battigalli and Dufwenberg, 2009), as the total utility of agents depends on the audience's beliefs about the agents' moral types.[12] Agents' strategies s map their type into a distribution over reports. Denote the probability of an agent of type (j,t) reporting a by $s(a|j,t)$. In the following, an agent is a *liar* if they choose a *dishonest* strategy where $s(a=j|j,t)=0$. To put it another way, an agent who never tells the truth is a liar. Conversely, a *truth-telling* agent is an agent with a strategy $s(a=j|j,t)=1$.

The following equilibrium definition invokes the standard conditions of utility maximization and that agents and the audience correctly apply Bayes' rule and have a common prior. This definition follows the literature and serves as a useful yardstick to think through strategic interdependencies. Since the maximum lying cost is high, every state is reported with a positive probability in equilibrium. This implies that Bayes' rule can be applied to calculate the equilibrium reputation of every state, obliterating the need for further equilibrium refinements to pin down beliefs that are off the equilibrium path.

Definition 3. An equilibrium is defined by strategies $s(a|j,t)$, where

- $s(a=j|j,t)\geq 0$, $s(a\neq j|j,t)\geq 0$, and $\sum_{k\in\mathcal{K}} s(a=k|j,t)=1$ for all j and t.
- $s(a|j,t)>0$ if and only if $a\in \arg\max_{a\in\mathcal{K}} u(j,t,a,\mathcal{R}_a^C)$.
- Agents and the audience hold the correct equilibrium beliefs

$$\mathcal{R}_a^C=\frac{\sum_{k\in\mathcal{K}}\int_{\underline{t}}^{\bar{t}} s(a|k,t)t f(t)\,\mathrm{d}t}{\sum_{k\in\mathcal{K}}\int_{\underline{t}}^{\bar{t}} s(a|k,t)f(t)\,\mathrm{d}t} \text{ for } a\in\mathcal{K}.$$

We are further going to focus the analysis in the main text on equilibria where agents play symmetric lying strategies. This refinement is motivated by the fact that the equilibrium definition above allows for a very rich variety of strategies that liars can play, some of which might appear "strange", or, at least, would require a considerable amount of coordination among liars. For example, with $K=4$, there can be an equilibrium in which some liars from 1 lie up to report 2 and some agents from state 3 lie down and also report 2. This equilibrium can be sustained if liars coordinate on their moral type; that is, the liars with the highest moral type report 2 while those with the lowest moral type report 4. Such behavior can be seen as problematic. Because lying costs are fixed, liars, conditional on lying, have the same preference ranking among reports. There is no a priori reason why a liar would report one state over another if they are indifferent over both. The degree to which liars have to coordinate to support such an equilibrium motivates a refinement that restricts agents to symmetric lying strategies, as defined below. Lemma 1 in the appendix presents more general properties that hold in any equilibrium of the game.[13]

Definition 4. Agents play symmetric lying strategies if $s(a=k|j,t), s(a=k|j',t')>0 \Rightarrow s(a=k|j,t)=s(a=k|j',t')$ for any $t,t'\in(\underline{t},\bar{t}]$, $j,j'\in\mathcal{K}\setminus\{k\}$.

Lying strategies are symmetric when the agents' type (j,t) determines whether they lie or not, but does not determine which state they report. A similar property is imposed by D&D ("uniform cheating") to obtain their main result. For the deed-based model, K&S prove that the equilibrium in symmetric strategies is unique. The refinement implies that liars randomize in the same way which states to report dishonestly. While there are few direct tests of mixed lying strategies, evidence from F&FH is seemingly in line with this refinement. They show that the reports of participants who take part in a die-roll experiment for a second time, and who reported the highest payoff in the first experiment, are indistinguishable from the second-time reports of participants who reported the second-highest payoff in the first experiment. If liars had further conditioned their reports on some intrinsic attributes, we would expect the reports of those who report the highest state to be systematically different from those who report the second-highest state.[14]

Solving the model in symmetric strategies gives the main existence and uniqueness result.

Proposition 1. *There exists a unique equilibrium in symmetric lying strategies with the following properties:*

(*i*) *There is a threshold state* $k^*\in\mathcal{K}\setminus\{K\}$ *and a vector of threshold lying costs* $\hat{t}^*=(\hat{t}_1^*,\dots,\hat{t}_K^*)$ *such that:*
  (*a*) $s(a|j,t)>0$ *for all* $a>k^*$, $t\leq\hat{t}_j^*$ *and* $\sum_{a=k^*+1}^{K} s(a|j,t)=1$ *for all* $t\leq\hat{t}_j^*$.
  (*b*) $s(a=j|j,t)=1$ *for all* $t>\hat{t}_j^*$.

[12] Battigalli and Dufwenberg (2009) provide a framework for games where players' utility can depend on *endogenously formed* beliefs. This extends the psychological game theory framework of Geanakoplos et al. (1989), who only allow *initial* beliefs to enter utility functions. Allowing for endogenous beliefs is crucial in games with image concerns because players update their beliefs throughout the game after observing actions.

[13] Appendix D gives an example of an asymmetric equilibrium where liars condition their strategies on their moral type.

[14] F&FH also show that participants who make reports lower than the second-highest payoff in the first experiment are more likely than others to make reports lower than the second-highest payoff in the second experiment, implying that decisions are to some extent consistent across both experiments.

(c) $\hat{t}_j^* = \underline{t}$ *for all* $j > k^*$.

(*ii*) $\mathcal{R}_a^C$ *is strictly decreasing in* a.

(*iii*) *If* $p(j) = 1/K$ *for all* $j \in \mathcal{K}$, *the report distribution is strictly increasing in* a.

The equilibrium of the game is of the following type: Agents lie only if they observe a state smaller or equal to some threshold state k^*. If they lie, they report a state larger than k^*. If $p(j)$ is uniform, state K is reported by most agents, followed by $K-1$, and so on. In what follows, I discuss the equilibrium properties while relegating the existence and uniqueness proof to Appendix A.

I will refer to states that are reported by liars as *high states* and states that are not reported by liars as *low states*. The set of high states is $\mathcal{H}$. Agents either report the state they observed or one of the high states. The parameter assumptions ensure that some type always reports K dishonestly with positive probability in equilibrium. Moreover, because lying costs are fixed, agents who lie are indifferent between reporting any of the high states. Because of this indifference, the decision problem becomes binary: agents prefer lying to telling the truth if and only if they prefer reporting K to reporting the state that they observed. Conditional on lying, they randomize over reporting any of the high states. Therefore, an agent of type (j,t) lies if and only if $t \leq \hat{t}_j$ for a cutoff $\hat{t}_j$ that, if interior, solves[15]

$$y(K) - \hat{t}_j + \mu \mathcal{R}_K^C = y(j) + \mu \mathcal{R}_j^C. \tag{1}$$

Truth-tellers therefore comprise the upper tail of the preference distribution and liars make up the lower tail. Truth-tellers and liars who observe j have an expected moral type of respectively

$$\begin{aligned} &\mathcal{M}^+(\hat{t}_j) \equiv \mathbb{E}(\boldsymbol{t} | \boldsymbol{t} > \hat{t}_j) \geq \mathbb{E}(\boldsymbol{t}), \\ &\mathcal{M}^-(\hat{t}_j) \equiv \mathbb{E}(\boldsymbol{t} | \boldsymbol{t} \leq \hat{t}_j) < \mathbb{E}(\boldsymbol{t}). \end{aligned} \tag{2}$$

The first term is larger than the second, which reflects that liars are stigmatized while truth-tellers are honored. Now suppose that some type lies after observing j. Because liars maximize utility, their report maximizes $y(a) + \mu \mathcal{R}_a^C$. If a type lies after observing j, this in turn implies that reporting j does not maximize $y(a) + \mu \mathcal{R}_a^C$. Therefore, no liar reports j and the reputation associated with reporting it is equal to $\mathcal{M}^+(\hat{t}_j) > \mathbb{E}(\boldsymbol{t})$. We collect the cutoffs $\hat{t}_j$ of each state in a vector $\hat{t}$ and define the *expected moral type of liars* as

$$\mathcal{L}(\hat{t}) \equiv \sum_{j \in \mathcal{K}} \text{P(observe } j | \text{lie}) \mathcal{M}^-(\hat{t}_j), \text{ with P(observe } j | \text{lie}) = \frac{F(\hat{t}_j)}{\sum_{k \in \mathcal{K}} F(\hat{t}_k)}. \tag{3}$$

Now consider a state j' that liars report with positive probability. Because liars maximize utility, j' must be among the reports that maximize $y(a) + \mu \mathcal{R}_a^C$. This suggests that agents who observe j' always tell the truth since lying is intrinsically costly and would decrease utility. In an equilibrium in symmetric lying strategies, all liars randomize in the same way between reporting any state that is reported by liars with positive probability. For this reason, the reputation associated with reporting j' is a weighted average between the expected moral type of truth-tellers (which equals the prior) and the expected moral type of liars (which is strictly smaller than the prior). Defining $r_j \equiv \text{P(honest} | a = j)$, it can be written as

$$\mathcal{R}_{j'}^C = r_{j'} \mathbb{E}(\boldsymbol{t}) + (1 - r_{j'}) \mathcal{L}(\hat{t}) \text{ if } j' \in \mathcal{H}. \tag{4}$$

The expression above is smaller than the audience's prior expectation $\mathbb{E}(\boldsymbol{t})$ as it is a convex combination of $\mathbb{E}(\boldsymbol{t})$ and $\mathcal{L}(\hat{t})$. This property is crucial to rule out candidate equilibria with downward lying. To see this, suppose by contradiction that there is an equilibrium where $s(a = j' | j, t) > 0$ for some t and $j > j'$. We have just shown that symmetric lying strategies imply $\mathcal{R}_{j'}^C < \mathbb{E}(\boldsymbol{t})$. Furthermore, we have also argued that $\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j) > \mathbb{E}(\boldsymbol{t})$. Therefore, an agent who reports j' after observing j reduces their image utility, their material payoff, and pays an intrinsic lying cost. This contradicts utility maximization. Therefore, symmetric lying strategies rule out equilibria with downward lying.

With these arguments in mind, we can determine why any equilibrium in symmetric strategy has to fulfill Property (*i*) that divides the state space at k^*. Rewrite Equation (1) as

$$\Delta(K, j) + \mu(\mathcal{R}_K^C - \mathcal{M}^+(\hat{t}_j)) - \hat{t}_j = 0. \tag{5}$$

This equation implies that $\hat{t}_j$ decreases among $j \leq k^*$. Therefore, if $\hat{t}_{k^*} \geq \underline{t}$ then $\hat{t}_j \geq \hat{t}_{k^*}$ for all $j < k^*$. For example, if there is a type who lies after observing k^*, there is also a type who lies after observing $j < k^*$. This implies Part (*i*) (*a*). Part (*i*) (*b*) then follows from earlier arguments suggesting that, in equilibrium, if a state is reported by liars, no agent who observes that state will lie.

Turning to Part (*ii*) in the proposition, decreasing reputations, it is useful to distinguish between low states and high states. Among the low states, reputations decrease as the material payoff of a state increases because, as the material payoff increases, agents have a smaller direct incentive to lie. For example, agents who report 1, despite having a high incentive to lie, send a higher signal about their intrinsic honesty than agents who report k^*. Reputations also intuitively decrease among high states because liars trade off material

[15] The assumption that $\bar{t} > \Delta(K,1) + \mu \mathbb{E}(\boldsymbol{t})$ ensures that the l.h.s. is smaller than the r.h.s. for some $t < \bar{t}$. If the l.h.s. is weakly smaller than the r.h.s. for $t = \underline{t}$, no agent lies after observing j. These are the high states for which $\hat{t}_j = \underline{t}$.

payoff against image utility. As the material payoff of reporting a high state decreases, the reputation associated with reporting it has to increase to ensure that liars are indifferent among high states.

If p is uniform, decreasing reputations imply increasing reporting frequencies; among low states, there is an inverse relation between the reputation of the state and the proportion of agents who report it. With symmetric lying strategies, the same relation holds among high states, as the reputation of any state is decreasing in the proportion of liars that are reporting it. Therefore, in the proposition, (iii) is a consequence of (ii).

The equilibrium described in Proposition 1 is similar to the equilibria of the deed-based analogs (GK&S; K&S). They share the same threshold state structure where agents lie only if they observe a state below a threshold and liars randomize between reporting any state above that threshold. Furthermore, they can all replicate the partial lying phenomenon observed in F&FH-type experiments.[16,17] Even the motive behind lying partially is essentially the same: In the equilibrium with symmetric lying strategies, all states in the lying range are reported by two groups of agents, those who are honest and have expected moral type $\mathbb{E}(\boldsymbol{t})$ and those who lie with expected moral type $\mathcal{L}(\hat{t}^*)$. Reputations associated with reporting a in the lying range only decrease in a because r_a, the probability of being honest conditional on reporting a, decreases in a. Therefore, agents tell partial lies to gain credibility. The same motive explains partial lying in the deed-based model. The differences between both models become apparent when we look at the determinants of reputation, as discussed in the next section.

3. Comparisons to the deed-based model

3.1. Determinants of image: credibility and the honor-stigma gap

Let us in this section delve deeper into the determinants of image in the character-based model. This is crucial to sharpen our intuitions about how image concerns determine behavior. Image concerns lead to strategic interdependencies between agents through the effects agents' actions have on equilibrium reputations. We will examine these strategic interactions by shifting the type of the marginal liar and evaluating behavioral spillovers.

I build up intuition for the results by focusing on the case with only two states. I will provide insights about the impact of reputation in the general case when discussing comparative statics in Section 3.2. From Proposition 1, we know that with two states there is an equilibrium in which agents always tell the truth after observing 2 and where some agents lie after observing 1.[18] Lying brings a material payoff gain of $\Delta(2,1)$ at a cost of t. In equilibrium, a fraction $F(\hat{t})$ lies after observing 1. Denoting the probability of observing 2 by p, the probability that an agent reporting 2 is truth-telling becomes $r(\hat{t}) = p/(p + (1-p)F(\hat{t}))$. Reporting 2 over 1 comes with a reputational penalty of size

$$\Psi(\hat{t}) = \underbrace{\mathcal{M}^+(\hat{t})}_{\text{Reputation from reporting 1}} - \underbrace{\left[r(t)\mathbb{E}(\boldsymbol{t}) + (1-r(\hat{t}))\mathcal{M}^-(\hat{t})\right]}_{\text{Reputation from reporting 2}}.$$

After a bit of algebra, we see that it can be equivalently formulated as[19]

$$\Psi(\hat{t}) = \frac{1}{1-p} \times (1-r(\hat{t}))(\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})). \tag{6}$$

This formula tells us that the stigma associated with the high report is proportional to the product of two terms. The term $1 - r(\hat{t})$ denotes the probability that a report of 2 is a lie. Therefore, the relative stigma of reporting 2 over 1 increases as it becomes more likely that reporting 2 is a lie. The term $\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})$ denotes the difference in the moral character of liars and non-liars among those agents who observed 1, i.e., among those agents that could lie to increase their material payoff. Therefore, the relative stigma of reporting 2 increases as lying becomes more diagnostic about moral character. In the two-state case, the equilibrium is pinned down by the threshold type $\hat{t}$ which is exactly indifferent between lying and truth-telling;

$$\Delta(2,1) - \hat{t} = \mu\Psi(\hat{t}).$$

[16] For exercises that calibrate the deed-based model to the data, see AN&R or K&S. Appendix B shows how the character-based can be calibrated to fit the available experimental evidence.

[17] The equilibrium prediction can also capture the student behavior described in the introductory example. Set $K = 3$ and interpret the states as the student being sick (observing 3), the student being healthy (observing 2), and the student having an emergency (observing 1). In an equilibrium where $\kappa^* = 2$, sick students always report being sick, while those who are healthy or in an emergency only report being sick if they have a sufficiently low lying cost. Moreover, students with an emergency are more likely to lie than students who are healthy. To further account for the fact that, for example, a student in an emergency may not benefit from reporting 2 instead of 1, the model would need to be extended to allow the material payoff to depend on the report and the observed state.

[18] With $K = 2$ we do not need the symmetric lying strategies refinement to obtain uniqueness.

[19] To see this, use the martingale property of beliefs, $\mathbb{E}(\boldsymbol{t}) = F(\hat{t})\mathcal{M}^-(\hat{t}) + (1-F(\hat{t}))\mathcal{M}^+(\hat{t})$, to replace $\mathbb{E}(\boldsymbol{t})$ in the stigma function:

$$\begin{aligned}\Psi(\hat{t}) &= \mathcal{M}^+(\hat{t}) - r(\hat{t})(F(\hat{t})\mathcal{M}^-(\hat{t}) + (1-F(\hat{t}))\mathcal{M}^+(\hat{t})) - (1-r(\hat{t}))\mathcal{M}^-(\hat{t}) \\ &= \left(\frac{p+(1-p)F(t)}{p+(1-p)F(t)} - \frac{p(1-F(t))}{p+(1-p)F(t)}\right)\mathcal{M}^+(t) - \left(\frac{pF(\hat{t})}{p+(1-p)F(\hat{t})} + \frac{(1-p)F(\hat{t})}{p+(1-p)F(\hat{t})}\right)\mathcal{M}^-(t) \\ &= \frac{F(\hat{t})}{p+(1-p)F(\hat{t})}(\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})) = \frac{1}{1-p}(1-r(\hat{t}))(\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})).\end{aligned}$$

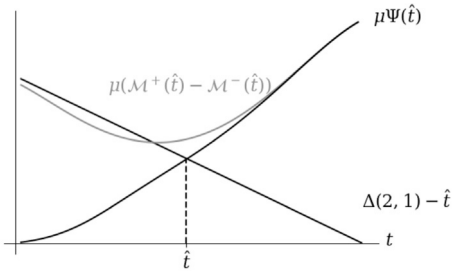


Fig. 2. Equilibrium for $K = 2$.

The left-hand side is decreasing in $\hat{t}$. Now consider the right-hand side. For small values of $\hat{t}$, the stigma function goes to zero as

$$\lim_{\hat{t}\to 0}\Psi(\hat{t}) = \frac{1}{1-p} \times (1 - r(0))(\mathcal{M}^+(0) - \mathcal{M}^-(0)) = \frac{1}{1-p} \times 0 \times (\mathcal{M}^+(0) - \mathcal{M}^-(0)) = 0.$$

As t increases, the stigma changes because of changes in the *credibility* of the report and in the *honor-stigma gap* between those who lie and those who tell the truth after having observed 1;

$$\Psi'(\hat{t}) = \frac{1}{1-p}\left[(1 - r(\hat{t}))\underbrace{(\mathcal{M}^{+\prime}(\hat{t}) - \mathcal{M}^{-\prime}(\hat{t}))}_{\substack{\text{Honor-stigma}\\ \text{effect } (\lesseqgtr 0)}} \underbrace{-r'(\hat{t})}_{\substack{\text{Credibility}\\ \text{effect } (>0)}} (\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t}))\right].$$

More agents reporting 2 makes it less credible that anyone reporting 2 is truth-telling. This effect leads to an increase in the stigma of reporting 2 after a marginal increase in $\hat{t}$. In addition, the types of those who lie and those who tell the truth change. The sign of this additional honor-stigma effect depends on the properties of $f(t)$, and we will discuss conditions for when it is positive or negative when discussing the role of non-observability below. However, independently of the moral type distribution, since the honor-stigma effect is weighted by $1 - r(t)$, it becomes less important as the probability of observing 2, p, increases. Therefore, if p is sufficiently high, the credibility effect dominates. This then implies that an increase in aggregate lying (an increase in $\hat{t}$) increases the relative stigma of reporting 2 over 1 (increases $\Psi(\hat{t})$). Lies are strategic substitutes; an increase in the lying of one agent crowds out the lying of other agents. An equilibrium obtains where the stigma function crosses the difference between the material payoff gain and the direct lying cost as displayed in Fig. 2.

Proposition 2. *If p is sufficiently large, the stigma function is increasing in $\hat{t}$. Lies are strategic substitutes.*

Relation to deed-based image concerns. I relate the findings to those of a deed-based model in which agents are esteemed for taking an honest action. In a model with such an image concern, agents receive a reputation that is proportional to the probability that they made a truthful report (see Definition 2). Therefore, image concerns in the deed-based model influence agents' behavior only through the credibility effect and not through the honor-stigma effect. The comparative statics of the stigma function with respect to $\hat{t}$ are therefore relatively straightforward; as $\hat{t}$ increases, reporting 2 becomes less credible. Lies are strategic substitutes. The following sections will explore cases where, due to the character-based model's honor-stigma effect, the qualitative predictions of the character- and deed-based models disagree.

The role of non-observability. Uncertainty about whether reporting 2 is a lie or not makes the model distinct from the standard B&T honor-stigma model. In these models, actions are usually perfectly observed so that the stigma from taking the "bad" over the "good" action is equal to $\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})$.[20] In the case of a single-peaked type distribution, this difference decreases for small $\hat{t}$ and increases for larger $\hat{t}$. Agents thus face the highest signaling incentives when the marginal type is either very small or very large. As Adriani and Sonderegger (2019) note, this intuitively happens because agents either want to separate themselves from the few "bad apples" that exist in the left tail of the distribution or because they want to belong to the "stars" in the right tail of the distribution. In the non-observed lying game, the reputational wedge of the standard Bénabou and Tirole honor-stigma model gets weighted by the probability that a report of 2 is a lie, as displayed in Equation (6). This reflects the audience's uncertainty about the state that remains after observing a report of 2. Intuitively, a small amount of "bad apples" barely affects the credibility of reporting 2 and provides agents with weak image incentives to separate to signal honesty. Put differently, truth-telling only pays off reputationally if the audience expects many agents to lie. Fig. 2 contrasts the stigma function in a non-observed lying game with the stigma function in a game where the audience can perfectly identify lies. The equilibrium threshold in the non-observed game is always larger than the threshold in the observed game because identified liars cannot reputationally benefit from pooling with truth-tellers.

[20] Most closely related are Bénabou and Tirole (2006), who provide a brief discussion of behavior under forced abstention of some agents (see their Proposition 7).

Table 1
Stigma functions and their derivatives, by image motive and degree of observability.

	Character-based	Deed-based
Non-observed	$\Psi(t) = \frac{1}{1-p}(1-r(t))(\mathcal{M}^+(t) - \mathcal{M}^-(t))$	$\Psi(t) = 1 - r(t)$
	$\Psi'(t) = \frac{1}{1-p}\left[(1-r(t))(\mathcal{M}^{+'}(t) - \mathcal{M}^{-'}(t)) - r'(t)(\mathcal{M}^+(t) - \mathcal{M}^-(t))\right]$	$\Psi'(t) = -r'(t)$
Observed	$\Psi(t) = \mathcal{M}^+(t) - \mathcal{M}^-(t)$	$\Psi(t) = 1$
	$\Psi'(t) = \mathcal{M}^{+'}(t) - \mathcal{M}^{-'}(t)$	$\Psi'(t) = 0$

Table 1 summarizes the stigma functions of models with character- and deed-based image concerns, for cases where lies are either non-observed or observed.[21] This paper is mostly concerned with the character-based/non-observed case as displayed in the upper-left quadrant. Deed-based models of lying (e.g., by GK&S and K&S) are in the upper-right quadrant. The character-based model with observed actions which, following B&T, is a standard model to explain, e.g., prosocial behavior such as charitable giving, is in the bottom-left quadrant. The bottom-right quadrant displays the deed-based model for the observed case. As only actions are stigmatized in the deed-based model, if these actions are observed, the degree of stigmatization will not depend on the behavior of others (i.e., the stigma function is flat). A direct implication of this is that beliefs about what others do should not matter for individuals with deed-based image concerns once their action is perfectly observed. This is a prediction which can be tested empirically. Le Maux et al. (2021) conduct an experiment that seems to contradict it. In their experiment, lies are observed. However, as participants receive information about how other participants in the experiment behave, their own behavior changes.

3.2. Comparative statics

We turn to comparative statics of the general model. I will focus on comparative statics concerning the type distributions $p(j)$ and $f(t)$. These comparative statics provide new insights relative to what we know from the deed-based model, in particular, relative to the comparative statics reported by K&S. Additional comparative statics with respect to the material payoff of K and the image sensitivity μ are relegated to the Online Appendix.

This section will mainly investigate changes in two outcomes: the likelihood that an agent lies and the threshold state k^*. The first outcome is a straightforward measure of lying. The second outcome is more intimately connected to the image motive. Recall that liars only report a state lower than K to derive a higher image utility. Therefore, if k^* increases, this suggests that liars have become more willing to trade off the material payoff gain against a decreased image, possibly because the relative stigma of reporting K decreased. Investigating changes in k^* is a useful way to illustrate how image concerns affect behavior.

3.2.1. Changing the state distribution

Image models of lying predict that reporting is sensitive to the distribution of states. Consider starting out with a uniform state distribution and reducing $p(K)$, the probability of observing K, from $1/K$ to $1/K - \tilde{\delta}$, equally redistributing probability mass $\tilde{\delta}$ among the remaining states. In equilibrium, honest agents and liars pool when reporting K and liars free-ride on the image provided to them by the higher expected moral type of honest agents. With a lower $p(K)$, the likelihood of an agent reporting K *honestly* decreases, which reduces the pooling advantage conferred to liars. Their reports become less credible. Liars will thus find other states to report, thereby expanding the range of states that are reported dishonestly. This is the intuition behind Proposition 3a below. Since the prediction works through the credibility effect, the deed-based model makes the same prediction.[22]

Proposition 3a. *Suppose that $k^* > 1$ and that states are uniformly distributed on $\mathcal{K}$. When agents hold character- or deed-based image concerns, redistributing probability mass $\tilde{\delta} \in (0, 1/K)$ away from K evenly towards all other states:*

(*i*) *Weakly decreases k^*.*
(*ii*) *Weakly decreases or increases the likelihood that an agent lies.*

The predictions of the character- and deed-based model may instead differ if we investigate shifts of the state distribution among the low states, i.e., among states that are observed but not reported by liars. To illustrate, consider an audience who observes a high report. In the character-based model, the audience now wonders about how likely it is that this report is a lie (to determine credibility) and what kind of types tell lies (to determine the expected moral type of liars). The answers to both questions will depend on the state distribution. For example, agents who lie after observing 1 are of a higher expected moral type than those who lie after observing k^*.[23] When we now distribute probability mass away from $p(k^*)$ and towards $p(1)$, the audience attaches a higher expected

[21] Note that, when moving from left to right in the table, we compare models that differ in the assumptions they make about the psychological underpinnings of image concerns. Instead, when moving up or down, we compare models that make different assumptions about the choice environment.

[22] See Proposition 6 in GK&S for a related result. GK&S also provide experimental evidence that shows that this prediction is borne out in the data.

[23] The fact that "small" lies are more severely stigmatized than "large" lies would be more ambiguous in a setting where agents' lying decisions have material payoff implications for a third party. In settings where agents cheat at the expense of others, it would be appropriate to introduce further moral dimensions, such as

moral type to liars—after all, they are now more likely to have observed 1. This decreases the honor-stigma gap, *drawing in* liars from k^* and any other low state. If this effect is strong enough, it may lead to an increase in the threshold k^*. However, since increasing $p(1)$ provides a share of agents with a higher material incentive to lie, overall lying increases, reducing credibility. The total effect is ambiguous. If the honor-stigma effect is too weak, the credibility effect dominates, and k^* decreases. This is the intuition behind Part (i) of the proposition below. In the deed-based model, only the credibility effect is present, implying that k^* decreases, as stated in Part (ii). This illustrates that in the character-based model, it not only matters *how many* lie but also *who* lies, which is not the case in the deed-based model.[24]

Proposition 3b. *Suppose that $k^* > 1$ and that states are uniformly distributed on $\mathcal{K}$. When redistributing probability mass $\tilde{\delta} \in (0, 1/K)$ away from k^* towards 1:*

(i) *With character-based image concerns, the likelihood that an agent lies increases while k^* may increase or decrease.*
(ii) *With deed-based image concerns, the likelihood that an agent lies increases while k^* weakly decreases.*

3.2.2. Changing the moral type distribution

We turn to comparative statics concerning the moral type distribution. K&S show that, in the deed-based model, an increase in the moral type distribution in the sense of first-order stochastic dominance (F.O.S.D.) decreases the likelihood that an agent lies and increases k^*.

Proposition 4a. *(Khalmetski and Sliwka, 2019, Proposition 7) Suppose that agents hold deed-based image concerns. If the distribution of lying costs F increases in the sense of F.O.S.D., then:*

(i) *The likelihood that an agent lies strictly decreases.*
(ii) *The threshold state k^* weakly increases.*

The intuition behind Part (i) is straightforward; an F.O.S.D. increase in the lying cost makes lying more costly throughout the type distribution, which reduces lying. Part (ii) is related to the credibility effect. Because fewer agents lie after the increase, reporting K becomes more credible, which increases the reputation associated with it. This in turn weakly decreases the range of states reported by liars, because they have less of a need to report a state different from K to disguise their lie.

The additional honor-stigma effect makes related comparative static comparisons of the character-based model more complex. In two examples below, I illustrate how it can lead to predictions that disagree with the deed-based model. Whenever discussing the character-based model in this subsection, I assume that the moral type distribution is uniform on $(\underline{t}, \bar{t})$, which I denote by $F_U(t, \underline{t}, \bar{t})$. I comment on the reasons and consequences of this assumption after presenting results.

As a first comparative static for the character-based model, consider a rightward shift of the moral type distribution. This shift essentially increases the lying cost of every agent by the same amount.

Proposition 4b. *Consider moving from a lying cost distribution $F_U(t, 0, \bar{t})$ to a lying cost distribution $F_U(t, c, \bar{t} + c)$, where $c \in (0, \hat{t}^*_{k^*})$:*

(i) *The likelihood that an agent lies strictly decreases.*
(ii) *The threshold state k^* weakly increases.*

As the proposition shows, shifting the lying cost distribution to the right in the character-based model has the same effect as increasing lying costs in the sense of F.O.S.D. in the deed-based model: Higher lying costs imply that fewer agents lie. This makes reporting K more credible. At the same time, while, for fixed reporting behavior, the reputation *levels* associated with reporting any state increase, the honor-stigma *gaps* between reporting different states remain the same. Therefore, the credibility effect dominates, which explains the increase in k^*. A rightward shift of the lying cost distribution increases lying costs in the sense of F.O.S.D. We conclude that a distributional shift that increases lying costs in the sense of F.O.S.D. can have the same effect in the character-based and deed-based models.

Now suppose increasing the upper truncation point $\bar{t}$ of the lying cost distribution, holding the lower truncation point constant. This also increases the lying cost distribution in the sense of F.O.S.D. However, the resulting comparative static prediction for the character-based model disagrees with the prediction of the deed-based model.

---

pro-sociality, into the model. The consequence might be that a "large" lie is more stigmatized than a "small" lie because agents who take from someone else signal that they care little about the welfare of others. See, e.g., Cohn et al. (2019) for further discussion and evidence that individuals are less likely to cheat for a large gain than for a small gain when they believe that someone else will suffer from it.

[24] A similar result can be derived for the case where probability mass is redistributed from K to 1. However, while all behavioral effects occur through changes in the type composition of liars when redistributing probability mass away from k^*, redistributing probability mass away from K *directly* reduces credibility because it decreases the probability of observing the highest state. This may make the theoretical prediction that the honor-stigma effect can dominate the credibility effect when redistributing probability mass from K to 1 empirically less relevant.

Proposition 4c. *Consider moving from a lying cost distribution $F_U(t,0,\bar{t})$ to a lying cost distribution $F_U(t,0,\bar{t}+c)$, where $c>0$:*

(*i*) *The likelihood that an agent lies strictly decreases.*
(*ii*) *The threshold state k^* weakly decreases.*

The intuition for Part (*i*) mirrors the preceding comparative statics; increasing lying costs reduces lying. Part (*ii*) is different because this comparative static only increases the lying cost of agents in the right tail of the distribution but not that of agents who are in the left tail. As a consequence, the shift increases the expected moral type of truth-tellers while leaving the expected moral type of liars, who are located in the left tail, unchanged. For fixed reporting behavior, this increases the reputation associated with reporting a low state by more than it increases the reputation associated with reporting a high state. Therefore, the honor-stigma gap increases, which broadens the range of states reported by liars. This shows how the comparative statics predictions of the character-based model with respect to an F.O.S.D. increase in the moral type distribution can differ from those of the deed-based model.

Remark: Consequences of assuming a uniform type distribution. The uniform assumption ensures that, for $j \leq k^*$, $\hat{t}_j$ and $F(\hat{t}_j)$ are both affine functions with respect to the lying cost distribution shifts investigated here. This implies that these shifts change the lying cost thresholds and the fraction of agents lying after observing j by the same amount for any $j \leq k^*$. This keeps the comparative statics analysis tractable. For other distributions, there may be an additional effect because shifts in the lying cost distribution might interact with the levels of $\hat{t}_j$ and $F(\hat{t}_j)$. These levels in turn depend on the model primitives, such as the shape of $y(a)$ and the value of μ which makes it difficult to derive any clear-cut predictions under more general distributional assumptions. While parts (*i*) of both comparative statics above can be shown to hold for more general distributions, one may worry about using the results of parts (*ii*) above when expecting highly nonlinear lying cost distributions. However, the results still illustrate how, due to the additional honor-stigma effect, the predictions of the character-based models can differ from those of the deed-based model when investigating shifts in the lying cost distribution.

Applications to belief-based interventions. Agents in the model prefer appearing as a high type over appearing as a low type. The reputational stigma of making a dishonest report depends on the distribution of types and on beliefs that agents and the audience hold about it. We will now apply the above results to investigate how changing beliefs about the type distribution affects behavior. In particular, we will consider an agent with type (j,t) who faces any of the three distribution shifts discussed above and ask how this type adjusts their behavior.

One interpretation of the following comparative statics is to think of moving an agent from a population with a certain preference distribution to a population with a different preference distribution and asking how this agent adjusts their behavior (see, e.g., Adriani and Sonderegger, 2019). However, the comparative statics also apply if we are willing to entertain a non-equilibrium solution concept where agents best respond to their subjective second-order belief about the audience's belief about the type distribution. Seen in this light, a comparative static that shifts a feature of the preference distribution while fixing the agent's type can be more literally interpreted as a shift in the agent's second-order belief. Such shifts might occur after a norms-based intervention that aims to correct agents' misperceptions about average behavior (Bénabou and Tirole, 2011). Alternatively, following Bénabou et al. (2020),[25] shifts in agents' second-order beliefs could be brought about by third parties who persuade agents to hold a certain belief about the preference distribution by using narratives.

The following result is a direct corollary of Propositions 4a–4c. Whenever the models predict that a shift in the moral type distribution weakly increases k^*, this suggests that agents have become more willing to trade off the material gain of reporting K against a decreased image. Therefore, if we fix a type (j,t) and ask how the likelihood that this type lies changes in response to shifts in the distribution, the comparative statics will go in the same direction as the comparative statics for k^*.

Corollary 1. *Consider an agent with type $\theta=(j,t)$.*

(*i*) *Suppose that the agent holds a deed-based image concern and consider the distribution shift discussed in Proposition 4a (F.O.S.D. increase in the distribution). Type θ becomes more likely to lie.*
(*ii*) *Suppose that the agent holds a character-based image concern.*
  (*a*) *Consider the distribution shift discussed in Proposition 4b (shifting the distribution to the right). Type θ becomes more likely to lie.*
  (*b*) *Consider the distribution shift discussed in Proposition 4c (increasing $\bar{t}$). Type θ becomes less likely to lie*

As an illustration of the results, consider an agent who is exposed to a narrative that "nobody is perfect". That is, everyone might lie if their incentives are strong enough. We can think about this narrative as reducing the agent's belief about the level of the highest moral type, redistributing some probability mass out of the right tail of the distribution (i.e., the part where the "perfect" types are

[25] Bénabou et al. (2020) study a case where narratives can shift agents' beliefs about the size of the externality of an action they take, while I look at narratives that shift agents' beliefs about the type distribution. The paper by Bénabou et al. (2020) is part of an emerging recent literature that investigates the effect of narratives on behavior. Other related papers are Eliaz and Spiegler (2020); Foerster and van der Weele (2021), and Schwartzstein and Sunderam (2021). Golman (2023) fully specifies the equilibrium of a game where agents express potentially controversial opinions and tailor interpretations of past data to increase their reputational utility.

located) to the left tail. Part (ii) (b) above suggests that, as a consequence, the agent becomes more likely to lie.[26] When being told the "nobody is perfect" narrative, the agent also comes to believe that the likelihood that other agents would lie increases; after all, the agent reduces their belief in the morality of others. This indicates that, in this example, beliefs about the actions of others and the agent's own action are complements. This is fundamentally different from the deed-based model which always predicts substitutability: The result in Part (i) shows that, with deed-based image concerns, an agent always becomes more likely to lie if they reduce their belief that others would lie.

It is, however, also not true that beliefs about the lying of others *always* complement an agent's own lying propensity in the character-based model. We can see this in Part (ii) (a). Here, shifting the moral type distribution to the right decreases the agent's belief that others would lie, yet it *increases* the agent's lying propensity. The reason for this is that the "nobody is perfect" narrative mainly affects behavior by reducing the honor-stigma gap between lying and truth-telling, while the rightward shift mainly affects behavior by increasing the credibility of high reports.[27]

Can we apply these insights to a setting that is not as stylized as in the model? In empirical applications, it would be difficult to measure the underlying preference distribution and beliefs about it. However, sometimes it is possible to observe past actions of others, be it by measuring lying in the lab and exposing future participants to that data or by estimating, e.g., the level of tax income misreporting from household consumption data. If evidence of high levels of cheating is interpreted as evidence that truth-telling is not very diagnostic of honor (as in the "nobody is perfect" narrative), this can reduce truth-telling. If an interpretation of the same data instead makes individuals aware of the high level of suspicion they will raise by making a report that is made by an implausibly high number of individuals, then it will increase truth-telling. We might thus expect different actors to make arguments that either justify lying by claiming that others would have behaved in the same unethical way in a similar situation or that encourage truth-telling by stressing the incredibility of high reports.

The role of type uncertainty. Moving beyond F.O.S.D., consider an audience that knows the history of agents' actions, which she can use to reduce her prior uncertainty about the agents' types. How do agents adjust their behavior in response to the audience's new beliefs? To study the role of decreased uncertainty about moral types, we will investigate the effects of reducing the dispersion of its distribution in the sense of a mean-preserving contraction. As before, we will do this for the family of uniform type distributions. To anticipate the intuition behind the following comparative static, think about taking noise out of an initial moral type distribution. The resulting less dispersed distribution will have thinner left and right tails than the initial distribution. As a consequence, the conditional expectations $\mathcal{M}^+(t)$ and $\mathcal{M}^-(t)$ will take on less extreme values. This in turn decreases the honor-stigma gap, which leads to the following result.

Proposition 5a. *Consider moving from a lying cost distribution $F_U(t,0,\bar{t})$ to a lying cost distribution $F_U(t,c,\bar{t}-c)$, where $c \in (0, min\{\hat{t}^*_{k^*}, \frac{\bar{t}}{2}\})$:*

(i) *The threshold state k^* weakly increases.*
(ii) *Fixing a type $\theta = (j,t)$, this type becomes more likely to lie.*

In the character-based model, agents want to convince the audience that they are of a high type. As the audience's prior becomes more certain, agents have less room to move the audience's prior by taking any particular action. Their actions in turn are less guided by image concerns. This makes partial lies less likely and increases any fixed type's incentive to lie. As this comparative static is driven by the honor-stigma effect, it is not predicted by the deed-based model. Instead, the predictions of the deed-based model are ambiguous as the credibility effect, depending on circumstances, can be positive, negative, or equal to zero.

Proposition 5b. *Suppose that agents hold deed-based image concerns and consider reducing the dispersion of F in the sense of a mean-preserving contraction.*

(i) *If the initial likelihood that an agent lies is sufficiently high:*
  (a) *The threshold state k^* weakly decreases.*
  (b) *Fixing a type $\theta = (j,t)$, this type becomes less likely to lie.*
(ii) *Otherwise:*
  (a) *The threshold state k^* weakly increases.*
  (b) *Fixing a type $\theta = (j,t)$, this type becomes more likely to lie.*

[26] Note that the distribution shift discussed in the text is the opposite of the distribution shift discussed in the corollary, such that a *reduction* in the highest type must lead to an *increase* in the likelihood of lying.

[27] As a referee helpfully points out, if we consider increasing the lower truncation point of the moral type distribution instead of decreasing the upper truncation point, the predictions of the character- and deed-based models coincide because the honor-stigma effect will go in the same direction as the credibility effect: Increasing the lower truncation point increases the expected moral type of liars, decreasing the honor-stigma gap. This makes lying more attractive. At the same time, it also shifts relative probability mass from lower to higher moral types, reducing lying and increasing the credibility of a high report. This also makes lying more attractive. Therefore, the character- and deed-based models predict that increasing the belief about the lower truncation point increases the likelihood of lying. This is an example of a case where agents become more likely to lie after they reduce their belief about the lying propensity of others.

There is an interesting connection between the most recent comparison and the influential criminological theory by Braithwaite (1989) (see also Makkai and Braithwaite, 1994). Braithwaite distinguishes between reintegrative and disintegrative shaming. Shaming is reintegrative if it condemns a moral transgression but does not make inferences about the personal traits of the transgressor based on the transgression (what we may call deed-based). Shaming is disintegrative if it generalizes from transgressions to the personal traits of the transgressor (what we may call character-based). In this theory, disintegrative shaming leads to worse outcomes as transgressors are labeled as deviants, and expectations about their deviant character stay attached to them. Transgressors in turn become more likely to re-offend. The comparison between the character- and deed-based models may be seen as giving a formal rationale for that distinction. The point is that in a population that mostly focuses on character-based image, signaling incentives, and thus truth-telling, decrease as the audience forms more precise priors.[28]

3.3. Applications

This section considers two applications of the model that embed it in a broader context. We first analyze the effects of different forms of lie detection and disclosure. Secondly, we consider a selection game, where agents first indicate whether they are interested in participating in a lying game or not before they possibly participate. As I will show, the character- and deed-based models make different predictions in these applications. In the first application, this has implications for optimal verification and disclosure design. The results from the second application could be especially useful for experimental researchers who aim to conduct sharp empirical tests of the character- and deed-based model.

In both applications, off-equilibrium beliefs do play a role in the analysis. I will use the refinement of Dufwenberg and Lundholm (2001) to determine off-equilibrium beliefs.

Definition 5. If a is an out-of-equilibrium action, then

$$\mathcal{R}_a^C = \tilde{t} \in (\tilde{j}, \tilde{t}) = \underset{(j,t)\in\mathcal{K}\times(\underline{t},\tilde{t}]}{\arg\min}\ u(j,t,a^*(j,t),\mathbb{E}(\boldsymbol{t}|a^*(j,t))) - u(j,t,a,t).$$

Under the refinement, the audience attributes the out-of-equilibrium action to the type that faces the smallest utility loss from deviating from their optimal to that action.

3.3.1. Verification and disclosure of lies

If individuals care about their image, they should react to threats of being verified and publicly exposed as a liar. This has motivated authors to promote raising the salience of caught lies in the policy mix to increase honesty (e.g., Abeler et al., 2019). Such policies are, for example, already used by some US States that maintain publicly accessible websites that list the names and addresses of individuals who accumulated tax debt (Perez-Truglia and Troiano, 2018). With character-based image, agents are sensitive to how their lies will be disclosed after verification. This section discusses how the type of disclosure policy might matter.

Consider an additional player in the game, the investigator. After reports are made, the investigator detects the state observed by any agent with probability $\pi \in (0,1)$. In its most basic form, the investigator could rely on *coarse disclosure* and disclose whether the agent lied or not, but not the state observed by a liar. Such a regime results in an image of $\mathcal{L}(\hat{t})$ for a disclosed liar. The expected reputation of a liar reporting K then becomes

$$\mathbb{E}[\boldsymbol{\mathcal{R}}_{\boldsymbol{K}}^{\boldsymbol{C}}|\text{observe } j < K] = (1-\pi)(r_K\mathbb{E}(\boldsymbol{t}) + (1-r_K)\mathcal{L}(\hat{t})) + \pi\mathcal{L}(\hat{t}).$$

As they gain a lower reputation when disclosed as a liar, agents prefer not being disclosed as a liar to being disclosed. Introducing verification and disclosure thus reduces lying because it reduces its (expected) credibility. It also makes partial lying less attractive as partial liars are as likely as full liars to be caught lying so the reputational advantage of partial over full-extent lying becomes smaller.[29]

Proposition 6a. *After an increase in the probability of lie detection π:*

(*i*) *The threshold state k^* weakly increases.*
(*ii*) *The likelihood that an agent lies decreases.*

[28] Experimental evidence suggests that lying becomes more prevalent in repeated environments. In their meta-study, AN&R report a small, but significantly positive, coefficient of the round of repetition on reporting. However, there are at least two concerns with interpreting this finding as being consistent with the character-based model: First, experimental participants usually know in advance that they will repeat the lying task, and it is not clear how forward-looking their behavior is. Second, the experimenter typically inspects the report sequences only after the experiment, so it would be wrong to think of the experimenter as an audience who updates her belief after every single report.

[29] An interesting extension of the model could consider an investigator who, faced with a distribution of reports, can choose to verify a fixed fraction of reports. If the goal is to maximize the lie detection rate the investigator should disproportionally focus on investigating reports of the highest state. This could, contrary to the present result, encourage partial lying.

Assume for the rest of the section that the prior type distribution is uniform.[30] Since the investigator observes the state, they could additionally commit in advance to disclosing it with some probability $\gamma \in (0,1)$. Such *contextualized disclosure* would result in an image of $\mathcal{M}^{-}(\hat{t}_j)$ for caught liars. Consider going from the coarse to the contextualized regime. For liars from the lowest states, the honor-stigma gap decreases because their expected moral type is larger than that of the average liar. They therefore become more likely to lie. The honor-stigma gap will instead increase for liars from higher states, who have a smaller expected moral type than the average liar. They become less likely to lie. These first-order effects lead to an increase in the average size of the lie.

Now consider an agent who observes one of the lowest states. The direct effect of introducing the contextualized disclosure regime encourages them to lie because they can separate from other liars in case of disclosure. Albeit this direct effect is there, it is also relatively small; agents from the lowest states are overrepresented among liars, so that, already under coarse disclosure, it is likely that a disclosed liar observed a low state. This makes their experienced decrease in the honor-stigma gap small. In contrast, the reputational penalty of going to contextualized disclosure is relatively harsher for agents from higher states as they only constitute a minority of liars. Therefore, the direct effect of going to contextualized disclosure has a larger behavioral effect on "small" liars who reduce their lying, than on "large" liars who increase their lying.

Choosing between coarse or contextualized disclosure thus constitutes a tradeoff between minimizing the total lying rate and the average size of lies.

Proposition 6b. *Suppose that lies are detected with probability $\pi > 0$ and that $t \sim U(0,\bar{t})$. After an increase in the probability of disclosing the state γ:*

(*i*) *The average size of lies increases.*
(*ii*) *The lying rate decreases.*

Relation to deed-based image concerns. Under deed-based image, introducing a nonzero verification probability also reduces lying. Notice how the expected reputation of a liar reporting K becomes

$$\mathbb{E}[\boldsymbol{\mathcal{R}}_{\boldsymbol{K}}^{\boldsymbol{D}} | \text{observed } j < K] = (1-\pi) \times r_K + \pi \times 0,$$

i.e., reporting K comes with a lower expected credibility as π increases. However, adding contextualized disclosure will not affect behavior because the audience does not differentiate between different types of liars—the equation above shows that the reputation awarded to a disclosed liar is equal to 0, *independent of their observed state j*. Therefore, providing additional context about the disclosed liar does not influence the audience's judgment. We summarize these insights in the following result.

Proposition 6c. *Suppose that agents have deed-based image concerns. After an increase in the probability of lie detection:*

(*i*) *The threshold state k^* weakly increases.*
(*ii*) *The likelihood that an agent lies decreases.*

Lying behavior is invariant to changes in the probability of disclosing the observed state γ.

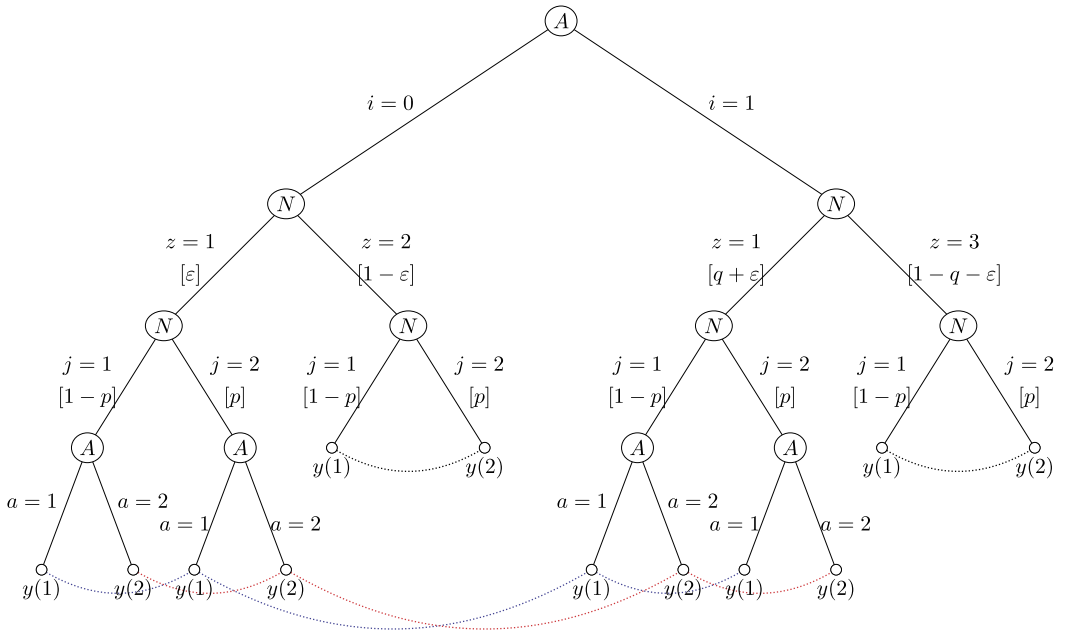
3.3.2. Selection into lying opportunities

This section studies lying in the context of selection. I will show that if agents have some control over whether they will participate in a lying game or not, the character-based model generates comparative statics that are different when compared to a model that studies selection with deed-based image concerns and when compared to a setup without selection. These results should be of interest to applied researchers who study lying in the context of selection and might be useful to experimentalists who want to distinguish between character- and deed-based image concerns in the lab.[31]

We embed the lying model in a selection game, where agents first make a participation decision $i \in \{0,1\}$ that indicates their interest in participating in the lying game ($i = 1$) or not ($i = 0$). To focus on essentials we consider a binary lying game with $\mathcal{K} = \{1,2\}$. After the participation decision, nature makes two moves. First, it makes a random move that determines whether the agent participates in the lying game or not—we will discuss this in detail below. Second, it draws a state $j \in \mathcal{K}$ with the probability of $j = 2$ being $p \in (0,1)$ and the probability of $j = 1$ being $1 - p$. If they take part in the lying game, agents then report $a \in \mathcal{K}$ and earn a material payoff $y(a)$. Agents who do not participate in the lying game do not report and instead earn a material payoff $y(j)$ that corresponds to the realization of nature's draw. This setup is similar to the design that Houdek et al. (2021) use to experimentally study selection into lying opportunities. In their study, participants are offered a lottery where they have to correctly predict the realization of a random number. They can choose to participate in a nonmanipulable version of the lottery where they first write

[30] This is mainly for ease of exposition. Similar results can be derived for different distributions.

[31] Selection has been studied in the context of public sector jobs with Hanna and Wang (2017) providing evidence that, in India, more dishonest individuals select into public sector jobs while Barfort et al. (2019) find that, in Denmark, more dishonest individuals instead select out of public sector jobs. These findings are possibly explained by differences in the corruption opportunities that the public sector offers in different countries. Lab studies have used selection tasks to structurally relate individual preferences to lying behavior. Konrad et al. (2021) elicit the WTP to participate in a lying game to recover individual lying costs, and Houdek et al. (2021) study selection into cheating opportunities to systematically relate individual-level characteristics to the selection choice and study interventions.



Note: End nodes show material payoffs.

Fig. 3. Selection game. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.)

down their prediction and then observe the realized number, or in a manipulable version of the lottery where they first observe the number and then write down their prediction. The manipulable version of the lottery is a lying game since individuals can adjust their predictions after observing the state. I complicate this setting by weakening the relationship between the participation decision and actual participation in the lying game. This will allow for a clear distinction between the character- and deed-based models. Consider the game tree in Fig. 3. As its first decision, nature draws a number z. If the agent does not indicate interest, nature chooses $z = 1$ with probability $\varepsilon \in (0, 1)$ and $z = 2$ with probability $1 - \varepsilon$. If an agent indicates interest, nature either chooses $z = 1$ or $z = 3$, with probabilities $q + \varepsilon \in (\varepsilon, 1)$ and $1 - q - \varepsilon$, respectively. Agents only participate in the lying game if nature chooses $z = 1$. Therefore, by indicating their interest, agents can increase the probability that they will participate in the lying game by q. After all decisions have been taken, the audience observes the realization of z and the agent's lying game report (if applicable) and makes an inference about the moral type.[32]

The information sets drawn at the end nodes of the game tree display the audience's information at the end of the game. Conditional on not participating in the lying game, the audience can tell whether an agent indicated interest (the realization of z is 3) or not (the realization of z is 2). Participation in the lying game is, in contrast, not a perfect signal of indicated interest, because $z = 1$ can occur after any participation decision. In the following, we will show results for the case where $\varepsilon \to 0$, i.e., where participation is an almost perfect signal of indicated interest. This setup allows us to sharply distinguish between different signaling motives.[33] As before, we assume that the image weight is small enough to ensure a unique equilibrium.

Proposition 7a. *As $\varepsilon \to 0$, the equilibrium of the selection game is characterized by two thresholds $\hat{t}_C < \hat{t}_{LG}$ which are defined in*

$$\Delta(2,1) - \hat{t}_C = \frac{\mu}{q(1-p)} \left(\mathcal{M}^+(\hat{t}_C) - \mathcal{M}^-(\hat{t}_C) \right),$$
$$\Delta(2,1) - \hat{t}_{LG} = \mu(\mathcal{M}^+(\hat{t}_{LG}) - \mathcal{M}^-(\hat{t}_C))$$

and agents choose

[32] To implement this game in a lab experiment, consider a design where participants first privately choose to draw from one of two urns. Urn 1 is filled with red and blue balls and Urn 2 is filled with red and yellow balls. Participants draw once and show their drawn ball to the experimenter, who lets them participate in a manipulable lottery if they show a red ball and in a nonmanipulable lottery if they show a blue or yellow ball. Including a fraction of ε red balls in Urn 1 and a fraction of $\varepsilon + q$ red balls in Urn 2 would implement the game tree in Fig. 3.

[33] It also helps us to sidestep an additional discussion about the possible nonexistence of equilibria. Suppose that $\varepsilon = 0$ and consider a candidate equilibrium where only agents that are planning to report 2 indicate their interest. The reputation associated with reporting 1 is not pinned down in the candidate equilibrium because no type reports it. While the equilibrium refinement from Definition 5 can be used to determine the off-equilibrium belief, whether an equilibrium exists or not will depend on properties of the moral type distribution $f(t)$. Allowing that $\varepsilon > 0$ but possibly very small instead ensures that the reputation associated with reporting 1 is pinned down by the expected moral type of an agent who by chance participated in the lying game, even though they did not indicate their interest.

$i = 1$, $a = 2$ *if* $t \leq \hat{t}_C$ *and for any* j,

$i = 0$, $a = 2$ *if* $t \in (\hat{t}_C, \hat{t}_{LG}]$ *and for any* j,

$i = 0$, $a = j$ *if* $t > \hat{t}_{LG}$ *and for any* j.

The equilibrium has the following properties:

(i) *The likelihood that an agent indicates interest in the lying game increases in* q.
(ii) *The likelihood that an agent indicates interest in the lying game decreases in* p.
(iii) *The likelihood that an agent lies in the lying game conditional on observing 1 is invariant in* p.

In equilibrium, agents indicate their interest in the lying game only if they are planning to cheat, i.e., if their moral type is sufficiently low. This makes the participation decision informative about an agent's type—among those who do not participate, those who indicated interest (where nature chose $z = 3$) consequently receive an image proportional to $\mathcal{M}^-(\hat{t}_C)$, which is lower than the image conferred to those who neither indicated interest nor participated in the lying game (where nature chose $z = 2$) and whose image is proportional to $\mathcal{M}^+(\hat{t}_C)$. Indicating interest is stigmatized. Agents trade off this indicated-interest stigma against the expected material payoff gain of indicating interest. This can be calculated as follows: Indicating interest raises the probability of participating in the lying game by q. Conditional on participation, agents gain in material payoff relative to not participating with probability $1 - p$, when they observe 1 and report 2. In all other cases, indicating interest has no consequences for the material payoff. Therefore, indicating interest only increases the material payoff with probability $q(1 - p)$. Parts (i) and (ii) of Proposition 7a now follow quite naturally: Increasing q increases the expected material payoff gain from indicating interest while increasing p decreases it. Changes that increase the expected material payoff motivate a larger fraction of agents to accept the associated stigma and indicate interest.

Part (iii) immediately follows since everyone who participates in the lying game lies. This prediction is very different from what the character-based model would predict without selection; there, an increase in p encourages those who observe 1 to lie. The reason for the difference is that, without selection, liars who report 2 pool with a random sample of non-selected types who are of a higher expected moral type. This random sample becomes relatively larger as p increases, which makes reporting 2 more credible and lying thus more attractive. In the selection game, in contrast, even those who honestly report 2 in the lying game have a low image because, by participating in the lying game, they already signal that their lying costs are low. Indeed, in the equilibrium of the selection game, among those who report 2, liars and truth-tellers are of the same expected moral type. Therefore, liars do not enjoy a pooling advantage when reporting 2 in the selection game, which mutes the credibility effect.

Relation to deed-based image concerns. The predictions of the character-based model are different from the predictions of the deed-based model, as Proposition 7b clarifies. Since only the reporting decision is a relevant signal about whether an agent tells the truth or lies, there is no stigma associated with indicating interest. Thus, agents do not face a tradeoff between expected material payoff gain and reputational stigma when making their participation decision. Participation is therefore independent of p and q, in contrast to the character-based model. As a second difference, the likelihood that an agent lies after observing $j = 1$ increases in p. Essentially, regardless of the selection process, agents with deed-based image concerns want the audience to believe that they are telling the truth. As p increases, reporting 2 becomes more credible, which makes lying more attractive.

Proposition 7b. *Suppose that agents have deed-based image concerns. As* $\varepsilon \to 0$*, the equilibrium of the selection game is characterized by a threshold* $\hat{t}_D$ *which is defined in*

$$\Delta(2,1) - \hat{t}_D = \mu\left(1 - \frac{p}{p + (1-p)F(\hat{t}_D)}\right)$$

and agents choose

$i = 1$, $a = 2$ *if* $t \leq \hat{t}_D$ *and for any* j,

$i = 1$, $a = j$ *if* $t > \hat{t}_D$ *and for any* j.

The equilibrium has the following properties:

(i) *The likelihood that an agent indicates interest in the lying game is invariant in* q.
(ii) *The likelihood that an agent indicates interest in the lying game is invariant in* p.
(iii) *The likelihood that an agent lies in the lying game conditional on observing 1 increases in* p.

With deed-based image concerns, all agents indicate interest in the lying game. They do this because they can only signal an honest action in a situation where they have the option to lie. As the image utility always adds something positive to agents' total utility, agents prefer sending any signal to not sending a signal. A different specification of the deed-based model where the image utility enters total utility as a cost (i.e., where agents face costs of being suspected of lying; Khalmetski and Sliwka, 2019) instead predicts the same selection pattern as the character-based model. Agents indicate interest only if they are planning to lie. However, also in this alternative deed-based model, only the lying game decision sends a signal to the audience. The Online Appendix formalizes that,

as a consequence, both variants of the deed-based model predict that the selection decision does not depend on p and q. Therefore, the predictions of the character-based model are also different from an alternative specification of the deed-based model that is closer to the character-based model in terms of selection behavior.

4. Discussion

This paper presented a model where agents derive reputational esteem from being perceived as an honest character. Such a model can explain many of the previous experimental results on lying games. Differences with other lying models emerge because agents' signaling motives (credibility vs. honor-stigma) differ. The results are illustrated in applications to the behavioral effects of norm interventions or narratives, they make predictions about the short- and long-term effects of different shaming conventions, have implications on how lies should be disclosed, and predict how selection into lying opportunities affects behavior.

4.1. Extensions

Two simplifying assumptions were maintained throughout the analysis; that intrinsic lying costs are fixed and that agents care to the same extent about image utility. I will now briefly discuss the consequences of relaxing these assumptions.

Behavior when lying costs are not fixed. Non-fixed lying costs have been studied in the context of the deed-based model by GK&S and K&S. Both papers provide results for the case where lying costs consist of a fixed, moral type-dependent and a variable, moral type-independent component. For example, K&S assume that lying costs increase linearly in the distance between the report and the observed state. They show that, compared to a model with only a fixed lying cost, all equilibrium features of the deed-based model remain qualitatively the same. It is relatively straightforward to show that the same results translate to the character-based model. As long as variable lying costs do not interact with the moral type, they will not fundamentally change equilibrium behavior. The Online Appendix provides formal results for the case where the moral type interacts with the variable lying cost. That is, agents face a higher marginal cost of lying if they are of a higher moral type. Under this assumption the equilibrium prediction is that agents report any state but 1 dishonestly with positive probability. The Online Appendix argues that this prediction, however, is not particularly realistic in light of the experimental evidence that we have on behavior in observed lying games.

Heterogeneous image concerns. The Online Appendix also provides results that relax the homogenous image concern assumption. Consequences of such a relaxation have been explored by Zakharov (2023) in the context of the deed-based model. When different agents care about their image to different extents in the character-based model, partial lying can still emerge in equilibrium, but it will be of a slightly different kind. Remember how in the baseline analysis, liars are indifferent between reporting any state that is reported dishonestly with positive probability in equilibrium. With heterogeneous image concerns, this is no longer the case: some agents will value image utility more than others, which leads them to strictly prefer partial to full lying. The resulting equilibrium thus predicts that liars separate by their image type. For example, the least image-concerned liars report K while more image-concerned liars report $K-1$. Heterogeneous image concerns can also lead to downward lying. Since a highly image-concerned agent will prefer honestly reporting $K-1$ over honestly reporting K, they might also prefer dishonestly reporting $K-1$ after observing K if their intrinsic lying cost is sufficiently low. With heterogeneous image concerns, the character-based model can also account for experimental results from die-roll games that in some cases document a report distribution where the mode is smaller than K.[34] Such a reporting pattern seems puzzling when seen through the lens of a deed-based model since deviating from the mode towards reporting K would increase the material payoff and lower the audience's suspicion. The motivation for agents with character-based image concerns to report the mode purely follows from the honor-stigma motive.

4.2. Going forward

In addition to offering new theoretical insights, the model also generates several testable predictions. Going ahead, I identify three types of possible future research that could be informed by the theoretical lessons from this paper.

First, future experiments could address specific behavioral mechanisms identified by the theory and measure their empirical relevance. Section 3.3.2 provides new comparative statics results for the selection game, where the character- and deed-based models make fundamentally different predictions. One could translate the selection game into a lab experiment to measure how participants respond to changes that make the participation decision more consequential (that increase q) and that change the state distribution (that increase p). Such an experiment could provide a clean test of the empirical relevance of the different propositions.[35]

[34] Out of 24 papers included in the AN&R meta-study that employ a one-shot die-roll lying game, 8 contain experiments where the highest state is not the modal report. Most of these experiments have been conducted outside traditional lab environments in settings where the social distance between the audience and participants is arguably lower and where the image motive thus might play a greater role. For example, Ruffle and Tobol (2014) conduct an experiment with Israeli soldiers who have to report the outcome of a die roll to an army official. The higher the reported die roll, the earlier the soldiers will be released from duty on one weekday afternoon. They find that some soldiers lie to the army official and that most of them report the second-highest state.

[35] The idea that experiments mainly serve to test the empirical relevance of different propositions is borrowed from Battigalli and Dufwenberg (2022), who argue that different theories are not necessarily mutually exclusive, with their relative empirical relevance depending on the situation. This paper takes a similar stance concerning character- and deed-based image motivations, as there are plausible contexts that would favor one or the other model, as argued in the introduction of this paper.

Studying lying in contexts where individuals make more than one decision is also of practical relevance. The character-based model cautions us that in situations where people with a "history" face a cheating opportunity, image-based interventions might have unintended consequences. If the side product of an image-based intervention is to record and facilitate access to individual decision histories, the character-based model suggests that this can increase lying if it reduces type uncertainty (as shown in Proposition 5a). In the selection context of Section 3.3.2, the character-based model suggests that an intervention that reduces the credibility of reporting a high state might not have the desired impact of increasing honesty if agents have some agency over whether they find themselves in a situation where they can make such a report.

Second, future experiments could not only try to identify preferences but also the strategic reasoning of individuals who hold these preferences. In the current context, experiments that reinforce or create certain signaling motives through monetary incentives seem attractive. For example, introducing an investigator who might disclose and punish liars could serve to increase the credibility motive. Conversely, giving participants instrumental motives to appear trustworthy, e.g., by including a stage after the lying game in which participants play a trust game, could increase participants' concern about the composition of types that their report pools them with.

Third, the paper's applications show that beliefs can influence lying behavior for numerous reasons. In the character-based model, in addition to the question of *how many* people lie, questions such as *who* lies and *why* become important. Designs that hold the objective statistical data provided to participants about reporting of others constant but change the interpretation of the data provided along with it (similarly to what Hillenbrand and Verrina, 2022, do in the context of a dictator-giving experiment) could test the behavioral relevance of narratives that aim to raise the credibility or honor-stigma effects.

Declaration of competing interest

The author declares that he has no relevant material or financial interests that relate to the research described in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Proofs

A.1. Proof of Proposition 1

We first provide three lemmas before proceeding with the proof.

Lemma 1 *(General properties of equilibria without the symmetry refinement). In an equilibrium*

(*i*) *If* $s(a=k|j,t)>0$ *and* $s(a=l|j,t)>0$ *for some type* (j,t) *with* $j\neq k$ *and* $j\neq l$*, then* $y(k)+\mu\mathcal{R}_k^C=y(l)+\mu\mathcal{R}_l^C$.
(*ii*) *If there is a type* (j,t) *with* $j\neq k$ *for which* $s(a=k|j,t)>0$*, then* $s(a=k|k,t)=1$ *for all types* (k,t) *and* $s(a=j|l,t)=0$ *for all types* (l,t) *with* $l\neq j$.
(*iii*) *There is a type* (j,t) *with* $j<K$ *for which* $s(a=K|j,t)>0$ *and for all types* (j,t) *with* $j>1$, $s(a=1|j,t)=0$.
(*iv*) *If* $K>2$ *and the ratio* $\Delta(K,K-1)/\mu$ *is sufficiently small, then there is a type who will lie and report a number different from* K.

Proof. (*i*) An agent who observes state j will lie if there is a state k such that

$$y(k)-t+\mu\mathcal{R}_k^C>y(j)+\mu\mathcal{R}_j^C. \tag{A.1}$$

Since $y(K)>y(j)$ for $j<K$, there cannot be an equilibrium where all agents tell the truth. In this case, the reputational payoff would not depend on the reported state, and there would be an agent of type $(j,\underline{t}+\varepsilon)$, where $\varepsilon>0$ is arbitrarily close to zero, who could gain by reporting K. Because lying costs are fixed, agents always can make a report a to gain a gross payoff before lying costs of size $a\in\arg\max_{a\in\mathcal{K}} y(a)+\mu\mathcal{R}_a^C$. These considerations imply point (*i*).

(*ii*) It is useful to define a set

$$\mathcal{H}=\left\{j\in\mathcal{K}|j\in\arg\max_{a\in\mathcal{K}} y(a)+\mu\mathcal{R}_a^C\right\}$$

that collects all states that are reported dishonestly with positive likelihood in equilibrium. If someone who observes j lies, this implies by utility maximization that $j\notin\mathcal{H}$. Therefore, no agent will lie and report j if $s(a\neq j|j,t)>0$ for some type. By the same reasoning, no agent will lie if they observe a state $j\in\mathcal{H}$, as lying is costly and does not lead to higher payoffs.

(*iii*) Consider again the incentive constraint (A.1) and note that the payoff from lying strictly decreases in the lying cost. It follows that an agent lies if their lying cost is sufficiently low. In particular, for each state j there will be a threshold lying cost $\hat{t}_j$ and agents (j,t) will lie if $t\leq\hat{t}_j$, where $\hat{t}_j\geq\underline{t}$. Now consider the reputations that are associated with agents who observed state j. Truth-tellers

comprise the upper tail of the preference distribution, while liars make up the lower tail. Truth-tellers and liars have an expected cost of respectively

$$\mathcal{M}^+(\hat{t}_j) \equiv \mathbb{E}(\boldsymbol{t}|\boldsymbol{t} > \hat{t}_j),$$
$$\mathcal{M}^-(\hat{t}_j) \equiv \mathbb{E}(\boldsymbol{t}|\boldsymbol{t} \leq \hat{t}_j).$$

Part (ii) implies that, if a state is not lied at, its reputation is equal to the expected type of agents who are above the threshold;

$$\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j) \text{ if } j \notin \mathcal{H}. \tag{A.2}$$

Claim 1: $K \in \mathcal{H}$. Suppose the contrary, $K \notin \mathcal{H}$. Then, for all states $j \in \mathcal{H}$,

$$y(j) + \mu\mathcal{R}_j^C > y(K) + \mu\mathcal{R}_K^C, \text{ and } y(K) > y(j). \tag{A.3}$$

This in particular implies that $\mathcal{R}_j^C > \mathcal{R}_K^C$ for all $j \in \mathcal{H}$. From (A.2) it follows that $\mathcal{R}_K^C \geq \mathbb{E}(\boldsymbol{t})$ and more generally $\mathbb{E}(\boldsymbol{t}|\text{report } j \notin \mathcal{H}) \geq \mathbb{E}(\boldsymbol{t})$. By the martingale property of beliefs, it then follows that $\mathbb{E}(\boldsymbol{t}|\text{report } j \in \mathcal{H}) \leq \mathbb{E}(\boldsymbol{t})$, which requires that $\mathcal{R}_j^C \leq \mathbb{E}(\boldsymbol{t})$ for some $j \in \mathcal{H}$.[36] Combining the inequalities, we arrive at $\mathcal{R}_K^C \geq \mathbb{E}(\boldsymbol{t}) \geq \mathcal{R}_j^C$ for some $j \in \mathcal{H}$, which is a contradiction to (A.3).

Claim 2: $1 \notin \mathcal{H}$. Suppose the contrary, $1 \in \mathcal{H}$. Then, for all states $j \notin \mathcal{H}$,

$$y(j) + \mu\mathcal{R}_j^C < y(1) + \mu\mathcal{R}_1^C, \text{ and } y(1) < y(j). \tag{A.4}$$

This in particular implies that $\mathcal{R}_1^C > \mathcal{R}_j^C$ for all $j \notin \mathcal{H}$. Since $\mathcal{R}_1^C$ is a convex combination of the prior and the reputation of liars, the highest value $\mathcal{R}_1^C$ can obtain is smaller than $\max\{\mathbb{E}(\boldsymbol{t}), \max\{\hat{t}\}\} < \mathbb{E}(\boldsymbol{t}|\boldsymbol{t} > \max\{\hat{t}\})$. Since $\mathcal{R}_j^C = \mathbb{E}(\boldsymbol{t}|\boldsymbol{t} > \max\{\hat{t}\})$ for some $j \notin \mathcal{H}$, we arrive at a contradiction to (A.4).

(iv) Consider an equilibrium where $\mathcal{H}$ is a singleton. It then holds that

$$y(K-1) + \mu\mathcal{R}_{K-1}^C < y(K) + \mu\mathcal{R}_K^C,$$

because every liar must prefer to report K over $K-1$. We can rearrange this inequality to

$$\mathcal{R}_{K-1}^C - \mathcal{R}_K^C \leq \frac{\Delta(K, K-1)}{\mu}. \tag{A.5}$$

Since $K-1 \notin \mathcal{H}$, it follows from (A.2) that $\mathcal{R}_{K-1}^C \geq \mathbb{E}(\boldsymbol{t})$. Furthermore, if $\mathcal{H}$ is a singleton then by the martingale property of beliefs, $\mathcal{R}_K^C < \mathbb{E}(\boldsymbol{t})$. The left-hand side of (A.5) is strictly positive. Thus, there is a contradiction if $\frac{\Delta(K,K-1)}{\mu}$ is sufficiently small. □

Consider the function

$$\mathcal{T}(\varphi, \hat{t}, j) = \Delta(K, j) + \mu(\varphi - \mathcal{M}^+(\hat{t})) - \hat{t} \equiv 0.$$

Denote the $\hat{t}$ that solves that equation by $\tilde{t}_j(\varphi)$ and define $\hat{t}_j(\varphi) \equiv \max\{\tilde{t}_j(\varphi), \underline{t}\}$. The function $\hat{t}_j(\varphi)$ implicitly defines a threshold type $(j, \hat{t}_j(\varphi))$ who is indifferent between reporting K and j if the reputation associated with reporting these states are $\mathcal{R}_K^C = \varphi$ and $\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j(\varphi))$, respectively.

Lemma 2 *(Properties of $\hat{t}_j(\varphi)$). The derivative $\frac{\partial \hat{t}_j(\varphi)}{\partial \varphi} \in (0,1)$ if $\hat{t}_j(\varphi) > \underline{t}$ and μ is small enough. The derivative is increasing in μ.*

Proof. $\hat{t}_j(\varphi)$ is implicitly defined in

$$\hat{t}_j + \mu\left[\mathcal{M}^+(\hat{t}_j) - \varphi\right] - \Delta(K, j) = 0. \tag{A.6}$$

Implicitly differentiating the equation brings

$$\frac{\partial \hat{t}_j(\varphi)}{\partial \varphi} = \frac{\mu}{1 + \mu\mathcal{M}^{+\prime}(\hat{t}_j(\varphi))} \text{ if } j \leq k^*,$$

where $\mathcal{M}^{+\prime}(t) > 0$. Therefore, the derivative is between 0 and 1 if μ is small (e.g. $\mu \leq 1$). It also gets clear from taking the cross-derivative with respect to μ that the derivative is increasing in μ. □

[36] The martingale property states that a Bayesian audience never changes her prior on average. In the present context, $\mathbb{E}[\mathbb{E}(\boldsymbol{t}|a)] = \mathbb{E}(\boldsymbol{t})$.

Defining a vector $\hat{t}(\varphi) = (\hat{t}_1(\varphi), \dots, \hat{t}_K(\varphi))$, we can define the expected moral type of liars as

$$\mathcal{L}(\hat{t}(\varphi)) = \sum_{j \in \mathcal{K}} \frac{F(\hat{t}_j(\varphi))}{\sum_{k \in \mathcal{K}} F(\hat{t}_k(\varphi))} \mathcal{M}^-(\hat{t}_j(\varphi)).$$

We now provide properties of $\mathcal{L}$.

Lemma 3 *(Properties of $\mathcal{L}(\hat{t}(\varphi))$). $\mathcal{L}(\hat{t}(\varphi))$ is (i) a continuous function in φ whenever some $\hat{t}_j > \underline{t}$ with (ii) $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$ if μ is small enough. There exists (iii) an interval $(\varphi^{min}, \mathbb{E}(\boldsymbol{t}))$, where*

$$\varphi^{min} = \begin{cases} \mathbb{E}(\boldsymbol{t}) - (\Delta(K,1) - \underline{t})/\mu & \text{if } \mathbb{E}(\boldsymbol{t}) > (\Delta(K,1) - \underline{t})/\mu \\ \xi & \text{otherwise} \end{cases}$$

and $\xi = \mathcal{L}(\hat{t}(\xi))$ is a fixed-point. For all φ on this interval, $\mathcal{L}$ is continuous and $\mathcal{L}(\hat{t}(\varphi)) < \varphi$.

Proof. (i) The functions $\hat{t}_j(\varphi)$ and $\mathcal{M}^-(t)$ are continuous functions. The threshold types $\hat{t}$ can take on values between $[\underline{t}, \bar{t})$ and the c.d.f. $F(t)$ is continuous on $\hat{t} \in (\underline{t}, \bar{t}]$. Since $F(\underline{t}) = 0$ and $\lim_{t \to \underline{t}} F(t) = 0$, $F(t)$ is also continuous on $[\underline{t}, \bar{t}]$. Taking products, sums, and (nonzero) quotients of continuous functions preserves continuity, which ensures that $\mathcal{L}(\hat{t}(\varphi))$ varies continuously with φ whenever some $\hat{t}_j > \underline{t}$, so that the quotient in $\mathcal{L}(\hat{t}(\varphi))$ is nonzero.

(ii) Write the derivative as

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} = \sum_{j=1}^{K} \frac{\partial \mathcal{L}}{\partial \hat{t}_j} \frac{\partial \hat{t}_j}{\partial \varphi}.$$

Lemma 2 shows that $\frac{\partial \hat{t}_j}{\partial \varphi}$ increases in μ. Therefore, there is a small enough μ so that $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$. Appendix E shows that $\mu \leq 1$ is sufficient in case $f(t)$ is log-concave.

(iii) Lemma 1 shows that there is always lying from 1 and that no liar reports 1. This implies that, in equilibrium, $y(K) + \mu\varphi - \underline{t} > y(1) + \mu\mathbb{E}(\boldsymbol{t})$ and therefore in any equilibrium $\varphi > \max\{\underline{t}, \mathbb{E}(\boldsymbol{t}) - (\Delta(K,1) - \underline{t})/\mu\}$. If $\mathbb{E}(\boldsymbol{t}) - (\Delta(K,1) - \underline{t})/\mu \geq \underline{t}$ then it follows that $\mathcal{L}(\hat{t}(\varphi^{\min} + \varepsilon)) < \varphi^{\min} + \varepsilon$ for an arbitrarily small $\varepsilon > 0$ (since $\hat{t}_1(\varphi^{\min}) = \underline{t}$). The assumptions on $\bar{t}$ ensure that agents with the highest moral type never lie even if $\varphi = \mathbb{E}(\boldsymbol{t})$. Therefore, $\mathcal{L}(\hat{t}(\mathbb{E}(\boldsymbol{t}))) < \mathbb{E}(\boldsymbol{t})$. Since $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$, it follows that $\mathcal{L}(\hat{t}(\varphi)) < \varphi$ for all $\varphi \in (\mathbb{E}(\boldsymbol{t}) - (\Delta(K,1) - \underline{t})/\mu, \mathbb{E}(\boldsymbol{t}))$. If $\mathbb{E}(\boldsymbol{t}) - (\Delta(K,1) - \underline{t})/\mu < \underline{t}$ then $\mathcal{L}(\hat{t}(\underline{t})) > \underline{t}$, since $\hat{t}_1(\underline{t}) > \underline{t}$. However, as $\mathcal{L}(\hat{t}(\mathbb{E}(\boldsymbol{t}))) < \mathbb{E}(\boldsymbol{t})$ and $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$ there exists a unique fixed-point $\xi > 0$ at which $\mathcal{L}(\hat{t}(\xi)) = \xi$. Therefore, $\mathcal{L}(\hat{t}(\varphi)) < \varphi$ for all $\varphi \in (\xi, \mathbb{E}(\boldsymbol{t}))$. □

Proof of Proposition 1. Throughout this and the following proofs, we use φ to denote the reputation associated with reporting K ($\mathcal{R}_K^C$).

Claim 1: If $s(a \neq j|j,t) > 0$ for some (j,t) then there is a $\hat{t}_j(\varphi) > \underline{t}$ such that $s(a = j|j,t') = 0$ if $t' \leq \hat{t}_j(\varphi)$, $s(a = j|j,t') = 1$ if $t' > \hat{t}_j(\varphi)$, and the reputation associated with reporting j is $\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j(\varphi))$. First, Lemma 1 (iii) implies that there is at least one type who lies to report K and Lemma 1 (i) implies that if there are two states reported by liars with positive probability, then these two states must yield the same utility gross of lying costs. Combining these observations we find that, if $s(a \neq j|j,t) > 0$, then

$$y(K) + \mu\varphi - t \geq y(j) + \mu\mathcal{R}_j^C.$$

The l.h.s. decreases in t, which suggests that there is a $\hat{t}_j$ such that the equation above holds with equality. Second, if $s(a \neq j|j,t) > 0$ for some (j,t), then Lemma 1 (ii) suggests that no agent who observed $j' \neq j$ reports j. This implies only agents of type $(j, t > \hat{t}_j)$ report j, which suggests that $\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j) > \mathbb{E}(\boldsymbol{t})$ and that $\hat{t}_j$ is defined by $\hat{t}_j(\varphi)$.

Claim 2: The threshold $\hat{t}_j(\varphi) = \underline{t}$ if and only if $s(a = j|j,t) = 1$ for all t. For the if-direction, observe that $s(a = j|j,t) = 1$ implies that

$$y(K) + \mu\varphi - \tilde{t} = y(j) + \mu\mathbb{E}(\boldsymbol{t}),$$

where $\tilde{t} \leq \underline{t}$. By the definition of $\hat{t}_j(\varphi)$, this implies that $\hat{t}_j(\varphi) = \underline{t}$. For the only-if-direction suppose by contradiction that $\hat{t}_j(\varphi) > \underline{t}$. This implies that, for $t \in (\underline{t}, \hat{t}_j(\varphi))$,

$$y(K) + \mu\varphi - t > y(j) + \mu\mathbb{E}(\boldsymbol{t}).$$

These types have a strict incentive to lie, yielding a contradiction.

Claim 3: The reputation $\mathcal{R}_j^C > \mathbb{E}(\boldsymbol{t})$ if and only if $\hat{t}_j(\varphi) > \underline{t}$. The reputation $\mathcal{R}_j^C = \mathbb{E}(\boldsymbol{t})$ if and only if $\hat{t}_j(\varphi) = \underline{t}$ and $s(j|j',t) = 0$ for all $j' \neq j$. The reputation $\mathcal{R}_j^C < \mathbb{E}(\boldsymbol{t})$ if and only if $\hat{t}_j(\varphi) = \underline{t}$ and $s(j|j',t) > 0$ for some $j' \neq j$. We first show the if-direction. First, if $\hat{t}_j(\varphi) > \underline{t}$ then Claim 1 suggests that $\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j(\varphi)) > \mathbb{E}(\boldsymbol{t})$. Second, Claim 2 suggests that, if $\hat{t}_j(\varphi) = \underline{t}$, then no one lies after observing j. In symmetric lying strategies, the equilibrium reputation associated with reporting a state j where no agent is lying from is a convex combination between the prior and the reputation of liars reporting j;

$$\mathcal{R}_j^C = r_j \mathbb{E}(\boldsymbol{t}) + (1 - r_j)\mathcal{L}(\hat{t}(\varphi)),$$

where $r_j \equiv \text{P(honest|report } j) \in (0,1]$. Since $\mathcal{L}(\hat{t}(\varphi)) < \mathbb{E}(\boldsymbol{t})$ by Lemma 3, this implies that $\mathcal{R}_j^C = \mathbb{E}(\boldsymbol{t})$ if $r_j = 1$, which is the case if $s(j|j',t) = 0$ for all $j' \neq j$ and all t, and $\mathcal{R}_j^C < \mathbb{E}(\boldsymbol{t})$ if $s(j|j',t) > 0$ for some $j' \neq j$. The only-if direction follows because the three cases lined out in the claim are mutually exclusive.

Claim 4: In any equilibrium, $\hat{t}_1(\varphi) > \hat{t}_2(\varphi) \geq \ldots \geq \hat{t}_K(\varphi) = \underline{t}$.

We are first going to show that $\hat{t}_j(\varphi)$ is weakly decreasing in j. Suppose that $\hat{t}_j(\varphi) = \underline{t}$. This implies that

$$y(K) + \mu\varphi - \underline{t} \leq y(j) + \mu\mathcal{R}_j^C.$$

Now suppose by contradiction that $\hat{t}_{j'}(\varphi) > \underline{t}$ for some $j' > j$. This implies that

$$y(K) + \mu\varphi - \underline{t} > y(j') + \mu\mathcal{R}_{j'}^C.$$

Combining both inequalities yields

$$y(j') + \mu\mathcal{R}_{j'}^C < y(j) + \mu\mathcal{R}_j^C.$$

By Claim 3, we know that $\mathcal{R}_{j'}^C > \mathbb{E}(\boldsymbol{t}) \geq \mathcal{R}_j^C$. Since $y(j') > y(j)$, we arrive at a contradiction. Therefore, if $\hat{t}_j(\varphi) = \underline{t}$, then $\hat{t}_{j'}(\varphi) = \underline{t}$ for all $j' > j$. Now consider the case where $\hat{t}_j(\varphi) > \underline{t}$. This implies that

$$y(K) + \mu\varphi = y(j) + \mu\mathcal{M}^+(\hat{t}_j(\varphi)) + \hat{t}_j(\varphi).$$

And suppose by contradiction that $\hat{t}_{j'}(\varphi) \geq \hat{t}_j(\varphi)$ for some $j' > j$. This implies that

$$y(K) + \mu\varphi = y(j') + \mu\mathcal{M}^+(\hat{t}_{j'}(\varphi)) + \hat{t}_{j'}(\varphi).$$

Combining both inequalities yields

$$y(j') + \mu\mathcal{M}^+(\hat{t}_{j'}(\varphi)) + \hat{t}_{j'}(\varphi) = y(j) + \mu\mathcal{M}^+(\hat{t}_j(\varphi)) + \hat{t}_j(\varphi).$$

However, since $y(j') > y(j)$, $\mathcal{M}^+(\hat{t}_{j'}(\varphi)) \geq \mathcal{M}^+(\hat{t}_j(\varphi))$, and $\hat{t}_{j'}(\varphi) \geq \hat{t}_j(\varphi)$, this yields a contradiction. Therefore, if $\hat{t}_j(\varphi) > \underline{t}$, then $\hat{t}_{j'}(\varphi) < \hat{t}_j(\varphi)$ for all $j' > j$. Combining both cases, we arrive at the conclusion that $\hat{t}_j(\varphi)$ is weakly decreasing in j.

Lemma 1 now suggests that (a) there is at least one type dishonestly reporting K and that (b) no agent who observes K lies. Point (a) implies that $\hat{t}_j(\varphi) > \underline{t}$ for some j, which, because $\hat{t}_j(\varphi)$ increases in j, implies that $\hat{t}_1(\varphi) > \underline{t}$. Point (b) implies that $\hat{t}_K(\varphi) = \underline{t}$ Combining everything, we arrive at the initial claim.

Claims 1-4 establish Part (i) of Proposition 1. I omit the proofs for parts $(ii) - (iii)$ in the proposition and instead focus on the existence and uniqueness of equilibrium.

Part (i) of the proposition uses k^* to denote the largest state that is not reported by any liar. Formally, define

$$k^* \equiv \max_{j \in \mathcal{K}} \{j | s(a = j|j',t) = 0 \text{ for all } j' \neq j \text{ and all } t\}.$$

We will further use j_L to denote the largest state for which the inequality $y(K) + \mu\varphi \geq y(j_L) + \mu\mathbb{E}(\boldsymbol{t})$ holds;

$$j_L \equiv \max_{j \in \mathcal{K}} \{j | y(K) + \mu\varphi \geq y(j) + \mu\mathbb{E}(\boldsymbol{t})\}.$$

The following claim establishes that k^* and j_L are equal.

Claim 5: $k^* = j_L$. The property that $s(a = k^*|j,t) = 0$ for all types $(j \neq k^*, t)$ has two implications that we are going to use in the proof. First, Claim 3 suggests that $\mathcal{R}_{k^*}^C \geq \mathbb{E}(\boldsymbol{t})$. Second, the property also suggests that

$$y(K) + \mu\varphi \geq y(k^*) + \mu\mathcal{R}_{k^*}^C.$$

There are two cases that we need to rule out to prove the claim. First, suppose by contradiction that $k^* > j_L$. This suggests that

$$y(K) + \mu\varphi < y(k^*) + \mu\mathbb{E}(\boldsymbol{t})$$

Combining the previous two inequalities suggests that

$$y(k^*) + \mu\mathcal{R}_{k^*}^C < y(k^*) + \mu\mathbb{E}(\boldsymbol{t}),$$

which is a contradiction to our initial observation that $\mathcal{R}^C_{k^*} \geq \mathbb{E}(t)$. Therefore, $k^* \leq j_L$. Second, suppose by contradiction that $k^* < j_L$. Since k^* is the *largest* j for which $s(a = j|j', t) = 0$ for types $(j' \neq j, t)$, $k^* < j_L$ suggests that there is a type such that $s(a = j_L|j, t) > 0$. We note two implications. First, by Claim 3, we know that $\mathcal{R}^C_{j_L} < \mathbb{E}(t)$. Second, since liars are indifferent between reporting any state reported by liars, we also know that

$$y(K) + \mu\varphi = y(j_L) + \mu\mathcal{R}^C_{j_L}.$$

Now recall from the definition of j_L that

$$y(K) + \mu\varphi \geq y(j_L) + \mu\mathbb{E}(t).$$

Combining the previous two inequalities suggests that $\mathcal{R}^C_{j_L} \geq \mathbb{E}(t)$, which contradicts our initial observation that $\mathcal{R}^C_{j_L} < \mathbb{E}(t)$. Therefore, $k^* \geq j_L$. Combining $k^* \geq j_L$ and $k^* \leq j_L$, we arrive at $k^* = j_L$.

Claim 6: For every $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, the fraction of agents who lie is a function $S(\varphi) = \sum_{j=1}^K p(j)F(\hat{t}_j(\varphi))$. The function S is continuous with $S'(\varphi) > 0$. The first part follows because agents only lie if their moral type is smaller than the threshold $\hat{t}_j(\varphi)$. Therefore, the fraction of agents who are liars is given by S. Continuity of S follows because $\hat{t}_j(\varphi)$ varies continuously between $\underline{t}$ and $\bar{t}$ on $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$ and because $F(t)$ is continuous on $[\underline{t}, \bar{t}]$. Moreover, $F'(t) > 0$ and $\hat{t}'_j(\varphi) \geq 0$, with strict inequality if $\hat{t}_j(\varphi) > \underline{t}$. Since $\hat{t}_1(\varphi) > \underline{t}$ for all $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, $S'(\varphi) > 0$.

Claim 7: For every $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, $D(\varphi) = \sum_{j=k^*+1}^K p(j)\frac{1-r_j(\varphi)}{r_j(\varphi)}$ is continuous with $D'(\varphi) < 0$. In equilibrium, $D(\varphi) = \mathrm{P}(lie)$. The fraction of liars that report a state larger than k^* is

$$\sum_{j=k^*+1}^K \mathrm{P}(\text{report } j) \times \mathrm{P}(\text{lie}|\text{report } j). \tag{A.7}$$

We defined $r_j = P(\text{honest}|\text{report } j)$. By Bayes' Rule,

$$r_j = \frac{\mathrm{P}(\text{report } j \wedge \text{honest})}{\mathrm{P}(\text{report } j)} \text{ for } j > k^*.$$

Observe that in equilibrium exactly $p(j)$ agents report each state $j > k^*$ truthfully. Thus, we can rearrange the above equation to

$$\mathrm{P}(\text{report } j) = \frac{p(j)}{r_j}.$$

Plugging into (A.7), we arrive at the following expression:

$$\sum_{j=k^*+1}^K \mathrm{P}(\text{report } j) \times \mathrm{P}(\text{lie}|\text{report } j) = \sum_{j=k^*+1}^K p(j)\frac{1-r_j}{r_j}. \tag{A.8}$$

We can derive an expression for r_j depending on φ by noting that,

$$\mathbb{E}(t|j) = r_j E(t) + (1 - r_j)\mathcal{L}(\hat{t}(\varphi)) \text{ for all } j > k^*$$

and use the indifference condition from Lemma 1 (i) to replace $\mathbb{E}(t|j) = \varphi + \frac{\Delta(K,j)}{\mu}$ to derive

$$r_j(\varphi) = \frac{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{t}(\varphi))}{\mathbb{E}(t) - \mathcal{L}(\hat{t}(\varphi))}. \tag{A.9}$$

Finally, we define

$$D(\varphi) \equiv \sum_{j=k^*+1}^K p(j)\frac{1-r_j(\varphi)}{r_j(\varphi)} = \sum_{j=k^*+1}^K p(j)\frac{\mathbb{E}(t) - (\varphi + \Delta(K,j)/\mu)}{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{t}(\varphi))}. \tag{A.10}$$

The function $D(\varphi)$ is continuous as $\varphi > \mathcal{L}(\hat{t}(\varphi))$ for $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$ and because the sum and quotient of continuous functions are continuous. $D(\varphi)$ is decreasing in φ: the numerators in the sum term of (A.10) decrease in φ while the denominators increase as long as

$$\frac{d\mathcal{L}}{d\varphi} < 1,$$

which was shown in Lemma 3.

Claim 8: There exists a unique $\varphi^* \in (\varphi^{\min}, \mathbb{E}(t))$ such that $D(\varphi^*) = S(\varphi^*)$. From the previous claims, it follows that $D(\varphi)$ and $S(\varphi)$ are both continuous functions with $D'(\varphi) < 0$ and $S'(\varphi) > 0$. The intermediate value theorem guarantees a unique φ^* such that

$D(\varphi^*) = S(\varphi^*)$. For existence of φ^*, observe that the parameter assumptions guarantee that $S(\varphi) \in (0,1)$ for all $\varphi \in (\varphi^{\min}, \mathbb{E}(\boldsymbol{t}))$. When $\varphi \to \varphi^{\min}$, $S(\varphi) = 0$ and $D(\varphi) > 0$. In the case where $\varphi \to \mathbb{E}(\boldsymbol{t})$, Claim 5 suggests that $k^* = K - 1$ and thus

$$\lim_{\varphi \to \mathbb{E}(\boldsymbol{t})} D(\varphi) = \lim_{\varphi \to \mathbb{E}(\boldsymbol{t})} p(K) \frac{\mathbb{E}(\boldsymbol{t}) - \varphi}{\varphi - \mathcal{L}(\hat{t}(\varphi))} = 0.$$

It follows that

$$\lim_{\varphi \to \varphi^{\min}} [D(\varphi) - S(\varphi)] > 0, \text{ and } \lim_{\varphi \to \mathbb{E}(\boldsymbol{t})} [D(\varphi) - S(\varphi)] < 0.$$

As the difference is continuous and strictly decreasing there exists a unique $\varphi^* \in (\varphi^{\min}, \mathbb{E}(\boldsymbol{t}))$ such that $D(\varphi^*) = S(\varphi^*)$. □

A.2. Proof of Proposition 2

The derivative

$$\Psi'(\hat{t}) = \frac{1}{1-p}\left[(1 - r(\hat{t}))(\mathcal{M}^{+'}(\hat{t}) - \mathcal{M}^{-'}(\hat{t})) - r'(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \mathcal{M}^{-}(\hat{t}))\right].$$

is positive if the term in brackets is positive. We use

$$r'(\hat{t}) = -\frac{(1-p)pf(\hat{t})}{(p + (1-p)F(\hat{t}))^2} = -\frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(1 - r(\hat{t}))$$

to rewrite the condition on the bracket term as

$$\mathcal{M}^{+'}(\hat{t}) - \mathcal{M}^{-'}(\hat{t}) + \frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \mathcal{M}^{-}(\hat{t})) > 0.$$

The second term becomes

$$\begin{aligned}
\frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \mathcal{M}^{-}(\hat{t})) &= \frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \hat{t} + \hat{t} - \mathcal{M}^{-}(\hat{t})) \\
&= \frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \hat{t}) + \frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\hat{t} - \mathcal{M}^{-}(\hat{t})) \\
&= \frac{f(\hat{t})}{F(\hat{t})} r(\hat{t})(\mathcal{M}^{+}(\hat{t}) - \hat{t}) + r(\hat{t})\mathcal{M}^{-'}(\hat{t}).
\end{aligned}$$

Plugging this into the condition on the bracket term and rearranging yields

$$\begin{aligned}
&(p + (1-p)F(\hat{t}))(\mathcal{M}^{+'}(\hat{t}) - \mathcal{M}^{-'}(\hat{t})) + p\left(\frac{f(\hat{t})}{F(\hat{t})}\mathcal{M}^{+}(\hat{t}) + \mathcal{M}^{-'}(\hat{t})\right) > 0 \\
\Rightarrow\; &F(\hat{t})\mathcal{M}^{+'}(\hat{t}) - F(\hat{t})\mathcal{M}^{-'}(\hat{t}) + p\left[(1 - F(\hat{t}))\mathcal{M}^{+'}(\hat{t}) + F(\hat{t})\mathcal{M}^{-'}(\hat{t}) + \frac{f(\hat{t})}{F(\hat{t})}(\mathcal{M}^{+}(\hat{t}) - \hat{t})\right] > 0 \\
\Rightarrow\; &p > \frac{F(\hat{t})\mathcal{M}^{-'}(\hat{t}) - F(\hat{t})\mathcal{M}^{+'}(\hat{t})}{F(\hat{t})\mathcal{M}^{-'}(\hat{t}) + (1 - F(\hat{t}))\mathcal{M}^{+'}(\hat{t}) + \frac{f(\hat{t})}{F(\hat{t})}(\mathcal{M}^{+}(\hat{t}) - \hat{t})}.
\end{aligned}$$

The r.h.s. is smaller than 1 for any $\hat{t}$. Therefore, there exists a p such that $\Psi'(\hat{t}) > 0$ for all $\hat{t}$.

A.3. Proof of Proposition 3a

Character-based image. Define a function

$$\tilde{p}(j, \delta) = \begin{cases} \frac{1}{K} + \frac{\delta}{K-1} & \text{if } j < K \\ \frac{1}{K} - \delta & \text{if } j = K. \end{cases}$$

This function returns the initial state distribution when evaluated at $\delta = 0$ ($\tilde{p}(j,0) = 1/K$) and the new state distribution when evaluated at $\delta = \tilde{\delta}$. The threshold values $\hat{t}_j(\varphi)$ are independent of δ. Therefore, for a given φ, the threshold state k^* is also independent of δ. It is weakly increasing in φ (see Claim 5 in the proof of Proposition 1). The expected lying cost of liars when the state distribution is $\tilde{p}(j, \delta)$ is equal to

$$\sum_{j=1}^{K} \frac{(1/K + \delta/(K-1))F(\hat{t}_j(\varphi))}{\sum_{l=1}^{K}(1/K + \delta/(K-1))F(\hat{t}_l(\varphi))} \mathcal{M}^{-}(\hat{t}_j(\varphi)) = \sum_{j=1}^{K} \frac{F(\hat{t}_j(\varphi))}{\sum_{l=1}^{K} F(\hat{t}_l(\varphi))} \mathcal{M}^{-}(\hat{t}_j(\varphi)) = \mathcal{L}(\hat{t}(\varphi)).$$

Therefore, the expected lying cost of liars is independent of δ. The equilibrium determining functions D and S can be written as

$$D(\varphi,\delta)=\sum_{j=k^*+1}^{K-1}\left(\frac{1}{K}+\frac{\delta}{K-1}\right)\frac{\mathbb{E}(t)-(\varphi+\Delta(K,j)/\mu)}{\varphi+\Delta(K,j)/\mu-\mathcal{L}(\hat{\imath}(\varphi))}+\left(\frac{1}{K}-\delta\right)\frac{\mathbb{E}(t)-\varphi}{\varphi-\mathcal{L}(\hat{\imath}(\varphi))},$$

$$S(\varphi,\delta)=\sum_{j=1}^{K}\left(\frac{1}{K}+\frac{\delta}{K-1}\right)F(\hat{t}_j(\varphi)).$$

It is easy to see that $S(\varphi,\tilde{\delta})>S(\varphi,0)$. To see that $D(\varphi,\tilde{\delta})<D(\varphi,0)$, take the derivative of $D(\varphi,\delta)$ with respect to δ;

$$\begin{aligned}\frac{\partial D}{\partial\delta}&=\frac{1}{K-1}\sum_{j=k^*+1}^{K-1}\frac{\mathbb{E}(t)-(\varphi+\Delta(K,j)/\mu)}{\varphi+\Delta(K,j)/\mu-\mathcal{L}(\hat{\imath}(\varphi))}-\frac{\mathbb{E}(t)-\varphi}{\varphi-\mathcal{L}(\hat{\imath}(\varphi))}\\&\leq\frac{K-1-k^*}{K-1}\frac{\mathbb{E}(t)-(\varphi+\Delta(K,K-1)/\mu)}{\varphi+\Delta(K,K-1)/\mu-\mathcal{L}(\hat{\imath}(\varphi))}-\frac{\mathbb{E}(t)-\varphi}{\varphi-\mathcal{L}(\hat{\imath}(\varphi))}<0\end{aligned}$$

The condition

$$D(\varphi,\delta)-S(\varphi,\delta)=0$$

implicitly defines a $\varphi^*(\delta)$ that denotes the equilibrium reputation associated with reporting K for a given δ. The following holds:

$$D(\varphi^*(\tilde{\delta}),\tilde{\delta})-S(\varphi^*(\tilde{\delta}),\tilde{\delta})=0=D(\varphi^*(0),0)-S(\varphi^*(0),0).$$

From the results above, we know that $D(\varphi^*(0),\tilde{\delta})-S(\varphi^*(0),\tilde{\delta})<0$. As $D-S$ decreases in φ, it follows that $\varphi^*(\tilde{\delta})<\varphi^*(0)$. As k^* is weakly increasing in φ, it follows that $k^*_{\delta=\tilde{\delta}}\geq k^*_{\delta=0}$. Also, while the probability of lying *conditional* on observing $j\leq k^*_{\delta=\tilde{\delta}}$ decreases, the overall lying probability might still increase, since, as δ increases, agents are more likely to observe $j\leq k^*_{\delta=\tilde{\delta}}$ in the first place.

Deed-based image. With deed-based image concerns, the equilibrium properties are that agents only lie if they observe $j\leq k^*<K$, where k^* is weakly increasing in the reputation of the highest state, φ. If they lie, they are indifferent between reporting any state larger than k^* (GK&S; K&S). With deed-based image concerns, the threshold that denotes the moral type who is indifferent between lying and telling the truth after observing j is equal to

$$\dot{\hat{t}}_j(\varphi)=\Delta(K,j)+\mu(\varphi-1),$$

where φ denotes the reputation of reporting K. In equilibrium, the reputation of K is equal to $r_K=\text{P(honest|report }K)$. It follows that

$$r_K(\varphi)=\varphi.$$

Since liars have to be indifferent, the reputation for reporting $j\in(k^*,K)$ can be derived from

$$\begin{aligned}y(K)+\mu r_K(\varphi)=&y(j)+\mu r_j\\\Rightarrow r_j(\varphi)=&\frac{\Delta(K,j)}{\mu}+r_K(\varphi).\end{aligned}$$

Similar arguments as those given in the proof of Proposition 1 imply that, in equilibrium, $\dot{D}(\varphi^D)=\dot{S}(\varphi^D)$, where

$$\dot{S}(\varphi)=\sum_{j=k^*}^{K}\frac{1}{K}F(\dot{\hat{t}}_j(\varphi))\text{ and }\dot{D}(\varphi)=\sum_{j=k^*+1}^{K}\frac{1}{K}\frac{1-r_j(\varphi)}{r_j(\varphi)}.$$

This function uniquely defines the equilibrium φ^D, and $S^{D'}(\varphi)>0$, $D^{D'}(\varphi)<0$.

When redistributing draw probabilities, $\dot{S}$ and $\dot{D}$ can be written as

$$\dot{S}(\varphi,\delta)=\sum_{j=1}^{K}\left(\frac{1}{K}+\frac{\delta}{K-1}\right)F(\dot{\hat{t}}_j(\varphi)),$$

$$\dot{D}(\varphi,\delta)=\sum_{j=k^*+1}^{K-1}\left(\frac{1}{K}+\frac{\delta}{K-1}\right)\frac{1-r_j(\varphi)}{r_j(\varphi)}+\left(\frac{1}{K}-\delta\right)\frac{1-r_K(\varphi)}{r_K(\varphi)}.$$

It is easy to see that $\dot{S}(\varphi,\tilde{\delta})>\dot{S}(\varphi,0)$ and $\dot{D}(\varphi,\tilde{\delta})<\dot{D}(\varphi,0)$ for $\delta>0$. The condition

$$\dot{D}(\varphi,\delta)-\dot{S}(\varphi,\delta)=0$$

determines a $\varphi^D(\delta)$ that denotes the equilibrium reputation associated with reporting K for a given δ. When comparing the equilibrium reputation $\varphi^D(\tilde{\delta})$ to the reputation $\varphi^D(0)$, we can use the fact that

$$\dot{D}(\varphi^D(0),0)-\dot{S}(\varphi^D(0),0)=0=\dot{D}(\varphi^D(\tilde{\delta}),\tilde{\delta})-\dot{S}(\varphi^D(\tilde{\delta}),\tilde{\delta}).$$

Combining the previous results with the fact that $\dot{D} - \dot{S}$ is decreasing in φ implies that $\varphi^D(\tilde{\delta}) < \varphi^D(0)$, which suggests that k^* weakly decreases. The effect on the likelihood of lying is ambiguous. While agents are less likely to lie conditional on observing $j \leq k^*_{\delta=\tilde{\delta}}$, agents are more likely to observe a $j \leq k^*_{\delta=\tilde{\delta}}$ in the first place.

A.4. Proof of Proposition 3b

Character-based image. As in the proof of the previous proposition, note that $\hat{t}(\varphi)$ is independent of the state distribution. For a given φ, the threshold state k^* is also independent of the state distribution and weakly increasing in φ. Define a function

$$\tilde{p}(j,\delta) = \begin{cases} \frac{1}{K} + \delta & \text{if } j = 1, \\ \frac{1}{K} - \delta & \text{if } j = k^*_{\delta=0}, \\ \frac{1}{K} & \text{otherwise,} \end{cases}$$

with $k^*_{\delta=0}$ denoting the threshold state when $\delta = 0$. This function returns the initial state distribution when evaluated at $\delta = 0$ ($\tilde{p}(j,0) = 1/K$) and the new state distribution when evaluated at $\delta = \tilde{\delta}$. It follows that

$$S(\varphi,\delta) = \sum_{j=1}^{K} \frac{1}{K} F(\hat{t}_j(\varphi)) + \delta(F(\hat{t}_1(\varphi)) - F(\hat{t}_{k^*_{\delta=0}}(\varphi))).$$

Denote the equilibrium reputation associated with reporting K when $\delta = 0$ by $\varphi^*(0)$. We find that $S(\varphi^*(0),\tilde{\delta}) > S(\varphi^*(0),0)$. Consider the expected moral type of liars, which, when evaluated at $\varphi^*(0)$, is equal to

$$\mathcal{L}(\hat{t}(\varphi^*(0)),\delta) = \frac{1}{S(\varphi^*(0),\delta)} \Bigg[\sum_{j=1}^{k^*_{\delta=0}} \frac{1}{K} F(\hat{t}_j(\varphi^*(0))) \mathcal{M}^-(\hat{t}_j(\varphi^*(0))) + \delta \left(F(\hat{t}_1(\varphi^*(0))) \mathcal{M}^+(\hat{t}_1(\varphi^*(0))) - F(\hat{t}_{k^*}(\varphi^*(0))) \mathcal{M}^-(\hat{t}_{k^*_{\delta=0}}(\varphi^*(0))) \right) \Bigg].$$

This function is increasing in δ, which suggests that $\mathcal{L}(\hat{t}(\varphi^*(0)),\tilde{\delta}) > \mathcal{L}(\hat{t}(\varphi^*(0)),0)$. Therefore, we have

$$D(\varphi^*(0),\tilde{\delta}) = \sum_{j=k^*_{\delta=0}+1}^{K} \frac{1}{K} \frac{\mathbb{E}(t) - (\varphi^*(0) + \Delta(K,j)/\mu)}{\varphi^*(0) + \Delta(K,j)/\mu - \mathcal{L}(\hat{t}(\varphi^*(0)),\tilde{\delta})} > D(\varphi^*(0),0).$$

The condition

$$D(\varphi,\delta) - S(\varphi,\delta) = 0$$

implicitly defines a $\varphi^*(\delta)$ that denotes the equilibrium reputation associated with reporting K for a given δ. To see that lying increases as we increase δ from 0 to $\tilde{\delta}$, consider a $\widetilde{\varphi}$ such that $S(\widetilde{\varphi},\tilde{\delta}) = S(\varphi^*(0),0)$. Now, $S(\varphi^*(0),\tilde{\delta}) > S(\varphi^*(0),0)$ and $\frac{\partial S}{\partial \varphi} > 0$ suggest that $\widetilde{\varphi} < \varphi^*(0)$. From $\frac{\partial D}{\partial \varphi} < 0$ it now follows that $D(\widetilde{\varphi},\tilde{\delta}) > D(\varphi^*(0),\tilde{\delta}) > D(\varphi^*(0),0)$. This implies that

$$D(\widetilde{\varphi},\tilde{\delta}) - S(\widetilde{\varphi},\tilde{\delta}) > D(\varphi^*(0),0) - S(\varphi^*(0),0) = D(\varphi^*(\tilde{\delta}),\tilde{\delta}) - S(\varphi^*(\tilde{\delta}),\tilde{\delta}) = 0,$$

and since $D - S$ is decreasing in φ we conclude that $\varphi^*(\tilde{\delta}) > \widetilde{\varphi}$. Therefore, $S(\varphi^*(\tilde{\delta}),\tilde{\delta}) > S(\widetilde{\varphi},\tilde{\delta}) = S(\varphi^*(0),0)$; the likelihood of lying increases. Whether k^* increases or decreases depends on whether $\varphi^*(\tilde{\delta})$ is smaller or larger than $\varphi^*(0)$, which is generally ambiguous.

Deed-based image. As in the character-based model, we can write

$$\dot{S}(\varphi,\delta) = \sum_{j=1}^{K} \frac{1}{K} F(\hat{t}_j(\varphi)) + \delta(F(\hat{t}_1(\varphi)) - F(\hat{t}_{k^{*D}_{\delta=0}}(\varphi))).$$

Denote the equilibrium reputation associated with reporting K when $\delta = 0$ by $\varphi^{*D}(0)$. We find that $\dot{S}(\varphi^{*D}(0),\tilde{\delta}) > \dot{S}(\varphi^{*D}(0),0)$. Now consider the function $\dot{D}$;

$$\dot{D}(\varphi^*(0),\tilde{\delta}) = \sum_{j=k^{*D}_{\delta=0}+1}^{K} \frac{1}{K} \frac{1 - r_j(\varphi^*(0))}{r_j(\varphi^*(0))} = \dot{D}(\varphi^*(0),0).$$

Consider a $\widetilde{\varphi}$ such that $\dot{S}(\widetilde{\varphi},\tilde{\delta}) = \dot{S}(\varphi^{*D}(0),0)$. Now, since $\frac{\partial S}{\partial \varphi} > 0$, the above results suggest that $\widetilde{\varphi} < \varphi^{*D}(0)$. Furthermore, since $\frac{\partial \dot{D}}{\partial \varphi} < 0$, $\dot{D}(\widetilde{\varphi},\tilde{\delta}) > \dot{D}(\varphi^{*D}(0),\tilde{\delta}) = \dot{D}(\varphi^{*D}(0),0)$. This implies that

$$\dot{D}(\widetilde{\varphi},\tilde{\delta}) - \dot{S}(\widetilde{\varphi},\tilde{\delta}) > \dot{D}(\varphi^{*D}(0),0) - \dot{S}(\varphi^{*D}(0),0) = \dot{D}(\varphi^{*D}(\tilde{\delta}),\tilde{\delta}) - \dot{S}(\varphi^{*D}(\tilde{\delta}),\tilde{\delta}) = 0,$$

and since $\dot{D}-\dot{S}$ is decreasing in φ we conclude that $\varphi^{*D}(\tilde{\delta})>\widetilde{\varphi}$. Therefore, $\dot{S}(\varphi^{*D}(\tilde{\delta}),\tilde{\delta})>\dot{S}(\widetilde{\varphi},\tilde{\delta})=\dot{S}(\varphi^{*D}(0),0)$; the likelihood of lying is larger under $\delta=\tilde{\delta}$ than under $\delta=0$. To see that $\varphi^{*D}(\tilde{\delta})<\varphi^{*D}(0)$, note that when evaluated at $\varphi^{*D}(0)$, $\dot{D}(\varphi^{*D}(0),\tilde{\delta})-\dot{S}(\varphi^{*D}(0),\tilde{\delta})<0$, so that φ needs to decrease in order to bring the difference to zero. This suggests that $k^{*D}_{\delta=\tilde{\delta}}\leq k^{*D}_{\delta=0}$.

A.5. Proofs of Propositions 4b–5b and Corollary 1

A.5.1. Preliminaries

If t is distributed according to $F_U(t,\underline{t},\bar{t})$, then

$$F_U(t,\underline{t},\bar{t})=\begin{cases}0 & \text{if } t\leq\underline{t}\\ \frac{t-\underline{t}}{\bar{t}-\underline{t}} & \text{if } t\in(\underline{t},\bar{t})\\ 1 & \text{if } t\geq\bar{t},\end{cases}$$

$$\mathcal{M}^{-}(t)=\begin{cases}\frac{t+\underline{t}}{2} & \text{if } t\in(\underline{t},\bar{t})\\ \frac{t+\bar{t}}{2} & \text{if } t\geq\bar{t},\end{cases}$$

and

$$\mathcal{M}^{+}(t)=\begin{cases}\frac{\underline{t}+\bar{t}}{2} & \text{if } t\leq\underline{t},\\ \frac{\bar{t}+t}{2} & \text{if } t\in(\underline{t},\bar{t}).\end{cases}$$

A.5.2. Proof of Proposition 4b

Denote by $F_X(t)=F_U(t,0,\bar{t})$ and by $F_Y(t)=F_U(t,c,\bar{t}+c)$. Using properties of the uniform distribution,

$$\mathcal{M}^{+}_Y(t+c)-M^{+}_X(t)=\mathcal{M}^{-}_Y(t+c)-M^{-}_X(t)=\mathbb{E}_Y(\boldsymbol{t})-\mathbb{E}_X(\boldsymbol{t})=c.$$

Use φ_X to denote the equilibrium reputation associated with reporting K under F_X and consider a $\widetilde{\varphi}$ such that $\hat{t}_{jY}(\widetilde{\varphi})=\hat{t}_{jX}(\varphi_X)+c$. Note that at such a threshold, $F_Y(\hat{t}_{jY}(\widetilde{\varphi}))=F_X(\hat{t}_{jX}(\varphi_X))$. It follows that

$$\begin{aligned}\hat{t}_{jX}(\varphi_X)+c&=\Delta(K,j)+\mu(\widetilde{\varphi}-\mathcal{M}^{+}_Y(\hat{t}_{jY}(\widetilde{\varphi})))\\ &=\Delta(K,j)+\mu(\varphi_X-\mathcal{M}^{+}_X(\hat{t}_{jX}(\varphi_X)))+\mu(\widetilde{\varphi}-\varphi_X)+\mu(\mathcal{M}^{+}_X(\hat{t}_{jX}(\varphi_X))-\mathcal{M}^{+}_Y(\hat{t}_{jY}(\widetilde{\varphi})))\\ &=\hat{t}_{jX}(\varphi_X)+\mu(\widetilde{\varphi}-\varphi_X)-\mu c.\\ \Rightarrow\widetilde{\varphi}&=\varphi_X+\frac{1+\mu}{\mu}c.\end{aligned}$$

Since $\widetilde{\varphi}$ is independent of j, it follows that $F_Y(\hat{t}_{jY}(\widetilde{\varphi}))=F_X(\hat{t}_{jX}(\varphi_X))$ for all j. Therefore, $S_Y(\widetilde{\varphi})=S_X(\varphi_X)$. Now consider

$$\begin{aligned}\mathcal{L}_Y(\hat{t})&=\sum_{j=1}^{K}\frac{p(j)F_Y(\hat{t})}{\sum_{l=1}^{K}p(j)F_Y(\hat{t})}\mathcal{M}^{-}_Y(\hat{t})\\ &=\sum_{j=1}^{K}\frac{p(j)F_X(\hat{t}-c)}{\sum_{l=1}^{K}p(j)F_X(\hat{t}-c)}(\mathcal{M}^{-}_X(\hat{t}-c)+c)=\mathcal{L}_X(\hat{t}-c)+c.\end{aligned}$$

This implies that $\mathcal{L}_Y(\hat{t}_Y(\widetilde{\varphi}))=\mathcal{L}_X(\hat{t}_X(\varphi_X))+c$. We combine these insights to show that $D_Y(\widetilde{\varphi})<D_X(\varphi_X)$. A sufficient condition is that

$$\frac{\mathbb{E}_Y(\boldsymbol{t})-(\widetilde{\varphi}+\Delta(K,j)/\mu)}{\widetilde{\varphi}+\Delta(K,j)/\mu-\mathcal{L}_Y(\hat{t}_Y(\widetilde{\varphi}))}<\frac{\mathbb{E}_X(\boldsymbol{t})-(\varphi_X+\Delta(K,j)/\mu)}{\varphi_X+\Delta(K,j)/\mu-\mathcal{L}_X(\hat{t}(\varphi_X))}$$

for all $j>k^{*}_X$. Plugging previous insights into the inequality and using $\widetilde{\varphi}=\varphi_X+\frac{1+\mu}{\mu}c$, we get

$$\frac{\mathbb{E}_X(\boldsymbol{t})+c-(\varphi_X+\frac{1+\mu}{\mu}c+\Delta(K,j)/\mu)}{\varphi_X+\frac{1+\mu}{\mu}c+\Delta(K,j)/\mu-\mathcal{L}_X(\hat{t}(\varphi_X))-c}<\frac{\mathbb{E}_X(\boldsymbol{t})-(\varphi_X+\Delta(K,j)/\mu)}{\varphi_X+\Delta(K,j)/\mu-\mathcal{L}_X(\hat{t}(\varphi_X))}.$$

It immediately follows that the numerator on the l.h.s. is smaller than the numerator on the r.h.s. and that the denominator on the l.h.s. is larger than the denominator on the r.h.s. We conclude that the inequality holds for all $j>k^{*}_X(\varphi_X)$. Therefore, $D_Y(\widetilde{\varphi})<D_X(\varphi_X)$. Combining the previous results, we find that

$$D_Y(\widetilde{\varphi}) - S_Y(\widetilde{\varphi}) < D_X(\varphi_X) - S_X(\varphi_X).$$

The conditions

$$D_X(\varphi_X) - S_X(\varphi_X) = 0 \text{ and } D_Y(\varphi_Y) - S_Y(\varphi_Y) = 0$$

define the equilibrium reputations φ_X and φ_Y associated with reporting K under F_X and F_Y, respectively. Since $D_X(\varphi_X) - S_X(\varphi_X) = 0$, it follows that $D_Y(\widetilde{\varphi}) - S_Y(\widetilde{\varphi}) < 0$. Since $D - S$ is decreasing in φ, it follows that $\varphi_Y < \widetilde{\varphi}$. As $\frac{\partial S}{\partial \varphi} > 0$, this suggests that $S_Y(\varphi_Y) < S_Y(\widetilde{\varphi}) = S_X(\varphi_X)$; lying is lower under F_Y than F_X.

Turning to Part (ii), we are going to show that $\hat{t}_{Yk_X^*}(\varphi_Y) > \hat{t}_{Xk_X^*}(\varphi_X)$, which suggests that $k_Y^* \geq k_X^*$ as $k_Y^* < k_X^*$ would imply that $\hat{t}_{Yk_X^*}(\varphi_Y) = \underline{t} \leq t_{Xk_X^*}(\varphi_X)$. Consider a $\widetilde{\varphi}'$ such that $\hat{t}_{Yk_X^*}(\widetilde{\varphi}') = \hat{t}_{Xk_X^*}(\varphi_X)$, implying that

$$\begin{aligned}
&\hat{t}_{Yk_X^*}(\widetilde{\varphi}') = \Delta(K, k_X^*) + \mu(\widetilde{\varphi}' - \mathcal{M}_Y^+(\hat{t}_{Yk_X^*}(\widetilde{\varphi}'))) = \Delta(K, k_X^*) + \mu(\varphi_X - \mathcal{M}_X^+(\hat{t}_{Xk_X^*}(\varphi_X))) = \hat{t}_{Xk_X^*}(\varphi_X),\\
\Rightarrow &\widetilde{\varphi}' = \varphi_X + \mathcal{M}_Y^+(\hat{t}_{Xk_X^*}(\varphi_X)) - \mathcal{M}_X^+(\hat{t}_{Xk_X^*}(\varphi_X)) = \varphi_X + \frac{c}{2}.
\end{aligned}$$

Since $\widetilde{\varphi}'$ is independent of j, it follows that $\hat{t}_Y(\widetilde{\varphi}') = \hat{t}_X(\varphi_X)$. Moreover, since $\widetilde{\varphi}' = \varphi_X + c/2 < \widetilde{\varphi} = \varphi + (1+\mu)/\mu c$, it follows that $S_Y(\widetilde{\varphi}') < S_Y(\widetilde{\varphi}) = S_X(\varphi_X)$. We are now going to argue that $D_Y(\widetilde{\varphi}') > D_X(\varphi_X)$. A sufficient condition is that, for all $j > k_X^*$,

$$\frac{\mathbb{E}_X(t) + c/2 - \varphi_X - \Delta(K,j)/\mu}{\varphi_X + \Delta(K,j)/\mu + c/2 - \mathcal{L}_Y(\hat{t}_X(\varphi_X))} > \frac{\mathbb{E}_X(t) - \varphi_X - \Delta(K,j)/\mu}{\varphi_X + \Delta(K,j)/\mu - \mathcal{L}_X(\hat{t}_X(\varphi_X))}. \tag{A.11}$$

The numerator on the l.h.s. is strictly larger than the numerator on the r.h.s. Therefore, this inequality holds if the denominator on the l.h.s. is smaller;

$$\begin{aligned}
&\varphi_X + \Delta(K,j)/\mu + c/2 - \mathcal{L}_Y(\hat{t}_X(\varphi_X)) \leq \varphi_X + \Delta(K,j)/\mu - \mathcal{L}_X(\hat{t}_X(\varphi_X)),\\
\Rightarrow &c/2 \leq \mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)).
\end{aligned}$$

Denote the probability of observing j conditional on lying by

$$q_j(\hat{t}) = \frac{p(j)F(\hat{t}_j)}{\sum_{l=1}^K p(l)F(\hat{t}_l)},$$

which, under F_X and F_Y, becomes

$$\begin{aligned}
q_{Xj}(\hat{t}) &= \frac{p(j)\hat{t}_j}{\sum_{l=1}^{k^*} p(l)\hat{t}_l},\\
q_{Yj}(\hat{t}) &= \frac{p(j)(\hat{t}_j - c)}{\sum_{l=1}^{k^*} p(l)(\hat{t}_l - c)}.
\end{aligned}$$

Taking the derivative of q_{Yj} with respect to c, we find that

$$\frac{\partial q_{Yj}(\hat{t})}{\partial c} = p(j)\frac{(\hat{t}_j - c) - \sum_{l=1}^{k^*} p(l)(\hat{t}_l - c)}{(\sum_{l=1}^{k^*} p(l)(\hat{t}_l - c))^2} = \frac{1}{\sum_{l=1}^{k^*} p(l)(\hat{t}_l - c)}\left[q_{Yj}(\hat{t}) - p(j)\right].$$

This suggests that $q_{Yj}(\hat{t}) > q_{Xj}(\hat{t})$ if and only if $q_{Yj}(\hat{t}) > p(j)$. Therefore, as $\hat{t}_{Xj}(\varphi_X)$ is decreasing in j, there is a $z \in [1, k_X^*)$ such that $q_{Yj}(\hat{t}) > q_{Xj}(\hat{t})$ if and only if $j \leq z$. We can now rewrite the difference in $\mathcal{L}$ as follows:

$$\begin{aligned}
\mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)) = &\sum_{j=1}^K q_{Yj}(\hat{t})(\mathcal{M}_Y^-(\hat{t}_{Xj}(\varphi_X)) - \mathcal{M}_X^-(\hat{t}_{Xj}(\varphi_X)))\\
&+ \sum_{j=1}^K (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))\mathcal{M}_X^-(\hat{t}_{Xj}(\varphi_X)).
\end{aligned}$$

The first sum term simplifies to $c/2$. The second sum term can be written as

$$\begin{aligned}
&\sum_{j=1}^z (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))\mathcal{M}_X^-(\hat{t}_{Xj}(\varphi_X)) + \sum_{j=z+1}^{k^*} (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))\mathcal{M}_X^-(\hat{t}_{Xj}(\varphi_X))\\
&> \sum_{j=1}^z (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))\mathcal{M}_X^-(\hat{t}_{Xz}(\varphi_X)) + \sum_{j=z+1}^{k^*} (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))\mathcal{M}_X^-(\hat{t}_{Xz+1}(\varphi_X))\\
&= \sum_{j=1}^z (q_{Yj}(\hat{t}_X(\varphi_X)) - q_{Xj}(\hat{t}_X(\varphi_X)))(\mathcal{M}_X^-(\hat{t}_{Xz}(\varphi_X)) - \mathcal{M}_X^-(\hat{t}_{Xz+1}(\varphi_X))) \geq 0.
\end{aligned}$$

Therefore, $\mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)) \geq c/2$, which suggests that the sufficient condition in (A.11) holds. We conclude that $D_Y(\widetilde{\varphi}') > D_X(\varphi_X) = S_X(\varphi_X) > S_Y(\widetilde{\varphi}')$. Therefore, $D_Y(\widetilde{\varphi}') - S_Y(\widetilde{\varphi}') > 0$ and, since $D - S$ is decreasing in φ, it follows that $\varphi_Y > \widetilde{\varphi}'$. This suggests that $\hat{t}_{Yk_X^*}(\varphi_Y) > \hat{t}_{Xk_X^*}(\varphi_X)$, which implies that $k_Y^* \geq k_X^*$.

A.5.3. Proof of Proposition 4c

Denote by $F_X(t) = F_U(t, 0, \bar{t})$ and by $F_Y(t) = F_U(t, 0, \bar{t} + c)$. By the uniform distribution,

$$\mathcal{M}_Y^+(t) - \mathcal{M}_X^+(t) = \frac{c}{2}$$

is independent of t.

The threshold type that lies after observing j is given by

$$\hat{t}_j(\varphi) = \Delta(K, j) + \mu(\varphi - \mathcal{M}^+(\hat{t}_j(\varphi))).$$

Denote the equilibrium reputation associated with reporting K under F_X by φ_X and consider

$$\hat{t}_{Yj}(\varphi) - \hat{t}_{Xj}(\varphi_X) = \mu(\varphi - \varphi_X + \mathcal{M}_X^+(\hat{t}_{Xj}(\varphi_X)) - \mathcal{M}_Y^+(\hat{t}_{Yj}(\varphi))).$$

Choose $\widetilde{\varphi}$ such that $\hat{t}_{Yj}(\widetilde{\varphi}) = \hat{t}_{Xj}(\varphi_X)$. This suggests that

$$\widetilde{\varphi} = \varphi_X + \frac{c}{2} > \varphi_X.$$

The $\widetilde{\varphi}$ is independent of j. Therefore, $\hat{t}_Y(\widetilde{\varphi}) = \hat{t}_X(\varphi_X)$. Note further that, because $F_Y(t) < F_X(t)$ for all t, $S_Y(\widetilde{\varphi}) < S_X(\varphi_X)$. Now consider

$$D_Y(\widetilde{\varphi}) = \sum_{k_X^*+1}^{K} p(j) \frac{\mathbb{E}_Y(\boldsymbol{t}) - (\varphi_X + c/2 + \Delta(K, j)/\mu)}{\varphi_X + c/2 + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))}.$$

We can replace $\mathbb{E}_Y(\boldsymbol{t}) = c/2 + \mathbb{E}_X(\boldsymbol{t})$. Therefore,

$$D_Y(\widetilde{\varphi}) = \sum_{k_X^*+1}^{K} p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + c/2 + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))}.$$

We are going to argue that $D_Y(\widetilde{\varphi}) < S_Y(\widetilde{\varphi})$. To see this, note that equilibrium requires that $D_X(\varphi_X) = S_X(\varphi_X)$, which we can rewrite as

$$\bar{t} \sum_{k_X^*+1}^{K} p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))} = \sum_{j=1}^{k_X^*} p(j) \hat{t}_Y(\widetilde{\varphi}).$$

We can plug this into the inequality $D_Y(\widetilde{\varphi}) < S_Y(\widetilde{\varphi})$ and rearrange it to

$$(\bar{t} + c) \sum_{k_X^*+1}^{K} p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + c/2 + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))} < \bar{t} \sum_{k_X^*+1}^{K} p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))}.$$

A sufficient condition for this inequality to hold is that, for all $j > k_X^*$,

$$\begin{aligned}
&(\bar{t} + c) p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + c/2 + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))} < \bar{t} p(j) \frac{\mathbb{E}_X(\boldsymbol{t}) - (\varphi_X + \Delta(K, j)/\mu)}{\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))}, \\
\Rightarrow &\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X)) < \frac{\bar{t}}{\bar{t} + c} (\varphi_X + c/2 + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))) \\
\Rightarrow &(1 - \frac{\bar{t}}{\bar{t} + c})(\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))) < \frac{\bar{t}}{\bar{t} + c} \frac{c}{2} \\
\Rightarrow &\frac{c}{\bar{t} + c} (\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X))) < \frac{\bar{t}}{\bar{t} + c} \frac{c}{2} \\
\Rightarrow &\varphi_X + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}_X(\varphi_X)) < \frac{\bar{t}}{2} = \mathbb{E}_X(\boldsymbol{t}).
\end{aligned}$$

This holds, since in l.h.s. is smaller than $\varphi_X + \Delta(K, j)/\mu$, which is equal to the equilibrium reputation associated with reporting $j > k_X^*$ under F_X, which is smaller than $\mathbb{E}_X(\boldsymbol{t})$. Therefore, $D_Y(\widetilde{\varphi}) < S_Y(\widetilde{\varphi})$. As equilibrium requires that $D_Y(\varphi_Y) = S_Y(\varphi_Y)$ and since D is decreasing and S is increasing in φ, it follows that $\varphi_Y < \widetilde{\varphi}$, which implies $S_Y(\varphi_Y) < S_Y(\widetilde{\varphi}) < S_X(\varphi_X)$. We conclude that the likelihood of lying is lower under $F_Y(t)$ than under $F_X(t)$.

To show that k^* weakly decreases, note that $\varphi_Y < \widetilde{\varphi}$ implies that $\hat{t}_{Yk_X^*+1}(\varphi_Y) = \hat{t}_{Yk_X^*+1}(\widetilde{\varphi}) = \hat{t}_{Xk_X^*+1}(\varphi_X) = \underline{t}$. As an increase in k^* would imply that $\hat{t}_{Yk_X^*+1}(\varphi_Y) > 0$, we conclude that $k_Y^* \leq k_X^*$.

Proof of Corollary 1. Note that $k_Y^* \geq k_X^*$ only if $\hat{t}_Y(\varphi_Y) \geq \hat{t}_X(\varphi_X)$. Therefore, all $\hat{t}_j$ weakly increase if k^* weakly increases. If all $\hat{t}_j$ increase, this implies that an agent with lying cost t becomes more likely to lie after observing any state. Therefore, a type (j,t) becomes more likely to lie if k^* weakly increases. All claims of the corollary follow from this observation.

A.5.4. Proof of Proposition 5a

Denote by $F_X(t) = F_U(t,0,\bar{t})$ and by $F_Y(t) = F_U(t,c,\bar{t}-c)$. Denote the equilibrium reputation associated with reporting K under F_X by φ_X. The threshold $\hat{t}_{1X}(\varphi_X)$, is implicitly defined in

$$\hat{t}_{1X}(\varphi_X) = \Delta(K,1) + \mu(\varphi_X - \mathcal{M}_X^+(\hat{t}_{1X}(\varphi_X))).$$

Consider a $\widetilde{\varphi}$ such that $\hat{t}_{1Y}(\widetilde{\varphi}) = \hat{t}_{1X}(\varphi_X)$. This implies that

$$\begin{aligned}\widetilde{\varphi} =& \varphi_X + \mathcal{M}_Y^+(\hat{t}_{1X}(\varphi_X)) - \mathcal{M}_X^+(\hat{t}_{1X}(\varphi_X))\\ =& \varphi_X + \frac{c}{2},\end{aligned}$$

where we used properties of the uniform distribution. Observe that $\widetilde{\varphi} > \varphi_X$ and that $\widetilde{\varphi}$ is independent of j, which suggests that $\hat{t}_Y(\widetilde{\varphi}) = \hat{t}_X(\varphi_X)$. We are going to show that $D_Y(\widetilde{\varphi}) > S_Y(\widetilde{\varphi})$. First, note that we can use the equilibrium condition $D_X(\varphi_X) = S_X(\varphi_X)$ to show that[37]

$$\sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \varphi_X - \Delta(K,j)/\mu}{\varphi_X + \Delta(K,j)/\mu - \mathcal{L}_X(\hat{t}_X(\varphi_X))} \times \bar{t} = \sum_{j=1}^{K} p(j)\hat{t}_{Xj}(\varphi_X).$$

The inequality $D_Y(\widetilde{\varphi}) > S_Y(\widetilde{\varphi})$ implies that

$$\sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \widetilde{\varphi} - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}_Y(\hat{t}_X(\varphi_X))} \times (\bar{t} - 2c) + c > \sum_{j=1}^{K} p(j)\hat{t}_{Yj}(\widetilde{\varphi}).$$

Combining both equations by using $\hat{t}_{Xj}(\varphi_X) = \hat{t}_{Yj}(\widetilde{\varphi})$, we find that

$$\sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \widetilde{\varphi} - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}_Y(\hat{t}_X(\varphi_X))} \times (\bar{t} - 2c) + c > \sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \varphi_X - \Delta(K,j)/\mu}{\varphi_X + \Delta(K,j)/\mu - \mathcal{L}_X(\hat{t}_X(\varphi_X))}\bar{t}.$$

Observe that we can rewrite the l.h.s. as

$$\sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \widetilde{\varphi} - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}_Y(\hat{t}_X(\varphi_X))} \times (\bar{t} - c) + c\left(1 - \underbrace{\sum_{j=k_X^*+1}^{K} p(j)\frac{\mathbb{E}(t) - \widetilde{\varphi} - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}_Y(\hat{t}_X(\varphi_X))}}_{<1}\right).$$

Therefore, a sufficient condition for this inequality to hold is that

$$\frac{\mathbb{E}(t) - \widetilde{\varphi} - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}_Y(\hat{t}_X(\varphi_X))} \times (\bar{t} - c) > \frac{\mathbb{E}(t) - \varphi_X - \Delta(K,j)/\mu}{\varphi_X + \Delta(K,j)/\mu - \mathcal{L}_X(\hat{t}_X(\varphi_X))} \times \bar{t}$$

for all $j > k_X^*$. We will from now on use $\tilde{\mathcal{R}}_j^C = \widetilde{\varphi} + \Delta(K,j)/\mu$ and $\mathcal{R}_j^C = \varphi_X + \Delta(K,j)/\mu$ to save notation. The sufficient condition can be rearranged to

$$\begin{aligned}&\left[\mathbb{E}(t) - \tilde{\mathcal{R}}_j^C\right]\left[\mathcal{R}_j^C - \mathcal{L}_X(\hat{t}_X(\varphi_X))\right](\bar{t} - c) > \left[\mathbb{E}(t) - \mathcal{R}_j^C\right]\left[\tilde{\mathcal{R}}_j^C - \mathcal{L}_Y(\hat{t}_X(\varphi_X))\right]\bar{t},\\ \Rightarrow &\Big[(\mathbb{E}(t) - \mathcal{L}_X(\hat{t}_X(\varphi_X)))(\mathcal{R}_j^C - \tilde{\mathcal{R}}_j^C) + (\mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)))(\mathbb{E}(t) - \mathcal{R}_j^C)\Big]\bar{t}\\ &> c\left[(\mathbb{E}(t) - \tilde{\mathcal{R}}_j^C)(\mathcal{R}_j^C - \mathcal{L}_X(\hat{t}_X(\varphi_X)))\right].\end{aligned}$$

We can rewrite the l.h.s. as

$$\begin{aligned}&\Big[(\mathbb{E}(t) - \tilde{\mathcal{R}}_j^C)(\mathcal{R}_j^C - \mathcal{L}_X(\hat{t}_X(\varphi_X))) + \mathcal{L}_X(\hat{t}_X(\varphi_X))\mathbb{E}(t) + \mathcal{R}_j^C\tilde{\mathcal{R}}_j^C\\ &+ (\mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)))(\mathbb{E}(t) - \mathcal{R}_j^C)\Big]\bar{t}.\end{aligned}$$

[37] Because F_Y is a mean-preserving contraction of F_X, $\mathbb{E}_X(t) = \mathbb{E}_Y(t)$. We can therefore use $\mathbb{E}(t)$ to denote the prior expectation under both distributions.

The same arguments as used in the proof of Proposition 4c imply that $\mathcal{L}_Y(\hat{t}_X(\varphi_X)) - \mathcal{L}_X(\hat{t}_X(\varphi_X)) > c/2$. Therefore, the sufficient condition holds; $D_Y(\widetilde{\varphi}) > S_Y(\widetilde{\varphi})$. Equilibrium requires that $D_Y(\varphi_Y) = S_Y(\varphi_Y)$. Since $D - S$ is decreasing in φ, this indicates that $\varphi_Y > \widetilde{\varphi}$ and, therefore, that $\hat{t}_{Yj}(\varphi_Y) \geq \hat{t}_{Yj}(\widetilde{\varphi}) = \hat{t}_{Xj}(\varphi_X)$ for all j. Parts (i) and Part (ii) follow immediately: First, $\hat{t}_{Yk^*_X+1}(\varphi_Y) \geq \hat{t}_{Xk^*_X+1}(\varphi_X)$ suggests that $k^*_Y \geq k^*_X$. Second, $\hat{t}_{Yj}(\varphi_Y) \geq \hat{t}_{Xj}(\varphi_X)$ for all j implies that any type (j,t) becomes more likely to lie.

A.5.5. Proof of Proposition 5b

Denote by $F_Y(t)$ the mean-preserving contraction of $F_X(t)$. First, note that $\dot{D}(\varphi) = \sum_{j=k^*+1}^{K} p(j)(1 - r_j(\varphi))/r_j(\varphi)$ is independent of F and so is $\dot{\hat{t}}_j(\varphi)$. Therefore, we can focus on $\dot{S}(\varphi)$. Take the difference

$$\dot{S}_Y(\varphi^D_X) - \dot{S}_X(\varphi^D_X) = \sum_{j=1}^{k^*} p(j)(F_Y(\dot{\hat{t}}_j(\varphi^D_X)) - F_X(\dot{\hat{t}}_j(\varphi^D_X))).$$

Since $F_Y(t)$ is a mean-preserving contraction of $F_X(t)$, this implies that there is one $\tilde{t}$ such that

$$F_X(t) > F_Y(t) \text{ if } t < \tilde{t},\ F_X(\tilde{t}) = F_Y(\tilde{t}), \text{ and } F_X(t) < F_Y(t) if\ t > \tilde{t}.$$

Therefore, $\dot{S}_Y(\varphi^D_X) - \dot{S}_X(\varphi^D_X) > 0$ if $\dot{\hat{t}}_j(\varphi^D_X) \geq \tilde{t}$ for all j and $\dot{S}_Y(\varphi^D_X) - \dot{S}_X(\varphi^D_X) < 0$ if $\dot{\hat{t}}_j(\varphi^D_X) \leq \tilde{t}$ for all j. More generally, the difference is increasing in φ. This suggests hat $\dot{D}(\varphi^D_X) - \dot{S}(\varphi^D_X) < 0$ if $\dot{\hat{t}}(\varphi^D_X)$ is sufficiently large. Then, since $\dot{D}(\varphi) - \dot{S}(\varphi)$ is decreasing in φ, this implies that $\varphi^D_Y < \varphi^D_X$. In this case $\dot{\hat{t}}(\varphi^D_Y) < \dot{\hat{t}}(\varphi^D_X)$, which implies that $k^*_Y \leq k^*_X$ and that a type (j,t) becomes less likely to lie. In other cases, if $\dot{\hat{t}}(\varphi^D_X)$ is not sufficiently large $\dot{D}(\varphi^D_X) - \dot{S}(\varphi^D_X) \geq 0$ and $\varphi^D_Y \geq \varphi^D_X$. This implies that $\dot{\hat{t}}(\varphi^D_Y) \geq \dot{\hat{t}}(\varphi^D_X)$, which implies that $k^*_Y \geq k^*_X$ and that a type (j,t) becomes more likely to lie.

A.6. Proof of Proposition 6a

With coarse disclosure, the audience observes the report and possibly a disclosure decision. If a liar is disclosed, the audience attaches reputation $\mathcal{L}(\hat{t}(\varphi))$. If not disclosed, the audience does not know for sure whether the report is a lie. The expected reputation of a liar reporting K is

$$\mathbb{E}(\boldsymbol{\mathcal{R}}^{\boldsymbol{C}}_{\boldsymbol{K}} | \text{observe } j < K) = \varphi = (1 - \pi)(r_K \mathbb{E}(\boldsymbol{t}) + (1 - r_K)\mathcal{L}(\hat{t}(\varphi))) + \pi \mathcal{L}(\hat{t}(\varphi)).$$

Solving for $(1 - r_K)/r_K$ yields

$$\frac{1 - r_K}{r_K} = \frac{(1 - \pi)\mathbb{E}(\boldsymbol{t}) + \pi \mathcal{L}(\hat{t}(\varphi) - \varphi}{\varphi - \mathcal{L}(\hat{t}(\varphi))}.$$

Using the indifference condition that $y(K) + \mu\varphi = y(j) + \mu\mathbb{E}(\boldsymbol{\mathcal{R}}^{\boldsymbol{C}}_{\boldsymbol{j}})$ for states j larger k^*, we get the function

$$D(\varphi, \pi) = \sum_{j=k^*+1}^{K} p(j) \frac{(1 - \pi)\mathbb{E}(\boldsymbol{t}) + \pi \mathcal{L}(\hat{t}(\varphi)) - (\varphi + \Delta(K, j)/\mu)}{\varphi + \Delta(K, j)/\mu - \mathcal{L}(\hat{t}(\varphi))}.$$

This function is decreasing in φ and π. Together with the function

$$S(\varphi) = \sum_{j=1}^{K} p(j) F(\hat{t}_j(\varphi)),$$

an equilibrium arises where $D(\varphi^*(\pi), \pi) - S(\varphi^*(\pi)) = 0$. This equation implicitly defines the equilibrium $\varphi^*(\pi)$ as a function of π. Consider two values π' and $\pi'' > \pi'$. It holds that

$$D(\varphi^*(\pi''), \pi'') - S(\varphi^*(\pi'')) = D(\varphi^*(\pi'), \pi') - S(\varphi^*(\pi')) = 0 > D(\varphi^*(\pi'), \pi'') - S(\varphi^*(\pi')).$$

As $D - S$ is decreasing in φ, it follows that $\varphi^*(\pi'') > \varphi^*(\pi')$. Since $S'(\varphi) > 0$, lying is higher under π' than under π'', which implies Part (ii).

To show Part (i), that k^* weakly increases, recall that the proof of Proposition 1 shows that k^* is the largest state to which a liar would not deviate to. Denote the threshold state under π' by $k^{*'}$. With a probability of lie detection π', this condition becomes

$$y(K) + \mu\varphi' \geq y(k^{*'}) + \mu[(1 - \pi)\mathbb{E}(\boldsymbol{t}) + \pi \mathcal{L}(\hat{t}(\varphi'))].$$

After increasing π, the reputation terms of both the r.h.s. and the l.h.s. will adjust. If the decrease in reputation on the r.h.s. is larger than the decrease in reputation on the l.h.s., this inequality becomes more binding, which implies that it potentially will also hold for $k^{*'} + 1$. If it holds for $k^{*'} + 1$, the threshold state increases. Consider a $\widetilde{\varphi} < \varphi'$ such that, under π'', this condition holds with equality;

$$y(K) + \mu\widetilde{\varphi} = y(k^{*'}) + \mu[(1 - \pi'')\mathbb{E}(\boldsymbol{t}) + \pi'' \mathcal{L}(\hat{t}(\widetilde{\varphi}))],$$

$$\Rightarrow \widetilde{\varphi} = (1-\pi'')\mathbb{E}(\boldsymbol{t}) + \pi''\mathcal{L}(\hat{t}(\widetilde{\varphi})) - \frac{\Delta(K,k^{*'})}{\mu}.$$

Plugging into D, we get

$$D(\widetilde{\varphi},\pi'') = \sum_{j=k^{*'}+1}^{K} p(j)\frac{\Delta(K,k^{*'})/\mu - \Delta(K,j)/\mu}{\widetilde{\varphi} + \Delta(K,j)/\mu - \mathcal{L}(\hat{t}(\widetilde{\varphi}))}.$$

Now consider φ'. From the threshold state condition under π', we know that

$$\varphi' \geq (1-\pi')\mathbb{E}(\boldsymbol{t}) + \pi'\mathcal{L}(\hat{t}(\varphi')) - \frac{\Delta(K,k^{*'})}{\mu}.$$

Therefore,

$$D(\varphi',\pi') \leq \sum_{j=k^{*'}+1}^{K} p(j)\frac{\Delta(K,k^{*'})/\mu - \Delta(K,j)/\mu}{\varphi' + \Delta(K,j)/\mu - \mathcal{L}(\hat{t}(\varphi'))}.$$

Comparing $D(\widetilde{\varphi},\pi'')$ and the r.h.s. above, we see that the numerators in all sum terms are the same while the denominators are always larger in $D(\varphi',\pi')$, since $\varphi' > \widetilde{\varphi}$. Therefore, $D(\widetilde{\varphi},\pi'') > D(\varphi',\pi')$. Since $D(\varphi'',\pi'') < D(\varphi',\pi')$ (the likelihood of lying decreases in π) and D is decreasing in φ, it follows that $\varphi'' > \widetilde{\varphi}$, suggesting that

$$y(K) + \mu\varphi'' > y(k^{*'}) + \mu[(1-\pi)\mathbb{E}(\boldsymbol{t}) + \pi\mathcal{L}(\hat{t}(\varphi''))];$$

the threshold state weakly increases in π.

A.7. Proof of Proposition 6b

Denote the investigator's policy by (π,γ). The variable γ denotes the probability reveals a liar's observed state. If $\gamma = 0$, we are under the coarse disclosure regime. Denote the part of a liar's expected reputation when reporting K that is independent of $\mathcal{M}^-(\hat{t}_j)$ as

$$\varphi^I = (1-\pi)[(r_K\mathbb{E}(\boldsymbol{t}) + (1-r_K)\mathcal{L}(\hat{t}))] + \pi(1-\gamma)\mathcal{L}(\hat{t}).$$

In equilibrium, as liars are indifferent between reporting states $j > k^*$, it holds that $y(K) + \mu\varphi^I = y(j) + \mu\mathcal{R}_j^I$, where $\mathcal{R}_j^I$ is the part of the expected reputation of reporting j that is independent of the liar's observed state (analogously to φ^I). The equilibrium described in Proposition 1 remains an equilibrium as long as an agent from a state $j > k^*$ does not have an incentive to lie, in which case their payoff depends on the (off-equilibrium) reputation of being a liar from a state $j > k^*$. We can determine this off-equilibrium belief using the refinement. Consider an agent of type (j,t) and who considers reporting j', with $j' > j > k^*$. Suppose that this is the agent with the strongest incentive to deviate from their equilibrium action. Then, the off-equilibrium refinement pins down the reputation after being disclosed at t. With this reputation, the agent sticks to the equilibrium strategy of being honest if

$$y(j) + \mu[(1-\pi)(r_j\mathbb{E}(\boldsymbol{t}) + (1-r_j)\mathcal{L}(\hat{t}(\varphi))) + \pi\mathbb{E}(\boldsymbol{t})] > y(j') - t + \mu(R_{j'}^I + \pi\gamma t)$$
$$\Rightarrow t(\mu\pi\gamma - 1) < \mu\pi[r_j\mathbb{E}(\boldsymbol{t}) + (1-r_j)((1-\pi)\mathcal{L}(\hat{t}(\varphi)) + \pi\mathbb{E}(\boldsymbol{t})) - (1-\gamma)\mathcal{L}(\hat{t}(\varphi))].$$

Since the r.h.s. is larger than zero, a sufficient condition for this to hold for all t is that $\mu\pi\gamma < 1$, which holds if μ is not too large. In this case, the type who has the strongest incentive to deviate is the one with $t \to 0$. Therefore, the off-equilibrium reputation is $\mathcal{M}^-(0)$.

A liar from a state $j \leq k^*$ reporting K has an expected reputation of $\varphi^I + \pi\gamma\mathcal{M}^-(\hat{t}_j)$. The threshold function becomes

$$\mathcal{T}(\Delta(K,j),\varphi^I,\pi,\gamma) \equiv t + \mu[\mathcal{M}^+(t) - \varphi^I - \pi\gamma\mathcal{M}^-(t)] - \Delta(K,j) = 0,$$

so that the threshold $\hat{t}_j(\varphi^I,\pi,\gamma)$ now depends on π and γ. We denote the equilibrium threshold vector by $\hat{t}^*$. Consider a marginal increase in γ. The thresholds change by

$$\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\gamma} = \frac{\partial\hat{t}_j}{\partial\varphi^I} \times \left(\frac{\mathrm{d}\varphi^{I*}}{\mathrm{d}\gamma} + \pi\mathcal{M}^-(\hat{t}_j^*)\right).$$

Under the uniform distribution, $\frac{\partial\hat{t}_j}{\partial\varphi^I} = \frac{\partial\hat{t}_k}{\partial\varphi^I} > 0$ for $j,k \leq k^*$ and zero otherwise. The aggregate lying rate is $\sum_{j=1}^{K} p(j)\frac{\hat{t}_j}{\bar{t}}$, so that it decreases after a marginal increase in γ if

$$-\sum_{j=1}^{k^*} p(j)\frac{\mathrm{d}\varphi^{I*}}{\mathrm{d}\gamma} > \sum_{j=1}^{K} p(j)\pi\mathcal{M}^-(\hat{t}_j^*). \tag{A.12}$$

To derive $\mathrm{d}\varphi^{C*}/\mathrm{d}\gamma$, consider that equilibrium can be characterized by the function

$$h(\varphi,\gamma)=\varphi-(1-\pi)(r_K(\hat{t}(\varphi,\gamma))\mathbb{E}(\boldsymbol{t})+(1-r_K(\hat{t}(\varphi,\gamma)))\mathcal{L}(\hat{t}(\varphi,\gamma)))-\pi(1-\gamma)\mathcal{L}(\hat{t}(\varphi,\gamma))$$

and where the equilibrium $\varphi^{C*}(\gamma)$ solves $h(\varphi^{C*}(\gamma),\gamma)=0$. Applying the implicit function theorem, we have

$$\frac{\mathrm{d}\varphi^{I*}}{\mathrm{d}\gamma}=-\frac{\partial h/\partial\gamma}{\partial h/\partial\varphi},$$

where the two partial derivatives are

$$\begin{aligned}\frac{\partial h}{\partial\varphi}=&1-(1-\pi)[\frac{\partial r_K}{\partial\varphi}(\mathbb{E}(\boldsymbol{t})-\mathcal{L}(\hat{t}(\varphi,\gamma)))+(1-r_K(\hat{t}(\varphi,\gamma)))\frac{\partial\mathcal{L}}{\partial\varphi}]-\pi(1-\gamma)\frac{\partial\mathcal{L}}{\partial\varphi}\\ \frac{\partial h}{\partial\gamma}=&-(1-\pi)[\frac{\partial r_K}{\partial\gamma}(\mathbb{E}(\boldsymbol{t})-\mathcal{L}(\hat{t}(\varphi,\gamma)))+(1-r_K(\hat{t}(\varphi,\gamma)))\frac{\partial\mathcal{L}}{\partial\gamma}]-\pi(1-\gamma)\frac{\partial\mathcal{L}}{\partial\gamma}+\pi\mathcal{L}(\hat{t}(\varphi,\gamma)).\end{aligned}$$

Therefore, Inequality (A.12) can be written as

$$\begin{aligned}&\sum_{j=1}^{k^*}p(j)\frac{\partial h}{\partial\gamma}>\frac{\partial h}{\partial\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\\ \Rightarrow&(1-\pi)(\mathbb{E}(\boldsymbol{t})-\mathcal{L}(\hat{t}(\varphi,\gamma)))\left(\sum_{j=1}^{k^*}p(j)\frac{\mathrm{d}r_K}{\mathrm{d}\gamma}-\frac{\mathrm{d}r_K}{\mathrm{d}\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\right)+\\ &((1-\pi)(1-r_K(\hat{t}(\varphi,\gamma)))+\pi(1-\gamma))\left(\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\gamma}\sum_{j=1}^{k^*}p(j)-\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\right)\\ &<\sum_{j=1}^{k^*}p(j)\pi\mathcal{L}(\hat{t}(\varphi,\gamma))-\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*).\end{aligned}$$

Consider the term

$$\begin{aligned}\sum_{j=1}^{k^*}p(j)\frac{\mathrm{d}r_K}{\mathrm{d}\gamma}-\frac{\mathrm{d}r_K}{\mathrm{d}\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)&=\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}\frac{\partial r_K}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\gamma}-\sum_{j=1}^{K}\frac{\partial r_K}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\\ &=\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}\frac{\partial r_K}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\pi\mathcal{M}^-(\hat{t}_j^*)\\ &\quad-\sum_{j=1}^{K}\frac{\partial r_K}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*).\end{aligned}$$

We observe that both $\partial r_K/\partial\hat{t}_j\times\partial\hat{t}_j/\partial\varphi=\partial r_K/\partial\hat{t}_1\partial\hat{t}_1/\partial\varphi$ for $j\le k^*$ and zero otherwise. Therefore, the derivative terms can be moved out of the first sum and the second derivative sum can be written as $\sum_{j=1}^{k^*}p(j)\partial r_K/\partial\hat{t}_1\times\partial\hat{t}_1/\partial\varphi$. It then becomes apparent that the whole term is equal to zero. Moving on to the term

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\gamma}\sum_{j=1}^{k^*}p(j)-\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)&=\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\gamma}-\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\\ &=\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\pi\mathcal{M}^-(\hat{t}_j^*)-\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\sum_{j=1}^{K}p(j)\pi\mathcal{M}^-(\hat{t}_j^*)\\ &=\pi\frac{\partial\hat{t}_1}{\partial\varphi}\left(\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}p(j)\mathcal{M}^-(\hat{t}_j^*)-\sum_{j=1}^{K}\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\sum_{j=1}^{K}p(j)\mathcal{M}^-(\hat{t}_j^*)\right)\\ &=\pi\frac{\partial\hat{t}_1}{\partial\varphi}\left(\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^*-\mathcal{L}(\hat{t})}{\sum_{l\in\mathcal{K}}p(l)\hat{t}_l}\frac{\hat{t}_j^*}{2}\right.\\ &\quad\left.-\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^*-\mathcal{L}(\hat{t})}{\sum_{l\in\mathcal{K}}p(l)\hat{t}_l}\sum_{j=1}^{K}p(j)\mathcal{M}^-(\hat{t}_j^*)\right)\\ &=\pi\frac{\partial\hat{t}_1}{\partial\varphi}\left(\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^{*2}}{2\sum_{l\in\mathcal{K}}\hat{t}_l}-\frac{\mathcal{L}(\hat{t})}{\sum_{l\in\mathcal{K}}p(l)\hat{t}_l}\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^*}{2}\right.\\ &\quad\left.-\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^*-\mathcal{L}(\hat{t})}{\sum_{l\in\mathcal{K}}\hat{t}_l}\sum_{j=1}^{K}p(j)\mathcal{M}^-(\hat{t}_j^*)\right)\end{aligned}$$

$$
\begin{aligned}
=&\pi\frac{\partial\hat{t}_1}{\partial\varphi}\left(\sum_{j=1}^{k^*}p(j)\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^{*2}}{2\sum_{l\in\mathcal{K}}p(l)\hat{t}_l}\right.\\
&\left.-\sum_{j=1}^{K}p(j)\frac{\hat{t}_j^{*}}{\sum_{l\in\mathcal{K}}p(l)\hat{t}_l}\sum_{j=1}^{K}p(j)\mathcal{M}^{-}(\hat{t}_j^{*})\right)\\
=&\pi\frac{\partial\hat{t}_1}{\partial\varphi}\sum_{j=1}^{k^*}p(j)\left(\mathcal{L}(\hat{t}(\varphi,\gamma))-\frac{1}{\sum_{j=1}^{k^*}p(j)}\sum_{j=1}^{K}p(j)\mathcal{M}^{-}(\hat{t}_j^{*})\right).
\end{aligned}
$$

The inequality therefore reduces to

$$
\begin{aligned}
&((1-\pi)(1-r_K(\hat{t}(\varphi,\gamma)))+\pi(1-\gamma))\pi\frac{\partial\hat{t}_1}{\partial\varphi}\sum_{j=1}^{k^*}p(j)\left(\mathcal{L}(\hat{t}(\varphi,\gamma))-\frac{1}{\sum_{j=1}^{k^*}p(j)}\sum_{j=1}^{K}p(j)\mathcal{M}^{-}(\hat{t}_j^{*})\right)\\
&<\pi\sum_{j=1}^{k^*}p(j)\left(\mathcal{L}(\hat{t}(\varphi,\gamma))-\frac{1}{\sum_{j=1}^{k^*}p(j)}\sum_{j=1}^{K}p(j)\mathcal{M}^{-}(\hat{t}_j^{*})\right)\\
\Rightarrow&((1-\pi)(1-r_K(\hat{t}(\varphi,\gamma)))+\pi(1-\gamma))\frac{\partial\hat{t}_1}{\partial\varphi}<1.
\end{aligned}
$$

This holds as long as $\frac{\partial\hat{t}_1}{\partial\varphi}<1$, which holds if μ is not too large (see Lemma 2). Therefore, Inequality (A.12) holds, which implies that the likelihood that an agent lies decreases in γ. The average size of the lie increases because $\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\gamma}$ decreases in j, so that there are relatively more liars who observed lower states.

A.8. Proof of Proposition 6c

With deed-based image concerns, the threshold that denotes the moral type who is indifferent between lying and telling the truth after observing j is equal to

$$\dot{\hat{t}}_j(\varphi)=\Delta(K,j)+\mu(\varphi-1),$$

where φ denotes the reputation of reporting K. In equilibrium, the expected reputation of a liar reporting K is equal to

$$\varphi=(1-\pi)\times r_K.$$

It follows that

$$\frac{1-r_K(\varphi,\pi)}{r_K(\varphi,\pi)}=\frac{1-\pi-\varphi}{\varphi}.$$

Since liars have to be indifferent, the reputation for reporting $j\in(k^*,K)$ can be derived from

$$
\begin{aligned}
y(K)+\mu r_K(\varphi,\pi)=&y(j)+\mu r_j\\
\Rightarrow\frac{1-r_j(\varphi,\pi)}{r_j(\varphi,\pi)}=&\frac{1-\pi-\varphi-\Delta(K,j)/\mu}{\varphi+\Delta(K,j)/\mu}.
\end{aligned}
$$

Similar arguments as those given in the proof of Proposition 1 imply that, in equilibrium $D-S=0$, where

$$
\begin{aligned}
\dot{S}(\varphi)&=\sum_{j=1}^{K}p(j)F(\dot{\hat{t}}_j(\varphi))\\
\dot{D}(\varphi,\pi)&=\sum_{j=k^*+1}^{K}p(j)\frac{1-\pi-\varphi-\Delta(K,j)/\mu}{\varphi+\Delta(K,j)/\mu}.
\end{aligned}
$$

The equilibrium condition

$$\dot{D}(\varphi^{D*}(\pi),\pi)-\dot{S}(\varphi^{D*}(\pi))=0$$

implicitly defines the equilibrium reputation associated with reporting K for a given π, $\varphi^{D*}(\pi)$. Consider two values $\pi''>\pi'$. It holds that

$$\dot{D}(\varphi^{D*}(\pi''),\pi'')-\dot{S}(\varphi^{D*}(\pi''))=\dot{D}(\varphi^{D*}(\pi'),\pi')-\dot{S}(\varphi^{D*}(\pi'))=0>\dot{D}(\varphi^{D*}(\pi'),\pi'')-\dot{S}(\varphi^{D*}(\pi')).$$

Since $D-S$ is decreasing in φ, we conclude that $\varphi^{D*}(\pi'')<\varphi^{D*}(\pi')$. We conclude (ii): lying is higher under π' than under π''.

(i). Under π', there is a threshold state $k^{*'}$ which is the largest integer such that

$$y(K) + \mu\varphi^{D'} > y(k^{*'}) + \mu(1-\pi).$$

Now consider a $\widetilde{\varphi}$ such that

$$y(K) + \mu\widetilde{\varphi} = y(k^{*'}) + \mu(1-\pi'')$$
$$\Rightarrow \widetilde{\varphi} = 1 - \pi'' - \frac{\Delta(K, k^{*'})}{\mu}.$$

We are going to check whether $\varphi^{D''} < \widetilde{\varphi}$. If this were the case, then the threshold state would decrease after an increase in π. Plugging into D, we find that

$$\dot{D}(\widetilde{\varphi}, \pi') = \sum_{j=k^{*'}+1}^{K} p(j) \frac{\Delta(K, k^{*'})/\mu - \Delta(K, j)/\mu}{1 - \pi'' - \Delta(K, k^{*'})/\mu + \Delta(K, j)/\mu}.$$

Now consider $\varphi^{D'}$. From the threshold state condition under π', we know that

$$\varphi^{D'} \geq 1 - \pi' - \frac{\Delta(K, k^{*'})}{\mu}.$$

Therefore,

$$\dot{D}(\varphi^{D'}, \pi') \leq \sum_{j=k^{*'}+1}^{K} p(j) \frac{\Delta(K, k^{*'})/\mu - \Delta(K, j)/\mu}{1 - \pi' - \Delta(K, k^{*'})/\mu + \Delta(K, j)/\mu}.$$

Comparing $\dot{D}(\widetilde{\varphi}, \pi'')$ and r.h.s. above, we see that the numerators are the same while the denominators are always larger in $\dot{D}(\varphi^{D'}, \pi')$. Therefore, $\dot{D}(\widetilde{\varphi}, \pi'') > \dot{D}(\varphi^{D'}, \pi')$. Since D is decreasing in φ and $\dot{D}(\varphi^{D''}, \pi'') < \dot{D}(\varphi^{D'}, \pi')$ (the likelihood of lying decreases in π) it follows that $\varphi^{D''} > \widetilde{\varphi}$. This suggests that the threshold state weakly increases in π.

A.9. Proof of Proposition 7a

Note that only agents who observed $z = 1$ can participate in the lying game. Therefore, $\mathbb{E}(\boldsymbol{t}|a=1, z=1) = \mathbb{E}(\boldsymbol{t}|a=1)$ and $\mathbb{E}(\boldsymbol{t}|a=2, z=1) = \mathbb{E}(\boldsymbol{t}|a=2)$. I.e., conditional on making a report in the lying game the realization of z does not add any additional information. Therefore, in the lying game, agents lie if and only if

$$\Delta(2,1) - t \geq \mu(\mathbb{E}(\boldsymbol{t}|a=1) - \mathbb{E}(\boldsymbol{t}|a=2)).$$

This equation suggests a threshold rule where agents lie if they are of a type $t \leq \hat{t}_{LG}$, where

$$\Delta(2,1) - \hat{t}_{LG} = \mu(\mathbb{E}(\boldsymbol{t}|a=1) - \mathbb{E}(\boldsymbol{t}|a=2)).$$

The utility that agents with $t \leq \hat{t}_{LG}$ expect to derive from indicating interest is

$$(\varepsilon + q)(y(2) - t(1-p) + \mu\mathbb{E}(\boldsymbol{t}|a=2)) + (1-\varepsilon-q)(py(2) + (1-p)y(1) + \mu\mathbb{E}(\boldsymbol{t}|z=3)).$$

Their expected utility from not indicating interest is

$$\varepsilon(y(2) - t(1-p) + \mu\mathbb{E}(\boldsymbol{t}|a=2)) + (1-\varepsilon)(py(2) + (1-p)y(1) + \mu\mathbb{E}(\boldsymbol{t}|z=2)).$$

Combining these equations, an agent with $t \leq \hat{t}_{LG}$ indicates interest if and only if

$$q\{(1-p)(\Delta(2,1) - t) + \mu(\mathbb{E}(\boldsymbol{t}|a=2) - \mathbb{E}(\boldsymbol{t}|z=3))\} \geq (1-\varepsilon)\mu(\mathbb{E}(\boldsymbol{t}|z=2) - \mathbb{E}(\boldsymbol{t}|z=3)). \tag{A.13}$$

An agent with $t > \hat{t}_{LG}$ will instead expect the following utility when indicating interest:

$$(\varepsilon + q)(py(2) + (1-p)y(1) + \mu(p\mathbb{E}(\boldsymbol{t}|a=2) + (1-p)\mathbb{E}(\boldsymbol{t}|a=1))) +$$
$$(1-\varepsilon-q)(py(2) + (1-p)y(1) + \mu\mathbb{E}(\boldsymbol{t}|z=3))$$

and the following when not indicating interest:

$$\varepsilon(py(2) + (1-p)y(1) + \mu(p\mathbb{E}(\boldsymbol{t}|a=2) + (1-p)\mathbb{E}(\boldsymbol{t}|a=1))) +$$
$$(1-\varepsilon)(py(2) + (1-p)y(1) + \mu\mathbb{E}(\boldsymbol{t}|z=2)).$$

Therefore, an agent with $t > \hat{t}_{LG}$ will indicate interest if and only if

$$q\{\mu(p\mathbb{E}(\boldsymbol{t}|a=2)+(1-p)\mathbb{E}(\boldsymbol{t}|a=1))-\mu\mathbb{E}(\boldsymbol{t}|z=3)\}\geq(1-\varepsilon)\mu(\mathbb{E}(\boldsymbol{t}|z=2)-\mathbb{E}(\boldsymbol{t}|z=3)). \tag{A.14}$$

These equations suggest that agents with $t\leq\hat{t}_{LG}$ follow a threshold rule when deciding to indicate interest or not while other agents do not. Denote this threshold by $\hat{t}_C$. There can essentially be four cases which differ in the relation between $\hat{t}_C$ and $\hat{t}_{LG}$ and whether agents with $t>\hat{t}_{LG}$ indicate interest. They constitute our candidate equilibria.

Case 1: $\hat{t}_C<\hat{t}_{LG}$, agents $t>\hat{t}_{LG}$ indicate interest. Suppose that this is the case. Then, the lying game threshold is equal to

$$\Delta(2,1)-\hat{t}_{LG}=\mu(\mathbb{E}(\boldsymbol{t}|a=1)-\mathbb{E}(\boldsymbol{t}|a=2)).$$

By $\hat{t}_C<\hat{t}_{LG}$, we know that Inequality (A.13) does not hold when evaluated at $\hat{t}_{LG}$. Combining this insight with the fact that Inequality (A.14) holds, we get

$$\begin{aligned}
&q\left\{(1-p)(\Delta(2,1)-t_{LG})+\mu(\mathbb{E}(\boldsymbol{t}|a=2)-\mathbb{E}(\boldsymbol{t}|z=3))\right\}\\
&<q\{\mu(p\mathbb{E}(\boldsymbol{t}|a=2)+(1-p)\mathbb{E}(\boldsymbol{t}|a=1))-\mu\mathbb{E}(\boldsymbol{t}|z=3)\}\\
\Rightarrow&(1-p)(\Delta(2,1)-t_{LG})+\mu\mathbb{E}(\boldsymbol{t}|a=2)<\mu(p\mathbb{E}(\boldsymbol{t}|a=2)+(1-p)\mathbb{E}(\boldsymbol{t}|a=1))\\
\Rightarrow&\Delta(2,1)-t_{LG}<\mu(\mathbb{E}(\boldsymbol{t}|a=1)-\mathbb{E}(\boldsymbol{t}|a=2)),
\end{aligned}$$

a contradiction. Therefore, this cannot be an equilibrium.

Case 2: $\hat{t}_C\geq\hat{t}_{LG}$, agents $t>\hat{t}_{LG}$ indicate interest. If all agents indicate interest, the off-equilibrium utility of an agent who does not indicate interest and does not participate in the lying game is

$$py(2)+(1-p)y(1)+\mu\mathbb{E}(\boldsymbol{t}|z=2).$$

Under the equilibrium refinement, the off-equilibrium belief is equal to the agent type who has the strongest incentive to deviate from indicating interest. Among agents with $t\leq\hat{t}_{LG}$, the type $\hat{t}_{LG}$ has the lowest expected utility from indicating interest and the highest expected utility from not indicating interest (under the refinement if $\hat{t}_{LG}$ has the strongest incentive of all agents, $\mathbb{E}(\boldsymbol{t}|z=2)=\hat{t}_{LG}$). Agents with $t>\hat{t}_{LG}$ have the same expected utility from indicating interest while the type $\bar{t}$ has the strongest utility from not indicating interest (under the refinement if $\bar{t}$ has the strongest incentive of all agents, $\mathbb{E}(\boldsymbol{t}|a=z=2)=\bar{t}$). Now note two things: First, the expected utility from indicating interest for $\hat{t}_{LG}$ is equal to the expected utility of indicating interest of $\bar{t}$, since $\hat{t}_{LG}$ is indifferent between lying and not lying. Second, the expected utility from not indicating interest is higher for $\bar{t}$ than for $\hat{t}_{LG}$. Therefore, $\bar{t}$ has the strongest incentive amongst all types to deviate from indicating interest. This implies that the off-equilibrium belief $\mathbb{E}(\boldsymbol{t}|z=2)$ is equal to $\bar{t}$. Plugging this into Inequality (A.14) yields

$$\begin{aligned}
&q\left\{\mu(p\mathbb{E}(\boldsymbol{t}|a=2)+(1-p)\mathcal{M}^+(\hat{t}_{LG}))-\mu\mathbb{E}(\boldsymbol{t})\right\}\geq(1-\varepsilon)\mu(\bar{t}-\mathbb{E}(\boldsymbol{t})).\\
\Rightarrow&q\left\{(p\mathbb{E}(\boldsymbol{t}|a=2)+(1-p)\mathcal{M}^+(\hat{t}_{LG}))-\mathbb{E}(\boldsymbol{t})\right\}-q(1-\varepsilon)\bar{t}+(1-q)(1-\varepsilon)(\mathbb{E}(\boldsymbol{t})-\bar{t})+\varepsilon\mathbb{E}(\boldsymbol{t})\geq0,
\end{aligned}$$

which does not hold as $\varepsilon\to0$, yielding a contradiction. Therefore, this cannot be an equilibrium.

Case 3: $\hat{t}_C\geq\hat{t}_{LG}$, agents $t>\hat{t}_{LG}$ do not indicate interest. In this case, the reputations become $\mathbb{E}(\boldsymbol{t}|z=2)=\mathcal{M}^+(\hat{t}_{LG})$, $\mathbb{E}(\boldsymbol{t}|z=3)=\mathcal{M}^-(\hat{t}_{LG})$, $\mathbb{E}(\boldsymbol{t}|a=1)=\mathcal{M}^+(\hat{t}_{LG})$,

$$\mathbb{E}(\boldsymbol{t}|a=2)=\frac{(q+\varepsilon)F(\hat{t}_{LG})\mathcal{M}^-(\hat{t}_{LG})+\varepsilon p(1-F(\hat{t}_{LG}))\mathcal{M}^+(\hat{t}_{LG})}{(q+\varepsilon)F(\hat{t}_{LG})+\varepsilon p(1-F(\hat{t}_{LG}))}.$$

As $\varepsilon\to0$ this last reputation term becomes $\mathbb{E}(\boldsymbol{t}|a=2)=\mathcal{M}^-(\hat{t}_{LG})$. Therefore, the threshold for the lying game solves

$$\Delta(2,1)-\hat{t}_{LG}=\mu(\mathcal{M}^+(\hat{t}_{LG})-\mathcal{M}^-(\hat{t}_{LG})).$$

When evaluated at $\hat{t}_{LG}$, the Inequality (A.13) has to hold strictly. Plugging in, this suggests that

$$\begin{aligned}
&q\left\{(1-p)(\Delta(2,1)-\hat{t}_{LG})+\mu(\mathcal{M}^-(\hat{t}_{LG})-\mathcal{M}^-(\hat{t}_{LG}))\right\}>(1-\varepsilon)\mu(\mathcal{M}^+(\hat{t}_{LG})-\mathcal{M}^-(\hat{t}_{LG}))\\
\Rightarrow&q(1-p)(\Delta(2,1)-\hat{t}_{LG})>(1-\varepsilon)\mu(\mathcal{M}^+(\hat{t}_{LG})-\mathcal{M}^-(\hat{t}_{LG}))\\
\Rightarrow&\Delta(2,1)-\hat{t}_{LG}>\frac{\mu}{q(1-p)}(1-\varepsilon)(\mathcal{M}^+(\hat{t}_{LG})-\mathcal{M}^-(\hat{t}_{LG})).
\end{aligned}$$

As $\varepsilon\to0$, this inequality does not hold, yielding a contradiction. Therefore, this cannot be an equilibrium.

Case 4: $\hat{t}_C<\hat{t}_{LG}$, agents $t>\hat{t}_{LG}$ do not indicate interest. The reputations become $\mathbb{E}(\boldsymbol{t}|z=2)=\mathcal{M}^+(\hat{t}_C)$, $\mathbb{E}(\boldsymbol{t}|z=3)=\mathcal{M}^-(\hat{t}_C)$, and $\mathbb{E}(\boldsymbol{t}|a=1)=\mathcal{M}^+(t_{LG})$. We can derive an expression for the reputation term

$$\mathbb{E}(\boldsymbol{t}|a=2)=\frac{(q+\varepsilon)F(\hat{t}_C)\mathcal{M}^-(\hat{t}_C)+\varepsilon[(1-p)(F(\hat{t}_{LG})-F(\hat{t}_C))\mathbb{E}(\boldsymbol{t}|\boldsymbol{t}\in(\hat{t}_C,\hat{t}_{LG}))+p(1-F(\hat{t}_C))\mathcal{M}^+(\hat{t}_C)]}{(q+\varepsilon)F(\hat{t}_C)+\varepsilon[(1-p)(F(\hat{t}_{LG})-F(\hat{t}_C))+p(1-F(\hat{t}_C))]},$$

which, taking the limit $\varepsilon \to 0$ becomes $\mathbb{E}(t|a=2) = \mathcal{M}^-(\hat{t}_C)$. Inequality (A.13) holds with equality when evaluated at $\hat{t}_C$, so that

$$\Delta(2,1) - \hat{t}_C = \frac{\mu}{q(1-p)}((\mathcal{M}^+(\hat{t}_C) - \mathcal{M}^-(\hat{t}_C)))$$

and likewise

$$\Delta(2,1) - \hat{t}_{LG} = \mu(\mathcal{M}^+(\hat{t}_{LG}) - \mathcal{M}^-(\hat{t}_C)).$$

This is the only candidate equilibrium that exists.

Properties.

(i). Taking the implicit derivative of

$$T(\hat{t}_C, q, p) = \hat{t}_C + \frac{\mu}{q(1-p)}(\mathcal{M}^+(\hat{t}_C) - \mathcal{M}^-(\hat{t}_C)) - \Delta(2,1) = 0 \tag{A.15}$$

with respect to q yields

$$\frac{\mathrm{d}\hat{t}_C}{\mathrm{d}q} = (1-p)\frac{\frac{\mu}{(q(1-p))^2}(\mathcal{M}^+(\hat{t}_C) - \mathcal{M}^-(\hat{t}_C))}{1 + \frac{\mu}{q(1-p)}(\mathcal{M}^{+\prime}(\hat{t}_C) - \mathcal{M}^{-\prime}(\hat{t}_C))} > 0$$

as long as we maintain the assumption that the denominator is positive, i.e., that the equilibrium is unique. Therefore, lying increases in q.

(ii). Taking the implicit derivative of Equation (A.15) with respect to p yields

$$\frac{\mathrm{d}\hat{t}_C}{\mathrm{d}p} = -q\frac{\frac{\mu}{(q(1-p))^2}(\mathcal{M}^+(\hat{t}_C) - \mathcal{M}^-(\hat{t}_C))}{1 + \frac{\mu}{q(1-p)}(\mathcal{M}^{+\prime}(\hat{t}_C) - \mathcal{M}^{-\prime}(\hat{t}_C))} < 0.$$

(iii). This follows as, conditional on $z = j = 1$, the likelihood that an agent lies is equal to 1, independent of p.

A.10. Proof of Proposition 7b

In the lying game, agents will lie if

$$\Delta(2,1) - t \geq \mu(\mathrm{P}(\text{honest}|a=2) - \mathrm{P}(\text{honest}|a=1)).$$

Since in equilibrium with deed-based image concerns there is no downward lying (Gneezy et al., 2018), $\mathrm{P}(\text{honest}|a=1) = 1$. This equation suggests a threshold rule where agents lie if they are of a type $t \leq \hat{t}_D$, where

$$\Delta(2,1) - \hat{t}_D = \mu(\mathrm{P}(\text{honest}|a=2) - 1).$$

An agent with $t \leq \hat{t}_D$ expects the following utility when indicating interest:

$$(\varepsilon + q)(y(2) - t(1-p) + \mu\mathrm{P}(\text{honest}|a=2)) + (1 - \varepsilon - q)(py(2) + (1-p)y(1)).$$

Their expected utility from not indicating interest is

$$\varepsilon(y(2) - t(1-p) + \mu\mathrm{P}(\text{honest}|a=2)) + (1-\varepsilon)(py(2) + (1-p)y(1)).$$

Combining these equations, an agent with $t \leq \hat{t}_{LG}$ will indicate interest if and only if

$$q\{(1-p)(\Delta(2,1) - t) + \mu\mathrm{P}(\text{honest}|a=2)\} \geq 0.$$

Since $\mathrm{P}(\text{honest}|a=2) > 0$, this equation always holds. An agent with $t > \hat{t}_D$ will instead expect the following utility when indicating interest:

$$(\varepsilon + q)(py(2) + (1-p)y(1) + \mu(p\mathrm{P}(\text{honest}|a=2) + (1-p))) +$$
$$(1 - \varepsilon - q)(py(2) + (1-p)y(1))$$

and the following when not indicating interest

$$\varepsilon(py(2) + (1-p)y(1) + \mu(p\mathrm{P}(\text{honest}|a=2) + (1-p))) +$$
$$(1-\varepsilon)(py(2) + (1-p)y(1)).$$

Therefore, an agent with $t > \hat{t}_D$ indicates interest if and only if

$$q\{\mu(p\mathrm{P}(\text{honest}|a=2)+(1-p))\}\geq 0.$$

This equation holds always. Therefore, all agents indicate interest. The equilibrium reputation becomes

$$\mathrm{P}(\text{honest}|a=2)=\frac{p}{p+(1-p)F(\hat{t}_D)}.$$

Properties. Parts (i) and (ii) follow from the fact that all agents always indicate interest.

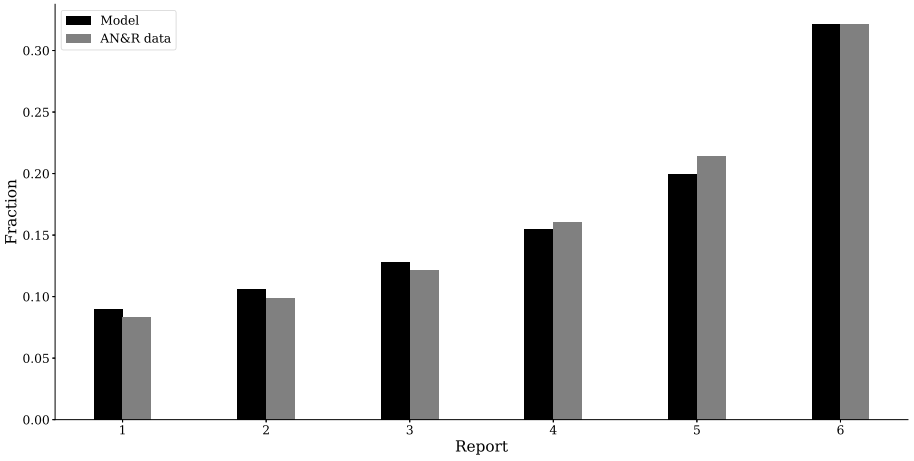
(iii). The threshold type for whom $z=j=1$ and who is indifferent between lying and truth-telling is defined in the equation

$$\Delta(2,1)-\hat{t}_D=\mu\left(1-\frac{p}{p+(1-p)F(\hat{t}_D)}\right).$$

In the threshold type equation, the r.h.s. is decreasing in p. Therefore, the threshold type is increasing in p, which implies the claim.

Appendix B. Calibration of the character-based model

Fig. B.4 compares the predicted equilibrium distribution for a calibrated version of the model to the data collected by AN&R. The model comes close to the observed frequency distribution and in particular can account for partial lying.



Note: Example equilibrium distribution of reports when lying costs follow a lognormal distribution, where log-costs have mean zero and standard deviation 1.1, where $y(a)-y(a-1)=1$ for all $a\in\{2,\ldots,K\}$, where $p(j)=1/K$ for all $j\in\mathcal{K}$, and where $\mu=2.1$.

Fig. B.4. Example equilibrium report distribution compared to the AN&R data.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geb.2024.08.006.

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87, 1115–1153.
Adriani, F., Sonderegger, S., 2019. A theory of esteem based peer pressure. Games Econ. Behav. 115, 314–335.
Akerlof, G.A., 1970. The market for "lemons": quality uncertainty and the market mechanism. Q. J. Econ. 84, 488–500.
Akın, Z., 2019. Dishonesty, social information, and sorting. J. Behav. Exp. Econ. 80, 199–210.
Barfort, S., Harmon, N.A., Hjorth, F., Olsen, A.L., 2019. Sustaining honesty in public service: the role of selection. Am. Econ. J.: Econ. Policy 11, 96–123.
Bašić, Z., Quercia, S., 2022. The influence of self and social image concerns on lying. Games Econ. Behav. 133, 162–169.
Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.
Battigalli, P., Dufwenberg, M., 2022. Belief-dependent motivations and psychological game theory. J. Econ. Lit. 60, 833–882.
Bénabou, R., Falk, A., Tirole, J., 2020. Narratives, Imperatives, and Moral Persuasion. Mimeo.
Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96, 1652–1678.
Bénabou, R., Tirole, J., 2011. Laws and Norms. NBER Working Paper.
Bicchieri, C., Dimant, E., Sonderegger, S., 2023. It's not a Lie if you believe the norm does not apply: conditional norm-following and belief distortion. Games Econ. Behav. 138, 321–354.
Braithwaite, J., 1989. Crime, Shame and Reintegration. Cambridge University Press, New York.
Cohn, A., Maréchal, M.A., Tannenbaum, D., Zünd, C.L., 2019. Civic honesty around the globe. Science 365, 70–73.
Crawford, V.P., Sobel, J., 1982. Strategic information transmission. Econometrica 50, 1431–1451.
Diekmann, A., Przepiorka, W., Rauhut, H., 2015. Lifting the veil of ignorance: an experiment on the contagiousness of norm violations. Ration. Soc. 27, 309–333.

Dufwenberg, M., Dufwenberg, M.A., 2018. Lies in disguise – a theoretical analysis of cheating. J. Econ. Theory 175, 248–264.
Dufwenberg, M., Lundholm, M., 2001. Social norms and moral hazard. Econ. J., 506–525.
Eliaz, K., Spiegler, R., 2020. A model of competing narratives. Am. Econ. Rev. 110, 3786–3816.
Feess, E., Kerzenmacher, F., 2018. Lying opportunities and incentives to Lie: reference dependence versus reputation. Games Econ. Behav. 111, 274–288.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise-an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.
Foerster, M., van der Weele, J.J., 2021. Casting doubt: image concerns and the communication of social impact. Econ. J. 131, 2887–2919.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Gibson, R., Tanner, C., Wagner, A.F., 2013. Preferences for truthfulness: heterogeneity among and within individuals. Am. Econ. Rev. 103, 532–548.
Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the Lie. Am. Econ. Rev. 108, 419–453.
Gneezy, U., Rockenbach, B., Serra-Garcia, M., 2013. Measuring lying aversion. J. Econ. Behav. Organ. 93, 293–300.
Golman, R., 2023. Acceptable discourse: social norms of beliefs and opinions. Eur. Econ. Rev. 160, 104588.
Haaland, I., Roth, C., Wohlfart, J., 2023. Designing information provision experiments. J. Econ. Lit. 61, 3–40.
Hanna, R., Wang, S.-Y., 2017. Dishonesty and selection into public service: evidence from India. Am. Econ. J.: Econ. Policy 9, 262–290.
Hillenbrand, A., Verrina, E., 2022. The asymmetric effect of narratives on prosocial behavior. Games Econ. Behav. 135, 241–270.
Houdek, P., Bahník, Š., Hudík, M., Vranka, M., 2021. Selection effects on dishonest behavior. Judgm. Decis. Mak. 16, 238–266.
Hursthouse, R., Pettigrove, G., 2018. Virtue ethics. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Winter 2018 Edition.
Kajackaite, A., Gneezy, U., 2017. Incentives and cheating. Games Econ. Behav. 102, 433–444.
Kartik, N., 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76, 1359–1395.
Khalmetski, K., Sliwka, D., 2019. Disguising lies—image concerns and partial lying in cheating games. Am. Econ. J. Microecon. 11, 79–110.
Konrad, K.A., Lohse, T., Simon, S.A., 2021. Pecunia non olet: on the self-selection into (dis)honest earning opportunities. Exp. Econ. 24, 1105–1130.
Le Maux, B., Masclet, D., Necker, S., 2021. Monetary incentives and the contagion of unethical behavior. SSRN Electron. J.
Makkai, T., Braithwaite, J., 1994. Reintegrative shaming and compliance with regulatory standards. Criminology 32, 361–385.
Perez-Truglia, R., Troiano, U., 2018. Shaming tax delinquents. J. Public Econ. 167, 120–137.
Rauhut, H., 2013. Beliefs about lying and spreading of dishonesty: undetected lies and their constructive and destructive social dynamics in dice experiments. PLoS ONE 8, e77878.
Ruffle, B.J., Tobol, Y., 2014. Honest on Mondays: honesty and the temporal separation between decisions and payoffs. Eur. Econ. Rev. 65, 126–135.
Schwartzstein, J., Sunderam, A., 2021. Using models to persuade. Am. Econ. Rev. 111, 276–323.
Weems, M.L., 1918. Birth and education. In: A History of the Life and Death, Virtues and Exploits of General George Washington. J. B. Lippincott, Philadelphia.
Zakharov, A., 2023. Lying with heterogeneous image concerns. Econ. Lett. 228, 111177.