Signaling motives in lying games

Tilman Fries*

March 29, 2023

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, in particular 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, and the disclosure of lies.

Keywords:. honesty, image concerns, lying, psychological game theory
JEL Codes:. D82, D91

*WZB Berlin Social Science Center, Reichpietschufer 50, D-10785 Berlin; e-mail: tilman.fries@wzb.eu. I am grateful to Johannes Abeler, Kai Barron, Christian Basteck, Daniele Caliari, Martin Dufwenberg, Dirk Engelmann, Hoa Ho, Agne Kajackaite, and Daniel Parra for helpful comments and discussions. I further 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.

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".[1] Religious texts and popular myths often stress the value of honesty.[2] 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 not just to ask if Bob was honest with her in the exchange they just had, but 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 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 that corresponded to 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, Nosenzo, and Raymond, 2019).

One robust feature in experiments that use the F&FH die-roll task is that some individuals lie and dishonestly report four when they could have earned more money by lying and reporting five. 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, Kajackaite, and Sobel (2018) and

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

[2]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 to also be honest.

Khalmetski and Sliwka (2019) provide theoretical models that formalize this intuition.[3] 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 observer, in which an agent draws a state (a number on a die, a coin flip) and makes a report to an observer. The report serves as a signal to the observer, 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 observer'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 observer 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 explain 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.[4] 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 both 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.[5]

I study the strategic implications of individuals signaling their moral character in a lying game. Agents draw a random number (by rolling a die, flipping a coin, etc.) and make a report to an observer. They are morally concerned and incur a cost if their report does not equal their draw. 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 be-

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

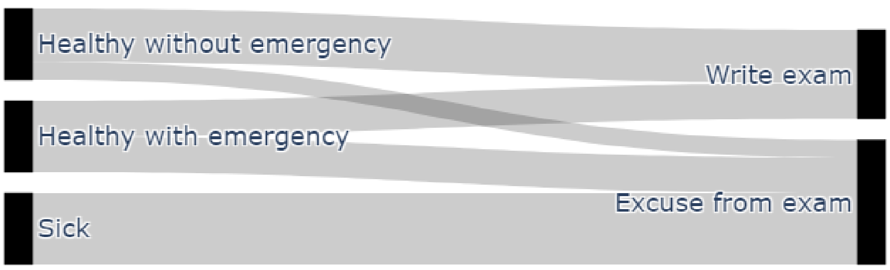
[4]GK&S and K&S introduce the image concern as either the probability to have told the truth, conditional on the report, or as the probability to have lied, conditional on the report. D&D further interact the conditional probability to have lied with the perceived size of the lie. For example, in D&D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

[5]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)).

cause worse moral types are more likely to dishonestly report a high number than better types. In the model, *credibility* of the report and the *honor-stigma gap* between those who do and do not lie influences 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 that 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. Figure 1 sketches out the sorting process from possible states of the worlds into student actions.

Figure 1. Sorting from states of the world into actions



The professor does not observe the real reason of 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 expected 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 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 they draw a number that 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 play symmetric strategies, this is the unique outcome of the game.

I apply the model to study the role of beliefs that agents hold about others and the disclosure of lies. A reoccurring 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.[6] This makes reporting this state reputationally less attractive and in turn reduces lying to that state—I call this the credibility effect. 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

[6]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.

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 drawing 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.

The character-based model can account for a number of 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 draw, 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 draw the lower state in a two-state lying game become less likely to lie when the probability of drawing 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 there are more participants who have the possibility to 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. When asked for their judgement, observers indicate that they would be less likely to lend money to participants who report high payoffs or to employ them. 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 introduces different experimental measures to shift beliefs about lying of others and measures behavior. In one experiment of that sort, AN&R measure behavior in a binary lying game where participants hold different beliefs about the fraction of others reporting the high state. They exogenously shift beliefs of participants using an anchoring technique and find that a smaller proportion of participants whose belief was exogenously increased lies. This effect, though insignificant, goes into the direction predicted by the deed-based model. Results from related experiments typically provide less direct evidence for deed-based models. Experiments reported in (Rauhut, 2013, Diekmann, Przepiorka, and Rauhut, 2015, 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 heterogenous 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 4 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, Masclet, and Necker (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 the only belief-based motive individuals hold.[7]

Fourth, the signaling motives implied by the character-based model can also account for findings from Bicchieri, Dimant, and Sonderegger (2023) who study the role of motivated beliefs in 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 presents the model. Parts 2.1 and 2.2 discuss the setup and equilibrium properties. I apply the model to investigate the determinants of reputation in Section 3. Section 4 applies previous insights to investigate the behavioral effects of interventions that change agents' beliefs and detect liars. Throughout this section, I contrast predictions of the character-based model with predictions from a deed-based model. The paper concludes in Section 5. Proofs of all formal results appear in Appendix A.

2 *Model*

2.1 *Setup*

Game form. Consider a game between a continuum of agents and an observer. Each agent draws a state $j \in \{1, ..., K\}$, which is randomly determined by nature. The agents can be thought to be participants in an economic experiment who are asked by the ex-

[7] Information provision experiments without an active control group provide little experimental control over how treated participants update to information, relative to control (Haaland, Roth, and Wohlfart, 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.

perimenter, who is the observer, to roll a die. In this case, the state would be the outcome of a die roll. An alternative interpretation of the setup could see agents as students who, at the day of an exam, are either sick or healthy and either are in an emergency or not. Throughout this section, we will focus on the first interpretation. In line with the die roll analogy, we make the simplifying assumption that the state is distributed uniformly on its domain.

After the draw, agents each make a report $a \in \mathcal{K} = \{1, ..., K\}$ to the observer and receive a total payoff consisting of direct and image payoffs, as described below. The observer is a passive player with no action whose payoff we do not further specify.

Direct payoffs. Agents know their state j and make a report a, which earns them a direct payoff $y(a)$, where $\Delta(a, a-1) \equiv y(a) - y(a-1) > 0$. The payoff scheme might reflect the experimenter's choice of rewards for reporting certain numbers of the die. Alternatively, the agent-as-student would always earn the highest payoff by claiming to be sick and excusing themselves from the exam.

Reporting $a \neq j$, agents incur cost t which is heterogenous 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, Tanner, and Wagner (2013), Gneezy, Rockenbach, and Serra-Garcia (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 sees individual draws 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.[8] The lying cost is unknown to the observer, who however knows that it is drawn from a distribution $F(t)$ with full support on $(0, \bar{t}]$ and which is independent of j. $\bar{t}$ 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.[9] Further morality dimensions, such as altruism, might be-

[8] I discuss how the model predictions would change in extensions of the model to more complex cost functions in Section 5.

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

come 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 payoffs. 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 prefer 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 for 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 observer. To make an inference, the observer 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* $\mathcal{R}_a^C \equiv \mathbb{E}(t|a)$.

Some 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* $\mathcal{R}_a^D \equiv \mathrm{P}(\textit{honest}|a)$.

The remainder of this section will be concerned with the model with character-based image concerns. The image payoff 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 disproportionally sensitive to changes in the image payoff.[10]

Utility. Direct and image payoffs add up to total payoffs, or utility. An agent of type (j, t) who reports a earns utility

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

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, regardless of the state they draw. One immediate consequence of the assumption is that the observer always puts a positive probability on any state being reported. This property is helpful when solving for the equilibrium, as described next.

[10]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* (Geanakoplos, Pearce, and Stacchetti, 1989, Battigalli and Dufwenberg, 2009), as the final payoffs of agents depend on the observer's beliefs about the agents' moral type. 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 observer 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 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* $a \in \underset{a \in \mathcal{K}}{\arg\max}\; y(a) - 1_{a \neq j} t + \mu \mathbb{E}(t|a)$.
- *Agents and the observer hold the correct equilibrium beliefs*

$$\mathcal{R}_j^C = \frac{\sum_{l \in \mathcal{K}} \int_0^{\bar{t}} s(j|l, t) t f(t)\, \mathrm{d}t}{\sum_{l \in \mathcal{K}} \int_0^{\bar{t}} s(j|l, t) f(t)\, \mathrm{d}t} \text{ for } j \in \mathcal{K}.$$

2.2.1 General results

Based on the definition, we can derive general properties that hold in any equilibrium of the game. Since they should be familiar to readers familiar with the literature, I relegate a formal discussion of them to the Appendix and give intuitions below.

Proposition 1. *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 than K.*

Point (i) says that, when ignoring type-dependent lying costs, every state that is dishonestly reported by some liars yields the same payoff. It can be seen by arguing by contradiction; if there were a state that paid a higher material plus reputational payoff than any other state, then agents would only dishonestly report that state. Point (ii) follows by a similar logic, stating that if someone lies to report k, then no agent lies after drawing k and that if someone lies after drawing j then no agent lies to report j. These results are driven by the assumptions that lying costs are fixed and that the image payoff weight μ is the same for every agent. Both assumptions taken together imply that liars have the same preferences over reporting any state after incurring the type-dependent lying cost. If multiple states are reported dishonestly, liars have to be indifferent between reporting any of them. Section 5 discusses how the model predictions change in an extension that allows for heterogenous image concerns.[11]

Point (iii) seems natural but contains a somewhat deeper point that is worth highlighting. Low moral types are more likely to lie than high moral types. Because of this, reporting a state that is reported by liars in equilibrium decreases the observer's prior belief in expectation. Reporting other states conversely increases the observer's prior. Suppose an equilibrium exists where no liar reports state K. Then, the observer would have to increase her prior expectation after hearing a report of K. Liars in such an equilibrium would always prefer to report K over any other state to gain the highest possible direct payoff and to simultaneously increase the observer's prior expectation. This contradicts utility maximization. A symmetric argument can be made to show that no liar will ever report 1.

The last point (iv) follows because, with image concerns, liars trade off direct payoffs with image payoffs. The image payoffs of states gets spoiled by the liars reporting it. It is therefore beneficial for liars to report more than one state to "smooth out" the image losses they create over multiple states.

2.2.2 *Equilibrium refinement, existence, and characterization*

The predictions above can be useful, but they are also relatively unspecific. One reason is that the equilibrium definition allows for a very rich variety of strategies that liars can play, some of which that 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

[11] Homogenous image concerns are commonly assumed in the literature. They offer tractability. In an experiment, Friedrichsen and Engelmann (2018) provide evidence for heterogenous image concerns, though less is known about whether participants take into account heterogenous image concerns in others.

moral type; that is, the liars with the highest intrinsic type report 2 while those with the lowest intrinsic 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.[12]

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* $t, t' \in (0, \bar{t}]$, $j, j' \in \mathcal{K} \setminus \{k\}$.

Lying strategies are symmetric when the agents' type (j, t) determines whether they lie or not, but does not determine which state they report. Similar properties are imposed by D&D ("uniform cheating") to obtain their main result and by K&S to generate comparative statics predictions. Symmetric lying strategies imply that liars randomize in the same way which state 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 participate 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.[13]

Solving the model under symmetric strategies gives the main result.

Proposition 2. *There exists a unique equilibrium when agents play symmetric lying strategies. It has the following properties:*

(i) The report distribution is strictly increasing in j.

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

(iii) No agent who draws j *reports a state* $k < j$.

(iv) $s(a \neq j|j, t) > 0$ *only if* $j \leq k^*$*, where* $k^* \in \mathcal{K} \setminus \{K\}$.

The equilibrium of the game is of the following type: Agents lie only if they draw a state smaller or equal than some threshold state k^*. If they lie, they report a state larger than k^*. State K is reported by most agents, followed by $K - 1$, and so on. In what follows, I discuss the equilibrium properties and provide a sketch of the proof. I relegate the technical details to Appendix A.

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

[13] 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.

Equilibrium properties. 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 will either report the state that they drew or one of the high states. Since liars are indifferent between reporting any of the high states in equilibrium, the decision problem becomes binary: agents will prefer lying over telling the truth if and only if they prefer reporting K over the state that they drew. Conditional on lying, they will randomize over reporting any of the high states. Therefore, an agent of type (j, t) will lie if and only if $t \leq \hat{t}_j$ for a cutoff $\hat{t}_j$ that, if interior, solves

$$y(K) - \hat{t}_j + \mu \mathcal{R}_K^C = y(j) + \mu \mathcal{R}_j^C.$$ [14]

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

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

The first term is larger than the second, which reflects that liars are stigmatized while truth-tellers are honored. It follows that the reputation of a low state j is equal to $\mathcal{M}^+(\hat{t}_j)$ and a fraction $F(\hat{t}_j)$ of agents who draw that state are liars. We collect all cutoffs $\hat{t}_j$ of each state in a vector $\hat{\boldsymbol{t}}$ and define the *expected moral type of liars* by

$$\mathcal{L}(\hat{\boldsymbol{t}}) \equiv \sum_{j \in \mathcal{K}} \mathrm{P}(\text{draw } j|\text{lie}) \mathcal{M}^-(\hat{t}_j), \text{ with } \mathrm{P}(\text{draw } j|\text{lie}) = \frac{F(\hat{t}_j)}{\sum_{k \in \mathcal{K}} F(\hat{t}_k)}. \tag{1}$$

The probability of a liar reporting j (conditional on lying) is α_j, with corresponding vector $\boldsymbol{\alpha}$. Any high state is reported honestly by a fraction $1/K$ of all agents and by a fraction of $1/K \times \alpha_j \sum_{j \in K} F(\hat{t}_j)$ liars. The probability that a randomly chosen agent reporting j is telling the truth is

$$r_j \equiv \mathrm{P}(\text{truth}|\text{report } j) = \frac{1}{1 + \alpha_j \sum_{j \in K} F(\hat{t}_j)}.$$

The reputation of any high state becomes a weighted average between the expected type of truth-tellers (which equals the prior) and the expected type of liars;

$$\mathcal{R}_j^C = r_j \mathbb{E}(t) + (1 - r_j) \mathcal{L}(\hat{\boldsymbol{t}}) \text{ if } j \in \mathcal{H}. \tag{2}$$

The above expression is smaller than the observer's prior expectation $\mathbb{E}(t)$ as it is a con-

[14]The assumption that $\bar{t} > \Delta(K, 1) + \mu \mathbb{E}(t)$ ensures that the l.h.s. will be 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 $t = 0$, no agent is going to lie after drawing j. These are the high states in which $\hat{t}_j = 0$.

vex combination of $\mathbb{E}(t)$ and $\mathcal{L}(\hat{\boldsymbol{t}})$. We have seen before that the reputation of low states is higher than the prior expectation. One immediate consequence is that there is no downwards lying in equilibrium (part (iii) of Proposition 2): Reporting a state that pays less than the initial draw would imply a lower image payoff and a lower direct payoff, which is inconsistent with utility maximization.[15] Part (iv) of the proposition is a direct implication of part (iii).

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 direct payoff of a state increases because, as the direct 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 direct payoffs for image payoffs. As the direct payoff of a high state decreases, its reputation has to increase to insure that liars are indifferent among high states.

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, (i) is a consequence of (ii).

Existence. Constructing the equilibrium is seemingly complicated because it involves a threshold state k^*, a vector of cutoff types $\hat{\boldsymbol{t}}$, and a vector of probabilities $\boldsymbol{r} = (r_1, ..., r_K)$ that each depend on one another. The key step in the proof is to realize that we can fix the reputation of state K, which is always reported dishonestly in equilibrium, at some level φ. We can then define a function which implicitly defines threshold types *as a function of* φ;

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

Since this equation is strictly decreasing in t there is always a unique solution for a given φ, which we denote as $\tilde{t}_j(\varphi)$. The threshold type of state j is $\hat{t}_j(\varphi) = \max\left\{\tilde{t}_j(\varphi), 0\right\}$.

Aggregating thresholds in a function

$$S(\varphi) \equiv \frac{1}{K}\sum_{j\in\mathcal{K}} F(\hat{t}_j(\varphi))$$

gives the fraction of agents that are willing to lie if the reputation of state K is φ. This function can be thought of as characterizing the supply of lies.

Plugging the threshold functions into (1), the expected moral type of liars inderectly

[15] Note the role of the symmetry refinement here. Many of the counterintuitive equilibria without symmetry emerge because without symmetry $\mathcal{R}_j^C > \mathbb{E}(t)$ is possible for some (but not all) states $j \in \mathcal{H}$.

depends on φ through $\boldsymbol{t}$;

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

We can then observe that the equilibrium reputation of state K, φ^*, must be between $\mathcal{L}(\boldsymbol{t}(\varphi^*))$ and $\mathbb{E}(t)$, as the reputation is a weighted average of the expected moral type of the liars and truth-tellers reporting it. We can use the expected moral type of liars and Equation (2) to write r_K as a function of φ. Liars are indifferent between all high states, which allows us to derive a function $r_j(\varphi)$ for all remaining states $j \in \mathcal{H}$.

Transforming $r_j(\varphi)$ to a likelihood ratio $lr_j(\varphi) \equiv \frac{1-r_j(\varphi)}{r_j(\varphi)}$ gives the ratio of liars to non-liars reporting j if the reputation of the highest state is φ. Adding up the likelihood ratios and normalizing by $1/K$, we arrive at a function

$$D(\varphi) \equiv \frac{1}{K} \sum_{j \in \mathcal{H}} lr_j(\varphi).$$

This function returns the proportion of agents who lie as a function of φ. It can be interpreted as a demand function, as it gives the fraction of liars that are needed to sustain an equilibrium for a given reputation of the highest state.

Figure 2 illustrates the functions S and D. The upper panel shows the individual threshold functions and the aggregate supply function. An increase in the reputation of state K makes it more attractive for agents to lie, which is why the function slopes upwards. The lower panel shows the demand side. These functions slope downwards. Intuitively, when φ is low, many liars will report states different from K to alleviate reputational losses. However, such behavior requires that a high proportion of agents lies to sustain the indifference conditions. Conversely, as φ approaches $\mathbb{E}(t)$, every liar will report K, which is only possible if a small proportion of agents lies.

Supply and demand have to coincide in equilibrium, which determines φ^*, which in turn pins down the equilibrium $\hat{\boldsymbol{t}}^*$, $\boldsymbol{r}^*$, and k^*. The conditions imposed on $F(t)$ ensure existence.

Equilibrium behavior is shaped by signaling motives and a number of insights follow:

Reporting in equilibrium. The equilibrium predicts a report density that is increasing in the direct payoff. We would obtain the same prediction from a model with only intrinsic lying costs. The key difference between both models is however, that, when they are image concerned, some agents might lie and report a non-payoff maximizing state. For example, the model can match the empirical findings from die roll games in laboratory experiments quite well. Figure 3 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.

Figure 2. Equilibrium

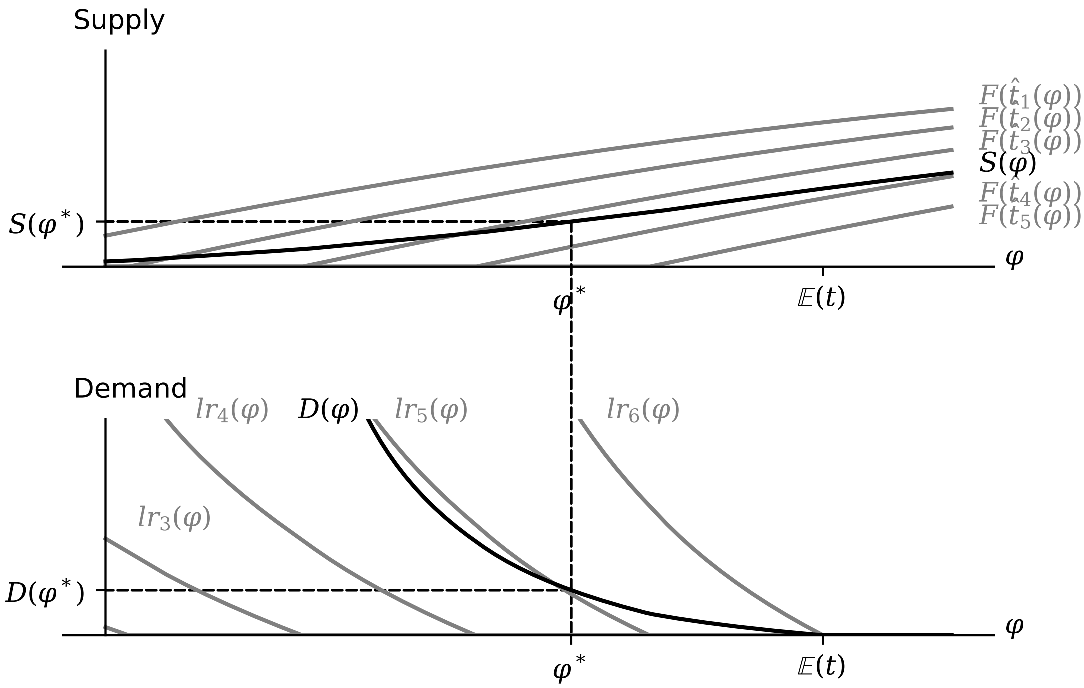
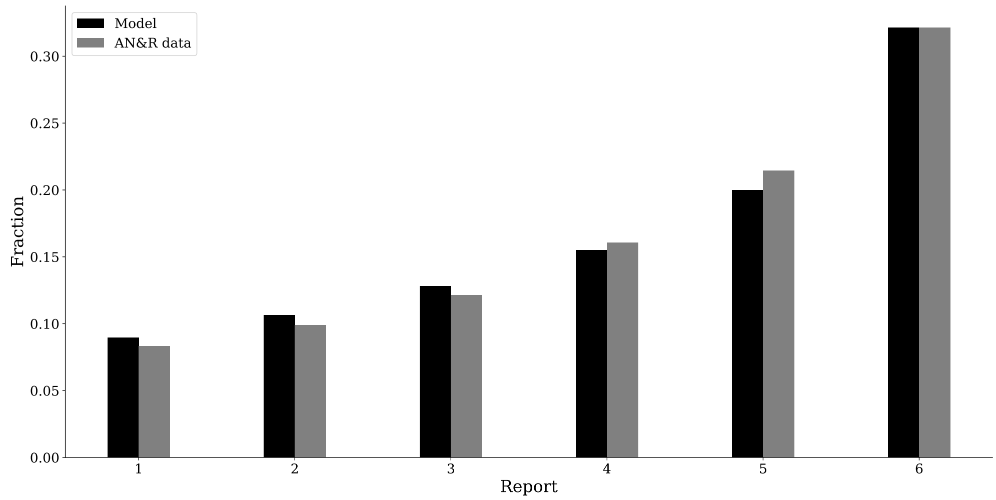


Figure 3. Example equilibrium report distribution compared to the AN&R data



Note: Example equilibrium distribution of reports when lying costs follow a log-normal distribution where log-costs have mean zero and standard deviation 1.1, for values $y(a) - y(a-1) = 1$, and $\mu = 2.1$.

Freeriding on reputation. Liars report a state different from the highest state only if they get a higher image payoff in return. In equilibrium, honest agents and liars pool, and liars free-ride on the honest agents' reputation. One necessary condition for this image enhancing effect is that every state which does not maximize direct payoffs, and which is reported by a liar, is also reported honestly by some agents. This also means that partial lying is sensitive to the underlying distribution of draws. In an extreme case where, e.g., the second-highest state has an initial drawing probability of zero, the model predicts that no agent would lie partially to report that state.

Image spoiling mechanisms. States that are reported dishonestly suffer a reputational penalty because of two factors; credibility and composition. If a state is reported by many liars, then any single agent reporting that state does not appear to have credibly done so truthfully. Since liars are of a worse moral type than the average agent, the reputation of a state suffers when more liars are reporting it. In addition, the reputation of a state also depends on the kind of liars that are reporting it. Liars are of higher reputation if they have a relatively high moral type. That is, they only lie if there are substantial utility gains from lying that give them good reasons to lie because the alternative would have been worse. The marginal liar in state j, who is of type $\hat{t}_j$, always has a higher reputation than the inframarginal liars, who are of expected type $\mathcal{M}^-(\hat{t}_j)$. This implies that the expected type of liars increases in their proportion. In the limit as $\hat{t}_j \to \bar{t}$, then $\mathcal{M}^-(\hat{t}_j) \to \mathbb{E}(t)$; there is no stigma associated with liars from state j. This reflects that bad behavior can be normalized because "everybody is doing it". If almost all agents are committing the bad deed, then doing so oneself is no longer a sign of low character, but merely a signal of mediocrity.

3 *Determinants of image: credibility and the honor-stigma gap*

Let us in this section delve deeper into the determinants of image. 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. Similar results are later derived for $K > 2$ states. From Proposition 2, we know that with two states there is an equilibrium in which agents always tell the truth after drawing 2 and where some agents lie after drawing 1.[16] Lying brings a direct gain $\Delta(2, 1)$ at a cost of t. In equilibrium, a fraction $F(\hat{t})$ lies after drawing 1. The probability that an agent reporting 2 is truth-telling is $r(\hat{t}) = 1/(1 + F(\hat{t}))$. Reporting 2 over 1 comes with a reputational

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

penalty of size

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

I will refer to this function as the *stigma function*. After a bit of algebra, we see that it can be equivalently formulated as

$$\Psi(t) = 2 \times (1 - r(t))(\mathcal{M}^{+}(t) - \mathcal{M}^{-}(t)).^{17} \tag{3}$$

This formulation tells us that the stigma associated with the high report is proportional to the product of two terms. The first term, $1 - r(t)$, 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 second term, $\mathcal{M}^{+}(t) - \mathcal{M}^{-}(t)$, denotes the difference in moral character of liars and non-liars among those agents who drew 1, i.e., among those agents that could lie to increase their direct utility. 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 $\hat{t}$ who is exactly indifferent between lying and truth-telling;

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

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

$$\lim_{t \to 0} \Psi(t) = 2 \times (1 - r(0))(\mathcal{M}^{+}(0) - \mathcal{M}^{-}(0)) = 2 \times 0 \times (\mathcal{M}^{+}(0) - \mathcal{M}^{-}(0)) = 0.$$

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

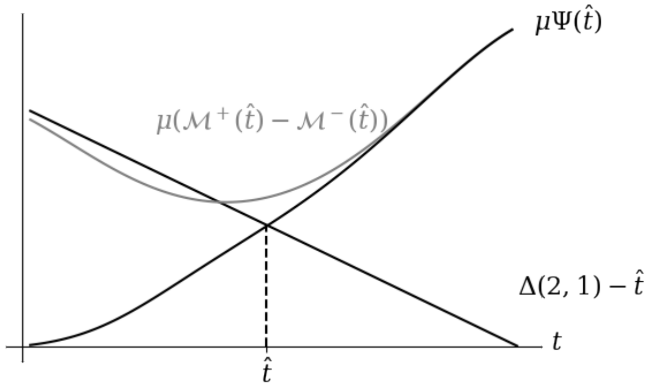
$$\Psi'(t) = 2\left[(1 - r(t)) \underbrace{(\mathcal{M}^{+\prime}(t) - \mathcal{M}^{-\prime}(t))}_{\substack{\text{Honor-stigma} \\ \text{effect } (\lesseqgtr 0)}} \underbrace{-r'(t)}_{\substack{\text{Credibility} \\ \text{effect } (> 0)}} (\mathcal{M}^{+}(t) - \mathcal{M}^{-}(t))\right].$$

More agents reporting 2 makes it less credible that anyone reporting 2 is truth-telling. The

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

$$\begin{aligned}
\Psi(t) &= \mathcal{M}^{+}(t) - r(t)(F(t)\mathcal{M}^{-}(t) + (1 - F(t))\mathcal{M}^{+}(t)) - (1 - r(t))\mathcal{M}^{-}(t) \\
&= \left(\frac{1 + F(t)}{1 + F(t)} - \frac{1 - F(t)}{1 + F(t)}\right)\mathcal{M}^{+}(t) - 2(1 - r(t))\mathcal{M}^{-}(t) \\
&= 2(1 - r(t))(\mathcal{M}^{+}(t) - \mathcal{M}^{-}(t)).
\end{aligned}$$

Figure 4. Equilibrium for $K = 2$



credibility effect leads to an increase in the stigma of reporting 2 after a marginal increase in t. In addition, the types of those who lie and those who tell the truth changes. The sign of this additional honor-stigma effect is ambiguous. The following result however shows that, independently of the sign of the honor-stigma effect and for a relatively broad class of distribution functions, the stigma function strictly increases with t.[18] This implies that, with only two states, 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 lying of one agent crowds out lying of other agents. An equilibrium obtains where the stigma function crosses the direct payoff as displayed in Figure 4.

Proposition 3a. *Suppose that $f(t)$ is strictly decreasing or log-concave. The stigma function $\Psi(t)$ is increasing. Lies are strategic substitutes.*

Relation to the deed-based model. 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 t are therefore relatively straightforward; as t increases, reporting 2 becomes less credible. As in the character-based model, lies in the deed-based model are thus strategic substitutes. The next section will explore cases where, due to the character-based model's honor-stigma effect, the qualitative predictions of the character- and deed-based model disagree.

The role of non-observability. That the stigma function increases with t is 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

[18] The result holds for all strictly decreasing distribution functions and for the family of log-concave distributions (e.g., the (truncated) normal, exponential, or uniform distributions). Log-concavity is a very common assumption in the signaling literature and the mathematical properties of log-concave distributions are well understood (see Bagnoli and Bergstrom, 2005, for an overview).

$\mathcal{M}^+(\hat{t}) - \mathcal{M}^-(\hat{t})$.[19] In case of a single-peaked type distribution, this difference is decreasing for small t and increasing for larger 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 Benabou2006 honor-stigma model gets weighted by the probability that a report of 2 is a lie, as displayed in Equation (3). This which reflects the uncertainty about the draw 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 another way, truth-telling reputationally only pays off if the observer expects many agents to lie. Figure 4 contrasts signaling incentives in a non-observed lying game with signaling incentives in an game where the observer 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.

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

	Character-based	Deed-based
Non-observed	$\Psi(t) = 2(1-r(t))(\mathcal{M}^+(t) - \mathcal{M}^-(t))$ $\Psi'(t) = 2\Big[(1-r(t))(\mathcal{M}^{+\prime}(t) - \mathcal{M}^{-\prime}(t)) - r'(t)(\mathcal{M}^+(t) - \mathcal{M}^-(t))\Big]$	$\Psi(t) = 1 - r(t)$ $\Psi'(t) = -r'(t)$
Observed	$\Psi(t) = \mathcal{M}^+(t) - \mathcal{M}^-(t)$ $\Psi'(t) = \mathcal{M}^{+\prime}(t) - \mathcal{M}^{-\prime}(t)$	$\Psi(t) = 1$ $\Psi'(t) = 0$

Table 1 summarizes the stigma functions of models with character and deed based image concerns, for cases where lies either non-observed or observed.[20] This paper is mostly concerned with the characer-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.,

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

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

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 which 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.

3.1 *More than two states*

We extend the analysis to more than two states. In this case, we can ask how a marginal increase in $\hat{t}_k$ changes the behavior of agents who draw a state j different from k. Other states are affected by increases in $\hat{t}_k$ because such increases have an impact on the image payoff from lying. If the reputation of the highest state increases in response to an increase in $\hat{t}_k$, then agents from other states will be encouraged to lie. Otherwise, they will be discouraged. The formal results from the two-state case quite naturally extend:

Proposition 3b. *With $K > 2$, lies are strategic substitutes with respect to the kth state if and only if*

$$(1 - \tilde{r}(\hat{\boldsymbol{t}}^*))\frac{\partial \mathcal{L}}{\partial \hat{t}_k} + \frac{\partial \tilde{r}}{\partial \hat{t}_k}(\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}^*)) < 0,$$

where

$$\tilde{r}(\hat{\boldsymbol{t}}) \equiv \frac{1}{K}\frac{1}{\sum_{j\in\mathcal{H}} \alpha_j^* \mathrm{P}(\textit{report } j)}.$$

The proposition shows that, similarly to the two state case, credibility and honor-stigma effects guide behavior when there are more than two states: the first term in the inequality above denotes the change in the reputation of liars caused by an increase in $\hat{t}_k$ while the second term denotes the change in that a high report is being made truthfully. To better see this, consider the expected credibility of a liar's report. In an equilibrium with partial lying, liars play mixed strategies, where they, conditional on lying, report a state j with probability α_j. Bayes' rule tells us that the credibility of any report of a high state is $P(\text{truth}|j) = 1/K \times 1/\mathrm{P}(\text{report } j)$. Therefore, the expected credibility a liar will get is

$$\mathbb{E}_{\alpha_j}(\mathrm{P}(\text{truth}|\text{report } j)) = \frac{1}{K}\sum_{j\in\mathcal{H}} \alpha_j^* \frac{1}{P(\text{report } j)},$$

which is approximately equal to the $\tilde{r}(\hat{\boldsymbol{t}})$ in the proposition.[21]

Differently from the two-state case, strategic complementarities can obtain. An increase in $\hat{t}_1$ will always lead to an increase in $\mathcal{L}(\boldsymbol{t})$, i.e., a positive honor-stigma effect;

[21] In an equilibrium without partial lying both terms exactly coincide since $\alpha_K = 1$. The reason that they do not coincide when there is partial lying is that the α_js themselves change after an exogenous increase in $\hat{t}_k$.

if this positive effect dominates the credibility effect, strategic complementarities obtain. The honor-stigma effect is smaller for larger k and can be negative if $t_k > \mathcal{L}(\hat{\boldsymbol{t}})$, e.g., after an increase in $\hat{t}_{k^*}$. Intuitively, only the lowest moral types would lie after drawing k^* while also higher types would lie after drawing 1. Liars from k^* are relatively cheap; they lie for a smaller utility gain than liars from 1, which makes them the liars with the lowest average reputation.[22] Another element that contributes to strategic complementarity is the baseline lying rate. If lying already is on a high level (so that $1 - \tilde{r}(\hat{\boldsymbol{t}}^*)$ is large) the honor-stigma effect gains in weight and actions are more likely strategic complements.

4 *Applications*

This section considers two applications of the model. We first consider the role of beliefs about others on behavior. Thereafter, we will turn to the effects of different forms of lie detection and disclosure.

4.1 *Changing beliefs*

Agents in the model prefer appearing as a high type over appearing as a low type. The reputational stigma of making a dishonest report closely depends on the distribution of types and on beliefs that agents and the observer hold about it. This part asks how a change in beliefs about the type distribution affects behavior.

One interpretation of the following comparative statics is to think of moving a single agent of a given type from a population with a certain preference distribution to a population with a different preference distribution and asking how the 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 observer's belief about the type distribution. Seen in this light, a comparative static that shifts a feature of the preference distribution can be more literally interpreted as a shift in the agent's second order belief. Such shifts might occur after a norms-based interventions which aims to correct agent's misperceptions about average behavior (Bénabou and Tirole, 2011). Alternatively, following Bénabou, Falk, and Tirole (2020),[23] shifts in agents' second-order beliefs could

[22] The fact that "small" lies are more severely stigmatized than "large" lies would be more ambiguous in a setting where agents' lying decisions have direct 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 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, Maréchal, Tannenbaum, and Zünd (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.)

[23] 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 which shift agents' beliefs about the type distri-

be brought about by third parties who persuade agents to hold a certain belief about the preference distribution by using narratives. I will use the narrative analogy in the first two comparative statics I will discuss. An additional interpretation which will be provided for the last comparative static in this section is that myopic agents and an observer repeatedly interact, with the observer's prior about agents' characters becoming more precise over time.

When studying how changes in beliefs affect behavior I focus on the two state case. This section will also assume that the conditions of Proposition 3a hold so that $\Psi(t)$ is increasing.

A "nobody is perfect" narrative. Consider agents who are 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 leading to a shift in the belief about the prior type distribution function, redistributing some probability mass out of the right tail of the distribution (i.e., the part where the highest moral types are located) to the left tail. One way to model this is by reducing the highest type $\bar{t}$ to a lower one, $\tau < \bar{t}$ (while maintaining the assumption that $\tau > \Delta(2,1) + \mu\mathbb{E}(t|t \leq \tau)$ to focus on interior equilibria).[24]

A reduction in $\bar{t}$ will affect the marginal type who is indifferent between truth-telling and lying through the stigma function $\Psi_{\bar{t}}(t)$. We will denote the stigma function under the new, right-truncated distribution (see the left panel in Figure 5) as

$$\Psi_\tau(\hat{t}) = 2 \times (1 - r_\tau(\hat{t}))(\mathcal{M}_\tau^+(\hat{t}) - \mathcal{M}_\tau^-(\hat{t})).$$

In the equation above, the τ subscripts indicate the new truncation point.[25] To investigate the effect of truncating the belief distribution, consider the derivative of $\Psi_\tau(\hat{t})$ with respect to τ;

$$\frac{\partial \Psi_\tau(\hat{t})}{\partial \tau} = 2 \times \left[(1 - r_\tau(\hat{t})) \underbrace{\frac{\partial \mathcal{M}_\tau^+(\hat{t})}{\partial \tau}}_{\substack{\text{Honor-stigma} \\ \text{effect } (>0)}} \underbrace{-\frac{\partial r_\tau}{\partial \tau}}_{\substack{\text{Credibility} \\ \text{effect } (<0)}} (\mathcal{M}_\tau^+(\hat{t}) - \mathcal{M}_\tau^-(\hat{t})) \right].$$

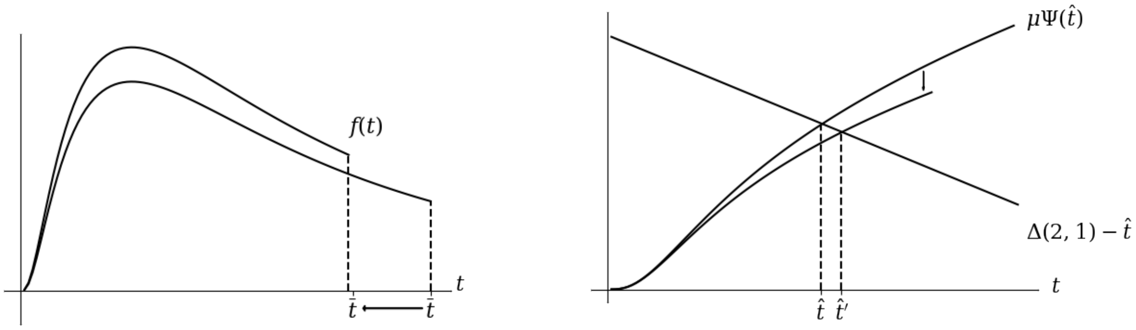
Again we see by now familiar credibility and honor-stigma effects. Decreasing $\bar{t}$ to τ decreases the stigma of lying decreases through the honor-stigma effect because, after

bution. The paper by Bénabou et al. (2020) is part of a emerging recent literature that investigate 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 (2021) fully specifies the equilibrium of a game where agents express potentially controversial opinions and tailor interpretations of past data to increase their reputational utility in front of others.

[24]If the reduction in $\bar{t}$ is larger, agents might expect everyone to lie after the reduction. Psychological versions of the intuitive criterion exist which could be used to pin down off-equilibrium beliefs in such a situation (Bernheim, 1994, Dufwenberg and Lundholm, 2001).

[25]That is, $\mathcal{M}_\tau^+(\hat{t}) = E(t|t \in (\hat{t}, \tau))$, $r_\tau(\hat{t}) = F(\tau)/(F(\tau) - F(\hat{t}))$. Note that $\mathcal{M}_\tau^-(\hat{t}) = \mathcal{M}^-(\hat{t})$ as long as $\hat{t} \leq \tau$.

Figure 5. Effect of a decrease of $\bar{t}$



knowing that "nobody is perfect", the reputation of truth tellers is stained. At the same time, the reputation of liars remains unaffected, as they are exclusively drawn from the left tail of the distribution. In the present case, the honor-stigma effect always dominates the credibility effect. Therefore, a decrease in $\bar{t}$ shifts the stigma function downwards. As illustrated in the right panel of Figure 5, agents become more likely to lie.

Proposition 4a. *If $f_X(t)$ is a density function with $f_X(t) > 0$ for $t \in (0, \bar{t}]$ and $f_Y(t)$ is a version of $f_X(t)$ truncated at a point τ, where $\Delta(2,1) + \mu\mathbb{E}(t|t \leq \tau) < \tau < \bar{t}$, then, under character-based image concerns, agents with belief $f_Y(t)$ are more likely to lie than agents with belief $f_X(t)$.*

An agent with character-based image concerns who decreases their parameter belief of the top type from $\bar{t}$ to τ becomes more likely to lie. Decreasing $\bar{t}$ also has the effect that the agent now believes that a higher fraction of other agents are lying (since none of them is of a very high type). That is, reporting 2 becomes more suspicious as measured by the credibility effect. Consequently, and in contrast to the model with character-based image concerns, agents with deed-based image concerns become less likely to lie after learning that "nobody is perfect".

Proposition 4a′. *Consider the same comparative static as described in Proposition 4a but now suppose that agents have deed-based image concerns. Then, agents with belief $f_Y(t)$ are less likely to lie than agents with belief $f_X(t)$.*

The predictions of the character- and deed-based model thus disagree. As the next result shows, it is however not generally true that an increase in the belief about the proportion of others lying increases one's own propensity to lie in the character-based model.

Shifting the preference distribution to the right. Consider shifting the whole belief density function $f_X(t)$ to the right using a positive parameter a so that $f_Y(t) = f_X(t-a)$. Such a shift results in agents believing that other agents are more honest as probability mass is shifted away from low moral types towards high types. Differently from the last comparative static, this shift happens for the whole distribution. Agents therefore come to believe that other agents are less likely to lie than before, making a high report more credible. This credibility effect dominates in the present case, so that agents become more likely to lie.

Proposition 4b. *If $f_X(t)$ is a density function with $f_X(t) > 0$ for $t \in (0, \bar{t}]$ and $f_Y(t)$ is a version of $f_X(t)$ with $f_Y(t) = f_X(t-a)$ where $a > 0$ then, under character-based image concerns, agents with belief $f_Y(t)$ are more likely to lie than agents with belief $f_X(t)$.*

As the credibility effect dominates, the predictions of the character- and deed-based models agree in this case.

Proposition 4b'. *Consider the same comparative static as described in Proposition 4b but now suppose that agents have deed-based image concerns. Then, agents with belief $f_Y(t)$ are more likely to lie than agents with belief $f_X(t)$.*

Even though the belief changes examined in the previous two comparative statics have seemingly opposite consequences–under the "nobody is perfect"-narrative agents come to believe that more other agents will lie while when shifting the preference distribution to the right agents come to believe that fewer other agents will lie–both comparative static results for the character-based model predict that agents themselves become more likely to lie after shifting their belief. This happens because the honor-stigma effect dominates in the first comparative static comparison while the credibility effect dominates in the second comparative static comparison.

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, it sometimes 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 reduces 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 making 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. Consider an observer who knows the past history of agents' actions, which she can use to reduce her prior uncertainty about the agents' types. How do agents adjust their behavior to the observer's new beliefs? To study the role of changing uncertainty about types, we will compare behavior under two type distributions that can be ordered according to the Unimodal Likelihood Ratio order, which was introduced by Ramos, Ollero, and Sordo (2000):

Definition 5. *Two distributions $F_X(t)$, $F_Y(t)$ satisfy the Unimodal Likelihood Ratio (ULR) order if the likelihood ratio $f_X(t)/f_Y(t)$ is unimodal and $\mathbb{E}_X(t) \geq \mathbb{E}_Y(t)$.*

The ULR order is a measure of the relative dispersion of probability distributions: The results of Ramos et al. (2000) imply that, if two distributions satisfy ULR and have the same mean, then $F_Y(t)$ is a mean-preserving spread of $F_X(t)$.[26] Comparing behavior under different type distributions which can be ordered according to ULR and whose mean coincides therefore addresses the question of changing type uncertainty. To anticipate the intuition behind the following comparative static, think about adding noise to an initial type distribution. The resulting more dispersed distribution will have fatter left and right tails than the initial distribution. As a consequence, the conditional expectations $\mathcal{M}^+(t)$ and $\mathcal{M}^-(t)$ will take on more extreme values under the more dispersed distribution. This in turn increases the stigma of reporting 2 instead of 1 for a given threshold type, which leads to the following result.

Proposition 4c. *Suppose the distributions $F_X(t)$ and $F_Y(t)$ satisfy the ULR order, that $\mathbb{E}_X(t) = \mathbb{E}_Y(t)$, and that both densities have full support on $(0, \bar{t}]$. Then agents with belief $f_Y(t)$ are less likely to lie than agents with belief $f_X(t)$.*

In the character-based model, agents want to convince the observer that they are of a high type. As the observer's prior becomes more certain, agents have less room to move the observer's prior by taking any particular action. Their actions in turn are less guided by image concerns, which makes them more likely 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 4c'. *Consider the same comparative static as described in Proposition 4c but now suppose that agents have deed-based image concerns. Then, there exists a unique critical value $\tilde{t} \in (0, \bar{t})$.*

(i) If, under the belief $f_X(t)$, the equilibrium threshold $\hat{t} < \tilde{t}$, agents with belief $f_Y(t)$ are less likely to lie than agents with belief $f_X(t)$.

(ii) If, under the belief $f_X(t)$, the equilibrium threshold $\hat{t} = \tilde{t}$, agents with belief $f_Y(t)$ are equally likely to lie than agents with belief $f_X(t)$.

(iii) If, under the belief $f_X(t)$, the equilibrium threshold $\hat{t} > \tilde{t}$, agents with belief $f_Y(t)$ are more likely to lie than agents with belief $f_X(t)$

The deed-based model does not make a clear prediction about how changing uncertainty about preferences affects behavior. Thus, signaling incentives in the deed-based model do not necessarily become weaker as the observer learns about agents from

[26]The ULR is a sufficient condition for second-order stochastic dominance that is fulfilled by many families of distributions. For example, within the family of normal or lognormal distributions, ULR and second-order stochastic dominance are equivalent conditions (Hopkins and Kornienko, 2007).

their past actions. There is an interesting connection between this last 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 personal traits of the transgressor based on the transgression (what we may call deed-based). Shaming is disintegrative if it generalizes from transgressions to personal traits of the transgressor (what we may call character-based). In his 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, decreases as observers form more precise priors.[27]

4.2 *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 who maintain publicly accessible websites which 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 original draw of any agent with probability π and discloses lies to the observer. In its most basic form, the investigator could rely on *coarse disclosure* that discloses lies, but not the original draw of the liar. Such a regime results in an image of $\mathcal{L}(\hat{\boldsymbol{t}})$ for a disclosed liar. The expected reputation of a liar reporting K then becomes

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

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. It also makes partial lying less attractive as partial liars are as likely as full liars to be

[27] 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 that it would be wrong to think of the experimenter as an observer who updates her belief after every single report.

caught lying, so that the reputational advantage of partial over full-extent lying becomes smaller.[28]

Proposition 4da. *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.

Assume for the rest of the section that the prior type distribution is uniform.[29] Since the investigator observes the state originally drawn by the liar, they could additionally commit in advance to disclosing it with some probability γ. 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. Liars from the lowest states, with a moral type larger than the average liar, will prefer contextualized disclosure to coarse disclosure. They become more likely to lie. However, liars from higher states, who have a type smaller than the average liar, will dislike contextualized disclosure as they can no longer benefit as much from being pooled with the liars from the lowest states upon detection. 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 draws 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 conditionally on being disclosed as a liar already under coarse disclosure it is likely that they drew a low state. 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 will have 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 4db. *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 initial draw γ*

(i) The average size of lies increases.

(ii) The lying rate decreases.

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

[28] 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.

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

K becomes

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

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

Proposition 4d'. *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 initial draw γ.

5 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 useful in applications to the behavioral effects of norm interventions or narratives, make predictions about the short- and long-term effects of different shaming conventions, and have implications on how lies should be disclosed.

5.1 Extensions

Two simplifying assumptions were maintained throughout the analysis; that intrinsic lying costs are fixed and that agents care to the same extent about the image payoff. 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 a 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 draw. 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. When different agents care about their image to different extents, partial lying can still emerge as part of an equilibrium but it will be of a slightly different kind. Remember how in the baseline analysis, liars are indifferent between 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 a higher image payoff 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 $K-1$. Heterogenous image concerns can also lead to downward lying. Since an highly image concerned agent will prefer honestly reporting $K-1$ over honestly reporting K, they might also prefer dishonestly reporting $K-1$ after drawing K if their intrinsic lying cost is sufficiently low. With heterogenous image concerns, the character-based model can also account for experimental results that document that, in some cases, a report distribution where the mode is smaller than K. 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 direct payoffs and lower the observer's suspicion. The motivation for agents with character-based image concerns to nonetheless report the mode purely follows from the honor-stigma motive.

5.2 *Going forward*

In addition to offering new theoretical insights, the model also generates a number of 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. For example, to measure a preference for appearing honestly, researchers could elicit the willingness to pay for participating in a lying game. The character-based model would predict that participants are willing to pay a premium to not to participate in the lying game to signal their honesty. An alternative question regards the heterogeneity of the image concern. Since such heterogeneity leads to downward lying, trying to find ways to (non-deceptively) study downward lying seems promising.[30] One approach could try to create a situation where the experimenter knows the individual draw but where participants report to a person who does not know it (e.g. another participant in the experiment). The experimenter would thus be able to collect information on both individual draws and reports while still keeping signaling motives that might motivate some to lie downward active. Finding that participants in this setting lie downward would be a strong indicator of heterogeneous image concerns.

Second, future experiments could not only try to identify preferences but also the strategic reasoning of individuals that 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. 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 through numerous reasons. In the character-based model, in addition to the question on *how many* people lie, questions such as *who* lies and *why* become important. Designs which 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 composition signaling motive.

[30] Experimental evidence on downwards lying so far has only been observed for selected samples and under special design features. In an experiment with a small sample of nuns, Utikal and Fischbacher (2013) find evidence that is consistent with downward lying. Barron (2019) finds systematic evidence that lab participants lie downwards on a small stakes die when they simultaneously have the opportunity to lie upwards on a high stakes die.

References

ABELER, J., D. NOSENZO, AND C. RAYMOND (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153.

ADRIANI, F. AND S. SONDEREGGER (2019): "A Theory of Esteem Based Peer Pressure," *Games and Economic Behavior*, 115, 314–335.

AKERLOF, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, 488–500.

AKIN, Z. (2019): "Dishonesty, Social Information, and Sorting," *Journal of Behavioral and Experimental Economics*, 80, 199–210.

BAGNOLI, M. AND T. BERGSTROM (2005): "Log-Concave Probability and Its Applications," *Economic Theory*, 26, 445–469.

BARRON, K. (2019): "Lying to Appear Honest," *WZB Discussion Paper SP II 2019-307*.

BAŠIĆ, Z. AND S. QUERCIA (2022): "The Influence of Self and Social Image Concerns on Lying," *Games and Economic Behavior*, 133, 162–169.

BATTIGALLI, P. AND M. DUFWENBERG (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1–35.

——— (2022): "Belief-Dependent Motivations and Psychological Game Theory," *Journal of Economic Literature*, 60, 833–882.

BÉNABOU, R., A. FALK, AND J. TIROLE (2020): "Narratives, Imperatives, and Moral Persuasion," *Mimeo*.

BÉNABOU, R. AND J. TIROLE (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

——— (2011): "Laws and Norms," *NBER Working Paper*.

BERNHEIM, B. D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102, 841–877.

BICCHIERI, C., E. DIMANT, AND S. SONDEREGGER (2023): "It's Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following and Belief Distortion," *Games and Economic Behavior*, 138, 321–354.

BRAITHWAITE, J. (1989): *Crime, Shame and Reintegration*, New York: Cambridge University Press.

BUTERA, L., R. METCALFE, W. MORRISON, AND D. TAUBINSKY (2022): "Measuring the Welfare Effects of Shame and Pride," *American Economic Review*, 112, 122–168.

COHN, A., M. A. MARÉCHAL, D. TANNENBAUM, AND C. L. ZÜND (2019): "Civic Honesty around the Globe," *Science*, 365, 70–73.

CRAWFORD, V. P. AND J. SOBEL (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–1451.

DIEKMANN, A., W. PRZEPIORKA, AND H. RAUHUT (2015): "Lifting the Veil of Ignorance: An Experiment on the Contagiousness of Norm Violations," *Rationality and Society*, 27, 309–333.

DUFWENBERG, M. AND M. A. DUFWENBERG (2018): "Lies in Disguise – A Theoretical Analysis of Cheating," *Journal of Economic Theory*, 175, 248–264.

DUFWENBERG, M. AND M. LUNDHOLM (2001): "Social Norms and Moral Hazard," *The Economic Journal*, 506–525.

ELIAZ, K. AND R. SPIEGLER (2020): "A Model of Competing Narratives," *American Economic Review*, 110, 3786–3816.

FEESS, E. AND F. KERZENMACHER (2018): "Lying Opportunities and Incentives to Lie: Reference Dependence versus Reputation," *Games and Economic Behavior*, 111, 274–288.

FISCHBACHER, U. AND F. FÖLLMI-HEUSI (2013): "Lies in Disguise-an Experimental Study on Cheating," *Journal of the European Economic Association*, 11, 525–547.

FOERSTER, M. AND J. J. VAN DER WEELE (2021): "Casting Doubt: Image Concerns and the Communication of Social Impact," *The Economic Journal*, 2887–2919.

FRIEDRICHSEN, J. AND D. ENGELMANN (2018): "Who Cares about Social Image?" *European Economic Review*, 110, 61–77.

GEANAKOPLOS, J., D. PEARCE, AND E. STACCHETTI (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60–79.

GIBSON, R., C. TANNER, AND A. F. WAGNER (2013): "Preferences for Truthfulness: Heterogeneity among and within Individuals," *American Economic Review*, 103, 532–548.

GNEEZY, U., A. KAJACKAITE, AND J. SOBEL (2018): "Lying Aversion and the Size of the Lie," *American Economic Review*, 108, 419–453.

GNEEZY, U., B. ROCKENBACH, AND M. SERRA-GARCIA (2013): "Measuring Lying Aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

GOLMAN, R. (2021): "Acceptable Discourse: Social Norms of Beliefs and Opinions," *Mimeo*.

HAALAND, I., C. ROTH, AND J. WOHLFART (2023): "Designing Information Provision Experiments," *Journal of Economic Literature*, 61, 3–40.

HARBAUGH, R. AND E. RASMUSEN (2018): "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10, 210–235.

HILLENBRAND, A. AND E. VERRINA (2022): "The Asymmetric Effect of Narratives on Prosocial Behavior," *Games and Economic Behavior*, 135, 241–270.

HOPKINS, E. AND T. KORNIENKO (2007): "Cross and Double Cross: Comparative Statics in First Price and All Pay Auctions," *The B.E. Journal of Theoretical Economics*, 7, 000010220219351704l366.

HURSTHOUSE, R. AND G. PETTIGROVE (2018): "Virtue Ethics," in *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta, vol. Winter 2018 Edition.

JEWITT, I. (2004): "Notes on the 'Shape' of Distributions." *Mimeo*.

KAJACKAITE, A. AND U. GNEEZY (2017): "Incentives and Cheating," *Games and Economic Behavior*, 102, 433–444.

KARTIK, N. (2009): "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76, 1359–1395.

KHALMETSKI, K. AND D. SLIWKA (2019): "Disguising Lies—Image Concerns and Partial Lying in Cheating Games," *American Economic Journal: Microeconomics*, 11, 79–110.

LE MAUX, B., D. MASCLET, AND S. NECKER (2021): "Monetary Incentives and the Contagion of Unethical Behavior," *SSRN Electronic Journal*.

MAKKAI, T. AND J. BRAITHWAITE (1994): "Reintegrative Shaming and Compliance with Regulatory Standards," *Criminology*, 32, 361–385.

METZGER, C. AND L. RÜSCHENDORF (1991): "Conditional Variability Ordering of Distributions," *Annals of Operations Research*, 32, 127–140.

PEREZ-TRUGLIA, R. AND U. TROIANO (2018): "Shaming Tax Delinquents," *Journal of Public Economics*, 167, 120–137.

RAMOS, H. M., J. OLLERO, AND M. A. SORDO (2000): "A Sufficient Condition for Generalized Lorenz Order," *Journal of Economic Theory*, 90, 286–292.

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.

SCHWARTZSTEIN, J. AND A. SUNDERAM (2021): "Using Models to Persuade," *American Economic Review*, 111, 276–323.

UTIKAL, V. AND U. FISCHBACHER (2013): "Disadvantageous Lies in Individual Decisions," *Journal of Economic Behavior & Organization*, 85, 108–111.

WEEMS, M. L. (1918): "Birth and Education," in *A History of the Life and Death, Virtues and Exploits of General George Washington*, Philadelphia: J. B. Lippincott.

A Proofs

A.1 Proof of Proposition 1

(i) An agent who draws state j will lie if there is a state k such that

$$y(k) - t + \mu\mathcal{R}_k^C > y(j) + \mu\mathcal{R}_j^C. \tag{4}$$

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, ε), where $\epsilon > 0$ 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 draws 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 draw a state $j \in \mathcal{H}$, as lying is costly and does not lead to higher payoffs.

(iii) Consider again the incentive constraint (4) 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 $\hat{t}_j > 0$ if $j \notin \mathcal{H}$ and $\hat{t}_j = 0$ otherwise. Now consider the reputations that are associated with agents who draw 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

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

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;

$$\mathcal{R}_j^C = \mathcal{M}^+(\hat{t}_j) \text{ if } j \notin \mathcal{H}. \tag{5}$$

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

$$y(j) + \mu\mathcal{R}_j^C > y(K) + \mu\mathcal{R}_K^C, \text{ and } y(K) > y(j). \tag{6}$$

This in particular implies that $\mathcal{R}_j^C > \mathcal{R}_K^C$ for all $j \in \mathcal{H}$. From (5) it follows that $\mathcal{R}_K^C > \mathbb{E}(t)$ and more generally $\mathbb{E}(t|\text{report } j \notin \mathcal{H}) > \mathbb{E}(t)$. By the martingale property of beliefs, it then follows that $\mathbb{E}(t|\text{report } j \in \mathcal{H}) < \mathbb{E}(t)$, which requires that $\mathcal{R}_j^C < \mathbb{E}(t)$ for some

$j \in \mathcal{H}$.[31] Combining the inequalities, we arrrive at $\mathcal{R}_K^C > \mathbb{E}(t) > \mathcal{R}_j^C$ for some $j \in \mathcal{H}$, which is a contradiction to (6).

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

$$y(j) + \mu\mathcal{R}_j^C < y(1) + \mu\mathcal{R}_1^C, \text{ and } y(1) < y(j). \tag{7}$$

This in particular implies that $\mathcal{R}_1^C > \mathcal{R}_j^C$ for all $j \in \mathcal{H}$. 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 $\max\{\mathbb{E}(t), \max\{\hat{\boldsymbol{t}}\}\} < \mathbb{E}(t|t > \max\{\hat{\boldsymbol{t}}\})$. Since $\mathcal{R}_j^C = \mathbb{E}(t|t > \max\{\hat{\boldsymbol{t}}\})$ for some $j \in \mathcal{H}$, we arrive at a contradiction to (7).

(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

$$\mathcal{R}_{K-1}^C - \mathcal{R}_K^C \leq \frac{\Delta(K, K-1)}{\mu}. \tag{8}$$

Since $K-1 \notin \mathcal{H}$, it follows from (5) that $\mathcal{R}_{K-1}^C > \mathbb{E}(t)$. Furthermore, if $\mathcal{H}$ is a singleton then by the martingale property of beliefs, $R_K > \mathbb{E}(t)$. The left-hand side of (8) is strictly positive. Thus, there is a contradiction if $\frac{\Delta(K,K-1)}{\mu}$ is sufficiently small.

A.2 *Proof of Proposition 2*

We first provide two lemmas before proceeding with the proof.

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

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

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

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. □

Lemma 2 (Properties of $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi))$)**.** $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi))$ *is* (i) *a continuous function in* φ *whenever some* $\hat{t}_j > 0$ *with* (ii) $\frac{d\mathcal{L}}{d\varphi} < 1$ *if* μ *is small enough. There exists* (iii) *an interval* $(\varphi^{min}, \mathbb{E}(t))$*, where*

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

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

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

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

(ii) Write the derivative as

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

Lemma 1 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 C shows that $\mu \leq 1$ is sufficient in case $f(t)$ is log-concave.

(iii) Proposition 1 shows that there is always lying from 1, which implies that, in equilibrium, $y(K) + \mu\varphi > y(1) + \mu\mathbb{E}(t)$ and therefore in any equilibrium $\varphi > \max\{0, \mathbb{E}(t) - \Delta(K,1)/\mu\}$. If $\mathbb{E}(t) > \Delta(K,1)/\mu$ then it follows that $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi^{\min} + \varepsilon)) < \varphi^{\min} + \varepsilon$ for an arbitrarily small $\varepsilon > 0$ (since $\hat{t}_1(\varphi^{\min}) = 0$). The assumptions on $\bar{t}$ ensure that agents with the highest moral type never lie even if $\varphi = \mathbb{E}(t)$. Therefore, $\mathcal{L}(\hat{\boldsymbol{t}}(\mathbb{E}(t))) < \mathbb{E}(t)$. Since $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$, it follows that $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) < \varphi$ for all $\varphi \in (\mathbb{E}(t) - \Delta(K,1)/\mu, \mathbb{E}(t))$. If $\mathbb{E}(t) \leq \Delta(K,1)/\mu$ then $\mathcal{L}(\hat{\boldsymbol{t}}(0)) > 0$, since $\hat{t}_1(0) > 0$. However, as $\mathcal{L}(\hat{\boldsymbol{t}}(\mathbb{E}(t))) < \mathbb{E}(t)$ and $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$ there exists a unique fixed-point $\xi > 0$ at which $\mathcal{L}(\hat{\boldsymbol{t}}(\xi)) = \xi$. Therefore, $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) < \varphi$ for all $\varphi \in (\xi, \mathbb{E}(t))$. □

Proof of Proposition 2. I omit the proofs for claims $(i) - (iv)$ in the proposition and instead focus on the existence and uniqueness of equilibrium.

Claim 1: For every φ there exists a unique threshold value k^* which is the maximum integer $j \in \{1, ..., K-1\}$ such that $y(K) + \mu\varphi \geq y(j) + \mu\mathbb{E}(t)$. Suppose by contradiction that there is a k^* for which $y(K) + \mu\varphi \leq y(k^*) + \mu\mathbb{E}(t)$. But then, individuals can profitably deviate and report k^*, as in such an equilibrium $\mathcal{R}^C_{k^*} > \mathbb{E}(t) > \varphi$, and hence

$$y(k^*) + \mu\mathcal{R}^C_{k^*} > y(K) + \mu\varphi.$$

This establishes that for any k^*, $y(K) + \mu\varphi \geq y(k^*) + \mu\mathbb{E}(t)$.

To see that k^* is the largest integer, consider a case where $k^* < k'$ and

$$y(K) + \mu\varphi \geq y(k') + \mu\mathbb{E}(t).$$

Since k' is now being lied at, $\mathcal{R}^C_{k'} < \mathbb{E}(t)$. The inequality is a contradiction to the condition that in any such equilibrium, $y(K) + \mu\varphi = y(k') + \mu\mathcal{R}^C_{k'}$.

Claim 2: For every $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, the fraction of agents who lie is a function $S(\varphi) = \frac{1}{K}\sum_{j \in \mathcal{K}} F(\hat{t}_j(\varphi))$. 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 $\hat{t}_j(\varphi)$ is varies continuously between 0 and $\bar{t}$ on $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$ and because $F(t)$ is continuous for $t \in [0, \bar{t}]$. Moreover, $F'(t) > 0$ and $\hat{t}'_j(\varphi) \geq 0$, with strict inequality if $\hat{t}_j(\varphi) > 0$. And since $\hat{t}_1(\varphi) > 0$ for all $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, $S'(\varphi) > 0$.

Claim 3: For every $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$, $D(\varphi) = \frac{1}{K}\sum_{j=k^*+1}^{K} \frac{1 - r_j(\varphi)}{r_j(\varphi)}$ is continuous with $D'(\varphi) <$

0. *In equilibrium,* $D(\varphi) = \text{P}(\textit{lie})$. The fraction of liars that report a state larger than k^* is

$$\sum_{j=k^*+1}^{K} \text{P}(\text{report } j) \times \text{P}(\text{lie}|\text{report } j). \tag{9}$$

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

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

Observe that in equilibrium exactly $\frac{1}{K}$ agents report each state $j > k^*$ truthfully. Thus, we can rearrange the above equation to

$$\text{P}(\text{report } j) = \frac{1}{K}\frac{1}{r_j}.$$

Plugging into (9), we arrive at the following expression

$$\sum_{j=k^*+1}^{K} \text{P}(\text{report } j) \times \text{P}(\text{lie}|\text{report } j) = \frac{1}{K}\sum_{j=k^*+1}^{K} \frac{1-r_j}{r_j}. \tag{10}$$

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

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

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

$$r_j(\varphi) = \frac{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))}{\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))}. \tag{11}$$

Finally, we define

$$D(\varphi) \equiv \frac{1}{K}\sum_{j=k^*+1}^{K} \frac{1-r_j(\varphi)}{r_j(\varphi)} = \frac{1}{K}\sum_{j=k^*+1}^{K} \frac{\mathbb{E}(t) - (\varphi + \Delta(K,j)/\mu)}{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))}. \tag{12}$$

The function $D(\varphi)$ is continuous as $\varphi > \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))$ 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 (12) decrease in φ while the denominators increase as long as

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

which was shown in Lemma 2.

Claim 4: 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 $\varphi \in (\varphi^{\min}, \mathbb{E}(t))$. When $\varphi \to \varphi^{\min}$, $S(\varphi) = 0$ and $D(\varphi) > 0$. In the case where $\varphi \to \mathbb{E}(t)$, $k^* = K - 1$ and thus

$$\lim_{\varphi \to \mathbb{E}(t)} D(\varphi) = \lim_{\varphi \to \mathbb{E}(t)} \frac{1}{K} \frac{\mathbb{E}(t) - \varphi}{\varphi - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))} = 0.$$

It follows that

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

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

A.3 Proof of Proposition 3a

The proof below makes use of properties of log-concave distribution functions. The following lemma states results for log-concave distribution functions that the proof will refer to.

Lemma 3. *Suppose density $f(t)$ is log-concave with $f(t) > 0$ for $t \in (0, \bar{t}]$.*

(i) The hazard rate $h(t) \equiv f(t)/(1 - F(t))$ increases in t and the inverse hazard rate $\iota(t) \equiv f(t)/F(t)$ decreases in t.

(ii) $\mathcal{M}^{+\prime}(t) \in (0, 1)$ and $\mathcal{M}^{-\prime}(t) \in (0, 1)$.

(iii) If f is strictly increasing, $\mathcal{M}^+(t) \leq 1/2 \leq \mathcal{M}^-(t)$ (for strictly decreasing f, $\mathcal{M}^+(t) \geq 1/2 \geq \mathcal{M}^-(t)$).

Proof. Point (i) follows because if $f(t)$ is log concave then $F(t)$ and $1 - F(t)$ are also log concave and because $g'(t)/g(t)$ is decreasing for any log concave function $g(t)$ (see, e.g., Bagnoli and Bergstrom, 2005).

For proofs of points (ii) and (iii), see Lemma 1 in Harbaugh and Rasmusen (2018). □

Proof of Proposition 3a. Define $v(t) \equiv 1 - r(t)$ and $w(t) \equiv \mathcal{M}^+(t) - \mathcal{M}^-(t)$. We can rewrite

$$\Psi(t) = 2v(t)w(t).$$

The function $v(t)$ is increasing. Jewitt (2004) shows that if $f(t)$ is always decreasing $w(t)$ is always increasing, if $f(t)$ is always increasing $w(t)$ is always decreasing and if $f(t)$ is first increasing and then decreasing then $w(t)$ is first decreasing and then increasing. The claim of the proposition immediately follows for $f(t)$ decreasing.
We further show the claim for log-concave distributions. Examine the logarithm of $\Psi(t)$. Its derivative with respect to t is

$$\frac{\partial \log(\Psi(t))}{\partial t} = \frac{1}{w(t)} \left[\frac{v'(t)}{v(t)} w(t) + w'(t) \right] = \frac{1}{w(t)} \left[\frac{f(t)}{F(t)(1 + F(t))} (\mathcal{M}^+(t) - M^-(t)) + \mathcal{M}^{+\prime}(t) - \mathcal{M}^{-\prime}(t) \right]. \tag{13}$$

The derivatives of the conditional expectation terms are

$$\mathcal{M}^{+\prime}(t) = h(t)(\mathcal{M}^+(t) - t),$$
$$\mathcal{M}^{-\prime}(t) = \iota(t)(t - \mathcal{M}^-(t)),$$

where $h(t)$ and $\iota(t)$ are as defined in Lemma 3. The derivative term (13) can only be nonpositive whenever the term in brackets is nonpositive. This condition can be rearranged to

$$\iota(t)(\mathcal{M}^+(t)-t)-\iota(t)(t-\mathcal{M}^-(t))+2\mathcal{M}^{+\prime}(t)\leq 0.$$

For this inequality to hold it is necessary that $t-\mathcal{M}^-(t)>\mathcal{M}^+(t)-t$. By part (ii) of Lemma 3, $t-\mathcal{M}^-(t)$ is increasing while $\mathcal{M}^+(t)-t$ is decreasing. Both terms cross once on $(0,\bar{t}]$. An additional necessary condition for $\Psi'(t)\leq 0$ is that $\mathcal{M}^{-\prime}(t')>\mathcal{M}^{+\prime}(t')$. Let $\tilde{t}$ denote the median of $f(t)$ and consider the following two cases.

Case 1: $\tilde{t}\leq\mathbb{E}(t)$. By the martingale property of beliefs, $\mathcal{M}^+(t)+\mathcal{M}^-(t)=\frac{\mathbb{E}(t)-(1-2F(t))\mathcal{M}^+(t)}{F(t)}$. Plugging in $\tilde{t}$, from $F(\tilde{t})=1/2$ it follows that $\mathcal{M}^+(\tilde{t})+\mathcal{M}^-(\tilde{t})=2E(t)$. Therefore, $\mathcal{M}^+(\tilde{t})-\tilde{t}\geq\tilde{t}-\mathcal{M}^-(\tilde{t})$. At $\tilde{t}$, $h(\tilde{t})=\iota(\tilde{t})$. Combined, these conditions imply that $\mathcal{M}^{+\prime}(\tilde{t})\geq\mathcal{M}^{-\prime}(\tilde{t})$. Log-concavity implies that there is one $\hat{t}'$ so that $\mathcal{M}^{-\prime}(t)>\mathcal{M}^{+\prime}(t)$ for $t<t'$ and $\mathcal{M}^{-\prime}(t)\leq\mathcal{M}^{+\prime}(t)$ otherwise. There thus is no t for which both $t-\mathcal{M}^-(t)>\mathcal{M}^+(t)-t$ and $\mathcal{M}^{-\prime}(t)>\mathcal{M}^{+\prime}(t)$ hold. We conclude that $\Psi'(t)>0$.

Case 2: $\tilde{t}>\mathbb{E}(t)$. Similar steps as above show that we cannot refute both necessary conditions when $\tilde{t}>\mathbb{E}(t)$, i.e., when $f(t)$ is left-skewed. We derive tighter conditions and show that the claim holds nonetheless in the case where f is always increasing (i.e., maximally left-skewed). Rearranging the bracket term in (13) and using $r(t)>1/2$ for $t\in(0,\bar{t})$, a necessary condition for $\Psi'(t)>0$ is that

$$\frac{1}{F(t)}\mathcal{M}^{+\prime}(t)>\mathcal{M}^{-\prime}(t)-\mathcal{M}^{+\prime}(t). \tag{14}$$

This inequality holds as $t\to 0$; the l.h.s. goes to infinity and the r.h.s. is always smaller than one. Consider the derivative $\mathcal{M}^{+\prime}(t)$ as $t\to\bar{t}$. Solving for the limit by repeatedly using l'hopital's rule:

$$\begin{aligned}\lim_{t\to\bar{t}}\mathcal{M}^{+\prime}(t)&=\lim_{t\to\bar{t}}\frac{f(t)\int_t^{\bar{t}}(1-F(s))\,\mathrm{d}s}{(1-F(t))^2}\\&=\lim_{t\to\bar{t}}\frac{f'(t)\int_t^{\bar{t}}(1-F(s))\,\mathrm{d}s-f(t)(1-F(t))}{-2f(t)(1-F(t))}\\&=\frac{1}{2}-\lim_{t\to\bar{t}}\frac{f'(t)\int_t^{\bar{t}}(1-F(s))\,\mathrm{d}s}{2f(t)(1-F(t))}\\&=\frac{1}{2}-\lim_{t\to\bar{t}}\frac{f''(t)\int_t^{\bar{t}}(1-F(s))\,\mathrm{d}s-f'(t)(1-F(t))}{2f(t)(1-F(t))-2f'(t)f(t)}=\frac{1}{2}.\end{aligned}$$

We use this result to show that inequality (14) holds as $t\to\bar{t}$, as

$$\lim_{t\to\bar{t}}\frac{1}{F(t)}\mathcal{M}^{+\prime}(t)=\frac{1}{2}>\lim_{t\to\bar{t}}(\mathcal{M}^{-\prime}(t)-\mathcal{M}^{+\prime}(t))=\underbrace{\overset{\text{By Lemma 3}}{\mathcal{M}^{-\prime}(\bar{t})}}_{<1}-\frac{1}{2}.$$

Inequality (14) thus holds at the extreme points of t. Suppose that it does not hold for some intermediate values of t. In this case the l.h.s. would have to cut the r.h.s. at least twice, once from above and once from below. We show in a last step that the l.h.s. can cross the r.h.s. only from above. Suppose there is a t' such that $\frac{1}{F(t')}\mathcal{M}^{+\prime}(t')=\mathcal{M}^{-\prime}(t')-\mathcal{M}^{+\prime}(t')$. If the l.h.s. cuts from above this means that the derivative of the l.h.s., evaluated at t', is smaller than the derivative of the r.h.s. evaluated at t'. Expressed formally,

$$\frac{1}{F(t')}\left[\mathcal{M}^{+\prime\prime}(t') - \iota(t')\mathcal{M}^{+\prime}(t')\right] < \mathcal{M}^{-\prime\prime}(t') - \mathcal{M}^{+\prime\prime}(t'). \tag{15}$$

The second derivatives of the conditional expectations are

$$\mathcal{M}^{+\prime\prime}(t) = \frac{f'(t)}{f(t)}\mathcal{M}^{+\prime}(t) + h(t)(2\mathcal{M}^{+\prime}(t) - 1),$$
$$\mathcal{M}^{-\prime\prime}(t) = \frac{f'(t)}{f(t)}\mathcal{M}^{-\prime}(t) + \iota(t)(1 - 2\mathcal{M}^{-\prime}(t)).$$

Plugging them into inequality (15) and rearranging,

$$\frac{f'(t)}{f(t)}\left[\frac{1}{F(t')}\mathcal{M}^{+\prime}(t') - (\mathcal{M}^{-\prime}(t') - \mathcal{M}^{+\prime}(t'))\right] <$$
$$\iota(t')(1-2\mathcal{M}^{-\prime}(t')) + h(t')(1-2\mathcal{M}^{+\prime}(t')) - \left[2h(t')\frac{1}{F(t')}\mathcal{M}^{+\prime}(t') - h(t') - \iota(t)\frac{1}{F(t')}\mathcal{M}^{+\prime}(t')\right].$$

The bracket term of the l.h.s. evaluated at t' is zero. We can replace $\frac{1}{F(t')}\mathcal{M}^{+\prime}(t')$ by $\mathcal{M}^{-\prime}(t') - \mathcal{M}^{+\prime}(t')$ on the r.h.s. and rearrange it to get to

$$0 < \iota(t')(1-\mathcal{M}^{-\prime}(t')) + h(t')(1-2\mathcal{M}^{+\prime}(t')) + (h(t') - \iota(t)\mathcal{M}^{+\prime}(t')).$$

The first and second terms are positive as $1 > \mathcal{M}^{-\prime}(t') \geq 1/2 > \mathcal{M}^{+\prime}(t')$ (part (iii) of Lemma 3). Using $\mathcal{M}^{+\prime}(t') = F(t')/(1+F(t')) \times \mathcal{M}^{-\prime}(t')$, the last term is positive as

$$h(t') - \iota(t')\frac{F(t')}{1+F(t')}\mathcal{M}^{-\prime}(t') = h(t') - \underbrace{\frac{f(t')}{1+F(t')}}_{<h(t')}\underbrace{\mathcal{M}^{-\prime}(t')}_{<1} > 0.$$

It follows that the l.h.s. of inequality (14) can cross the r.h.s. at most once. Since the inequality holds as $t \to 0$ and at $\bar{t}$ we conclude that it also holds for all t on $(0, \bar{t})$. □

A.4 Proof of Proposition 3b

Decisions are strategic substitutes if and only if

$$\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\hat{t}_k} = \frac{\partial \hat{t}_j}{\partial \varphi}\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k} < 0.$$

Since $\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} \geq 0$, the inequality holds only if $\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k} < 0$.

An equilibrium obtains when

$$h(\varphi, \hat{t}_k) = K(S(\varphi, \hat{t}_k) - D(\varphi, \hat{t}_k)) = 0,$$

where

$$S(\varphi, \hat{t}_k) = \frac{1}{K}\sum_{j \neq k} F(\hat{t}_j(\varphi)) + F(\hat{t}_k),$$

$$D(\varphi, \hat{t}_k) = \frac{1}{K} \sum_{j\in\mathcal{H}} \frac{\mathbb{E}(t) - \varphi - \Delta(K,j)/\mu}{\varphi - \mathcal{L}(\varphi, \hat{t}_k) + \Delta(K,j)/\mu}.$$

Applying the implicit function theorem to h, we find that

$$\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k} = -\frac{\partial h/\partial \hat{t}_k}{\partial h/\partial \varphi}.$$

Consider

$$K\frac{\partial D}{\partial \hat{t}_k} = \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \sum_{j\in\mathcal{H}} \frac{\mathbb{E}(t) - \varphi + \Delta(K,j)/\mu}{(\varphi - \mathcal{L}(\varphi, \hat{t}_k) + \Delta(K,j)/\mu)^2}.$$

Using Equation (11), we can replace $1/(\varphi - \mathcal{L}(\varphi, \hat{t}_k) + \Delta(K,j)/\mu) = 1/(\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) \times 1/r_j$. Also, note that

$$\frac{\mathbb{E}(t) - \varphi + \Delta(K,j)/\mu}{\varphi - \mathcal{L}(\varphi, \hat{t}_k) + \Delta(K,j)/\mu} = \frac{1-r_j}{r_j}.$$

Therefore, the derivative of D is equal to

$$K\frac{\partial D}{\partial \hat{t}_k} = \frac{1}{\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)} \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \sum_{j\in\mathcal{H}} \frac{1}{r_j}\frac{1-r_j}{r_j}$$

To further simplify, consider

$$r_j = \frac{1}{1+\alpha_j \sum_{l\in\mathcal{K}} F(\hat{t}_l)} \quad \Rightarrow \quad \frac{1}{r_j} = 1 + \alpha_j \sum_{l\in\mathcal{K}} F(\hat{t}_l), \quad \frac{1-r_j}{r_j} = \alpha_j \sum_{l\in\mathcal{K}} F(\hat{t}_l). \tag{16}$$

Therefore, we can make the following replacement:

$$\begin{aligned}
\sum_{j\in\mathcal{H}} \frac{1}{r_j}\frac{1-r_j}{r_j} &= \sum_{j\in\mathcal{H}} (1+\alpha_j \sum_{l\in\mathcal{K}} F(\hat{t}_l))\alpha_j \sum_{l\in\mathcal{K}} F(\hat{t}_l) \\
&= \sum_{l\in\mathcal{K}} F(\hat{t}_l) + \sum_{j\in\mathcal{H}} \alpha_j^2 \left(\sum_{l\in\mathcal{K}} F(\hat{t}_l)\right)^2 \\
&= \sum_{l\in\mathcal{K}} F(\hat{t}_l)(1 + \sum_{j\in\mathcal{H}} \alpha_j^2 \sum_{l\in\mathcal{K}} F(\hat{t}_l)).
\end{aligned}$$

Defining $\tilde{r} \equiv \frac{1}{1+\sum_{j\in\mathcal{H}} \alpha_j^2 \sum_{l\in\mathcal{K}} F(\hat{t}_l)}$, the derivative of h becomes

$$\frac{\partial h}{\partial \hat{t}_k} = \frac{1}{\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)} \left(f(\hat{t}_k)(\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) - \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \sum_{l\in\mathcal{K}} \frac{F(\hat{t}_l)}{\tilde{r}} \right).$$

Multiplying and dividing the r.h.s. by $\tilde{r}^2 \sum_{j\in\mathcal{K}} \alpha_j^2$;

$$\frac{\partial h}{\partial \hat{t}_k} = \underbrace{\frac{1}{\tilde{r}^2 \sum_{j\in\mathcal{K}} \alpha_j^2 (\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k))}}_{=\beta>0} \left(\underbrace{\tilde{r}^2 \sum_{j\in\mathcal{K}} \alpha_j^2 f(\hat{t}_k)}_{=\frac{\partial \tilde{r}}{\partial \hat{t}_k}} (\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) - \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \underbrace{\tilde{r} \sum_{j\in\mathcal{K}} \alpha_j^2 \sum_{l\in\mathcal{K}} F(\hat{t}_l)}_{=1-\tilde{r}} \right).$$

We arrive at

$$\frac{\partial h}{\partial \hat{t}_k} = -\beta \left[\frac{\partial \tilde{r}}{\partial \hat{t}_k} (\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) + (1 - \tilde{r}) \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \right].$$

Therefore,

$$\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k} = \frac{\beta \left[\frac{\partial \tilde{r}}{\partial \hat{t}_k} (\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) + (1 - \tilde{r}) \frac{\partial \mathcal{L}}{\partial \hat{t}_k} \right]}{\partial h / \partial \varphi}.$$

From the proof of Proposition 2 we know that the denominator in the equation above is positive, which implies that, for $\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k}$ to be negative, the numerator must be negative. We conclude that $\frac{\mathrm{d}\varphi^*}{\mathrm{d}\hat{t}_k} < 0$ if and only if

$$(1 - \tilde{r}) \frac{\partial \mathcal{L}}{\partial \hat{t}_k} + \frac{\partial \tilde{r}}{\partial \hat{t}_k} (\mathbb{E}(t) - \mathcal{L}(\varphi, \hat{t}_k)) < 0.$$

A.5 Proof of Proposition 4a

We show that the derivative

$$\frac{1}{2} \frac{\partial \Psi_\tau(t)}{\partial \tau} = (1 - r_\tau(t)) \frac{\partial \mathcal{M}^+_\tau(t)}{\partial \tau} - \frac{\partial r_\tau(t)}{\partial \tau} (\mathcal{M}^+_\tau(t) - \mathcal{M}^-_\tau(t))$$

is positive. The τ subscript denotes the upper bound under the new distribution and we will use $F_\tau(t) = F(t)/F(\tau)$. The upper tail expectation is $\mathcal{M}^+_\tau(t) = \frac{F(\tau)}{F(\tau) - F(t)} \int_t^\tau t f(t) \, \mathrm{d}t$, which has derivative

$$\frac{\partial \mathcal{M}^+_\tau(t)}{\partial \tau} = \frac{f(\tau)}{1 - F(t)} \left(\tau - F(t) \mathcal{M}^+_\tau(t) \right).$$

The derivative of $r_\tau(t)$ with respect to τ is

$$\frac{\partial r_\tau(t)}{\partial \tau} = \frac{f(\tau)(F(\tau) + F(t)) - f(\tau)F(\tau)}{(F(\tau) + F(t))^2} = \frac{f(\tau)F(t)}{(F(\tau) + F(t))^2} = f(\tau) r_\tau(t)(1 - r_\tau(t)).$$

Therefore,

$$\begin{aligned}
\frac{1}{2} \frac{\partial \Psi_\tau(t)}{\partial \tau} &= (1 - r_\tau(t)) \left[\frac{\partial \mathcal{M}^+_\tau(t)}{\partial \tau} - f(\tau) r_\tau(t) (\mathcal{M}^+_\tau(t) - \mathcal{M}^-_\tau(t)) \right] \\
&= f(\tau)(1 - r_\tau(t)) \left[\frac{1}{1 - F(t)} [\tau - F(t) \mathcal{M}^+_\tau(t)] - r_\tau(t) (\mathcal{M}^+_\tau(t) - \mathcal{M}^-_\tau(t)) \right] \\
&= f(\tau)(1 - r_\tau(t)) \left[\frac{F(t)}{1 - F(t)} [\tau - \mathcal{M}^+_\tau(t)] + \tau - r_\tau(t) \mathcal{M}^+_\tau(t) + r_\tau(t) \mathcal{M}^-_\tau(t) \right] > 0.
\end{aligned}$$

The derivative is positive. After an increase in τ, the stigma function is larger for any t. As $\tau \to \bar{t}$, $\Psi_\tau(t) \to \Psi(t)$, it follows that the stigma function under $\tau < \bar{t}$ cuts the direct payoff at a larger t, i.e, $\hat{t}$ increases.

A.6 Proof of Proposition 4a′

We will show that the derivative of the stigma function of the deed-based model with respect to τ,

$$\frac{\partial \Psi_\tau^D(t)}{\partial \tau} = -\frac{\partial r_\tau(t)}{\partial \tau},$$

is negative. From the proof of Proposition 4a, we know that

$$\frac{\partial r_\tau(t)}{\partial \tau} = f(\tau) r_\tau(t)(1 - r_\tau(t)) > 0.$$

Therefore $\frac{\partial \Psi_\tau^D(t)}{\partial \tau} < 0$. It follows that the stigma function under $\tau < \bar{t}$ cuts the direct payoff at a smaller t, i.e., $\hat{t}$ increases.

A.7 Proof of Proposition 4b

Decompose the stigma function into two parts like in the proof of Proposition 3a;

$$\Psi_X(t) = v_X(t) w_X(t)$$

under $f_X(t)$ and analogously for $f_Y(t)$. We know that $v_Y(t) = v_X(t-a)$ and $w_Y(t) = w_X(t-a)$. Therefore, $\Psi_Y(t) = \Psi_X(t-a)$. Since $\Psi'_X(t), \Psi'_Y(t) > 0$ and $a > 0$, $\Psi_Y(t) < \Psi_X(t)$ for all $t \in (0, \bar{t} - a]$. This implies that $\Psi_Y(t)$ cuts the direct payoff at a larger t than $\Psi_X(t)$ and the result follows.

A.8 Proof of Proposition 4b′

Since $F_X(t) > F_Y(t)$ for $t \in (0, \bar{t})$,

$$\Psi_X^D(t) = \frac{F_X(t)}{1 + F_X(t)} > \frac{F_Y(t)}{1 + F_Y(t)} = \Psi_Y^D(t).$$

This implies that $\Psi_Y^D(t)$ cuts the direct payoff at a larger t than $\Psi_X^D(t)$ and the result follows.

A.9 Proof of Proposition 4c

The proof relies on a Lemma on the properties of the Unimodal Likelihood Ratio.

Lemma 4 (Metzger and Rüschendorf (1991), Theorems 2.3 and 2.3 (c))**.** *If $f_X(t)/f_Y(t)$ is unimodal with maximum at $\tilde{t}_1$, then $F_X(t)/F_Y(t)$ is unimodal with maximum at $\tilde{t}_2 > \tilde{t}_1$ and $(1 - F_X(t))/(1 - F_Y(t))$ is unimodal with a maximum at $\tilde{t}_3 < \tilde{t}_1$.*

Proof of Proposition 4c
Claim 1: Consider the inverse hazard rates $\iota_X(t)$ and $\iota_Y(t)$. There is a $\tilde{t} \in (0, \bar{t})$ such that $\iota_X(t) > \iota_Y(t)$ for $t < \tilde{t}$ and $\iota_X(t) \leq \iota_Y(t)$ for $t \geq \tilde{t}$. By Lemma 4, the ratio $F_X(t)/F_Y(t)$ will be unimodal (first increasing and then decreasing) on $(0, \bar{t}]$. This implies that the sign of $f_X(t)/F_X(t) - f_Y(t)/F_Y(t)$ changes once from positive to negative, which implies the claim.
Claim 2: Consider the hazard rates $h_X(t)$ and $h_Y(t)$. There is a $\tilde{t} \in (0, \bar{t})$ such that $h_X(t) > h_Y(t)$ for $t < \tilde{t}$ and $h_X(t) \leq h_Y(t)$ for $t \geq \tilde{t}$. By Lemma 4, the ratio $(1 - F_X(t))/(1 - F_Y(t))$ will

be unimodal (first increasing and then decreasing) on $(0, \bar{t}]$. This implies that the sign of $f_X(t)/(1 - F_X(t)) - f_Y(t)/(1 - F_Y(t))$ changes once from positive to negative, which implies the claim.

Claim 3: $\mathcal{M}^-_X(t) \geq \mathcal{M}^-_Y(t)$ *for all* $t \in (0, \bar{t}]$. At $t = \bar{t}$, since the means of $f_Y(t)$ and $f_X(t)$ coincide, $\mathcal{M}^-_Y(\bar{t}) = \mathbb{E}_Y(t) = \mathbb{E}_X(t) = \mathcal{M}^-_X(\bar{t})$. Also, as $t \to 0$, both $\mathcal{M}^-_X(t)$ and $\mathcal{M}^-_Y(t)$ go to zero. Consider

$$\mathcal{M}^{-\prime}_X(t) - \mathcal{M}^{-\prime}_Y(t) = \iota_X(t)(\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)) + (t - \mathcal{M}^-_Y(t))(\iota_X(t) - \iota_Y(t)).$$

Evaluated at $\bar{t}$, $\mathcal{M}^{-\prime}_X(\bar{t}) - \mathcal{M}^{-\prime}_Y(\bar{t}) = (\bar{t} - \mathbb{E}_Y(t))(\iota_X(\bar{t}) - \iota_Y(\bar{t}))$. By Claim 1, it follows that $\mathcal{M}^{-\prime}_X(\bar{t}) - \mathcal{M}^{-\prime}_Y(\bar{t}) < 0$, i.e., $\mathcal{M}^-_Y(\bar{t})$ cuts $\mathcal{M}^-_X(\bar{t})$ from below. Consider gradually decreasing t, starting at $\bar{t}$. As long as $\mathcal{M}^{-\prime}_X(t) - \mathcal{M}^{-\prime}_Y(t) < 0$, it holds that $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t) > 0$. Observe that $\mathcal{M}^{-\prime}_X(t) - \mathcal{M}^{-\prime}_Y(t)$ can only be zero if $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)$ and $\iota_X(t) - \iota_Y(t)$ have opposite signs. It follows that the largest value for t where $\mathcal{M}^{-\prime}_X(t) - \mathcal{M}^{-\prime}_Y(t)$ is zero is where $\iota_X(t) - \iota_Y(t) > 0$ and $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t) < 0$. Since the difference $\iota_X(t) - \iota_Y(t)$ changes its sign only once from positive to negative for possible values of t and $\iota_X(\bar{t}) - \iota_Y(\bar{t}) < 0$ (Claim 1), this is also the unique point where $\mathcal{M}^{-\prime}_X(t) - \mathcal{M}^{-\prime}_Y(t)$ is zero. This shows that $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)$ is quasiconcave, which, taken together with the fact that $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)$ is zero as $t \to 0$ and $t = \bar{t}$, implies the initial claim.

Claim 4: $\mathcal{M}^+_X(t) \leq \mathcal{M}^+_Y(t)$ *for all* $t \in (0, \bar{t}]$. At $t = \bar{t}$, since the means of $f_Y(t)$ and $f_X(t)$ coincide, $\mathcal{M}^+_Y(\bar{t}) = \bar{t} - \mathbb{E}_Y(t) = \bar{t} - \mathbb{E}_X(t) = \mathcal{M}^+_X(\bar{t})$. Also, as $t \to 0$, both $\mathcal{M}^+_X(t)$ and $\mathcal{M}^+_Y(t)$ go to $\mathbb{E}_X(t) = \mathbb{E}_Y(t)$. Consider

$$\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_X(t) = h_X(t)(\mathcal{M}^+_Y(t) - \mathcal{M}^+_X(t)) + (\mathcal{M}^+_Y(t) - t)(h_Y(t) - h_X(t)).$$

As $t \to 0$, $\mathcal{M}^{+\prime}_X(t) - \mathcal{M}^{+\prime}_Y(t) = \mathbb{E}_X(t)(h_Y(t) - h_X(t))$. By Claim 2, it follows that $\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_Y(t) > 0$. Therefore, $\mathcal{M}^+_Y(t) > \mathcal{M}^+_X(t)$ for small t. Consider gradually increasing t starting from zero. As long as $\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_Y(t) > 0$, it holds that $\mathcal{M}^+_X(t) > \mathcal{M}^+_Y(t)$. Observe that $\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_Y(t)$ is only zero if $\mathcal{M}^+_Y(t) - \mathcal{M}^+_X(t)$ and $h_Y(t) - h_X(t)$ hold the opposite sign. The smallest value t where $\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_Y(t)$ can be zero is where $\mathcal{M}^+_Y(t) - \mathcal{M}^+_X(t) > 0$ and where $h_Y(t) - h_X(t) < 0$. Since the difference $h_Y(t) - h_X(t)$ changes its sign only once from positive to negative for possible values of t (Claim 2), this is also the unique point where $\mathcal{M}^{+\prime}_Y(t) - \mathcal{M}^{+\prime}_Y(t)$ is zero. This shows that $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)$ is quasiconcave, which, taken together with the fact that $\mathcal{M}^-_X(t) - \mathcal{M}^-_Y(t)$ is zero as $t \to 0$ and $t = \bar{t}$, implies the initial claim.

Claim 5: $\Psi_Y(t) - \Psi_X(t) \geq 0$ *for all* $t \in (0, \bar{t}]$. As before, we use the definition

$$\Psi_Y(t) \equiv 2 \underbrace{v_Y(t)}_{\equiv \frac{F_Y(t)}{1+F_Y(t)}} \overbrace{w_Y(t)}^{\equiv \mathcal{M}^+_Y(t) - \mathcal{M}^-_Y(t)}$$

and symmetrically for $\Psi_X(t)$. The condition of the claim implies

$$\Psi_Y(t) - \Psi_X(t) \geq 0 \Rightarrow (v_Y(t) - v_X(t))w_X(t) + v_Y(t)(w_Y(t) - w_X(t)) \geq 0.$$

Consider that

$$v_Y(t) - v_X(t) = \frac{F_Y(t) - F_Y(t)}{(1 + F_Y(t))(1 + F_X(t))}.$$

Plugging this into the initial condition and simplifying, we have

$$\begin{aligned}
&\frac{F_Y(t)-F_X(t)}{1+F_X(t)}w_X(t)+F_Y(t)(w_Y(t)-w_X(t))\geq 0\\
\Rightarrow\ &\frac{F_Y(t)}{1+F_X(t)}w_X(t)-F_Y(t)w_X(t)+F_Y(t)w_Y(t)-\frac{F_X(t)}{1+F_X(t)}w_X(t)\geq 0\\
\Rightarrow\ &w_X(t)\left(\frac{F_Y(t)}{1+F_X(t)}-F_Y(t)\right)+w_Y(t)\left(F_Y(t)-\frac{F_Y(t)}{1+F_X(t)}\right)+\frac{F_Y(t)}{1+F_X(t)}w_X(t)\\
&\qquad -\frac{F_X(t)}{1+F_X(t)}w_X(t)\geq 0\\
\Rightarrow\ &\left(F_Y(t)-\frac{F_Y(t)}{1+F_X(t)}\right)(w_Y(t)-w_X(t))+\frac{1}{1+F_X(t)}\left(F_Y(t)w_Y(t)-F_X(t)w_X(t)\right)\geq 0.
\end{aligned}$$

The first term is nonnegative as $F_Y(t)\geq F_Y(t)/(1+F_X(t))$ and $w_Y(t)\geq w_X(t)$ by claims 3 and 4. By the martingale property of beliefs,

$$\mathbb{E}(t)=F(t)\mathcal{M}^-(t)+(1-F(t))\mathcal{M}^+(t)\Rightarrow\mathcal{M}^+(t)-\mathcal{M}^-(t)=\frac{\mathcal{M}^+(t)-\mathbb{E}(t)}{F(t)}.$$

We can substitute this equality into the second term of the inequality above;

$$\frac{1}{1+F_X(t)}\left(\mathcal{M}_Y^+(t)-\mathbb{E}_Y(t)-\mathcal{M}_X^+(t)+\mathbb{E}_X(t)\right).$$

It then follows from Claim 4 that this term is also nonnegative, with strict inequality for values on $(0,\bar{t})$. We conclude that the inequality holds for all possible t, which proves the claim.

The claims imply that $\Psi_Y(t)-\Psi_X(t)\geq 0$, with strict inequality for interior values of t. This implies that $\hat{t}_Y<\hat{t}_X$ when comparing two interior equilibria, which proves the proposition.

A.10 *Proof of Proposition 4c′*

Lemma 4 states that unimodality of $f_X(t)/f_Y(t)$ implies unimodality of $F_X(t)/F_Y(t)$. Furthermore, we know that $F_X(\bar{t})/F_Y(\bar{t})=1$. Going to the other extreme, suppose that $\lim_{t\to 0}F_X(t)/F_Y(t)\geq 1$. But then, because of unimodality of $F_X(t)/F_Y(t)$, $F_X(t)$ and $F_Y(t)$ would never cross. This implies first-order stochastic dominance of $F_Y(t)$ over $F_X(t)$, which is inconsistent with the property that $\mathbb{E}_Y(t)=\mathbb{E}_X(t)$. Therefore, $\lim_{t\to 0}F_X(t)/F_Y(t)<1$. Unimodality of $F_X(t)/F_Y(t)$ then implies that there is a unique point on $(0,\bar{t})$ where $F_X(t)/F_Y(t)=1$, i.e., $F_X(t)$ and $F_Y(t)$ cross exactly once on for interior values of t. Denote this crossing point by $\tilde{t}$. We have

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

Observe that the stigma function in the deed-based model is equal to $\Psi^D t=F(t)/(1+F(t))$. It immediately follows that

$$\Psi_X^D(t)<\Psi_Y^D(t)\text{ if }t<\tilde{t},\ \Psi_X^D(t)=\Psi_Y^D(t)\text{ if }t=\tilde{t},\text{ and }\Psi_X^D(t)>\Psi_Y^D(t)\text{ if }t>\tilde{t}.$$

Therefore, if under belief $f_X(t)$, the equilibrium threshold respectively is smaller, equal to, or larger than $\tilde{t}$, the stigma function under $f_Y(t)$ will cross the direct payoff respectively at a larger, the same, or smaller t. This directly implies the statement in the proposition.

A.11 Proof of Proposition 4da

With coarse disclosure, we have

$$r_j(\varphi) = \frac{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))}{(1-\pi)(\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi)))},$$
$$1 - r_j(\varphi) = \frac{(1-\pi)\mathbb{E}(t) + \pi\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) - (\varphi + \Delta(K,j)/\mu)}{(1-\pi)(\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi)))} \text{ for } j \in \mathcal{H}.$$

Equilibrium is characterized by a function

$$h(\varphi,\pi) \equiv \frac{1}{K}\sum_{j\in\mathcal{K}} F(\hat{t}_j(\varphi)) - \frac{1}{K}\sum_{j\in\mathcal{H}} \frac{(1-\pi)\mathbb{E}(t) + \pi\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) - (\varphi + \Delta(K,j)/\mu)}{\varphi + \Delta(K,j)/\mu - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))} = 0.$$

This function implicitly defines the equilibrium $\varphi^*(\pi)$. It increases in φ and in π. Consider two values π and $\pi' > \pi$. It holds that

$$h(\varphi^*(\pi'),\pi') = h(\varphi^*(\pi),\pi) = 0 < h(\varphi^*(\pi),\pi').$$

Therefore, $\varphi^*(\pi') < \varphi^*(\pi)$. Since $S'(\varphi) > 0$ lying is higher under π than under π', which implies (ii).

To show point (i), that k^* weakly increases, consider that the proof of Proposition 2 shows that k^* is the largest state to which a liar would not deviate to. With an initial probability of lie detection π, this condition becomes

$$y(K) + \mu\varphi^*(\pi) \geq y(k^*) + \mu[(1-\pi)\mathbb{E}(t) + \pi\mathcal{L}(\hat{\boldsymbol{t}}(\varphi^*(\pi)))].$$

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. We thus have to show that the difference

$$(1-\pi)\mathbb{E}(t) + \pi\mathcal{L}(\hat{\boldsymbol{t}}(\varphi^*(\pi))) - \varphi^*(\pi)$$

decreases in π. Plugging in, the difference becomes

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

Taking the derivative with respect to π;

$$-(1-r_K)\underbrace{(\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi^*(\pi))))}_{>0} - \underbrace{\frac{\mathrm{d}r_K}{\mathrm{d}\pi}}_{>0}(1-\pi)(\mathbb{E}(t) - \mathcal{L}(\boldsymbol{t}(\varphi^*(\pi)))) + \underbrace{\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\pi}}_{<0}(1-r_K)(1-\pi) < 0.$$

Therefore, the threshold state weakly increases.

A.12 Proof of Proposition 4db

One issue with multiplicity arises as the equilibrium definition does not pin down the reputation of an agent from a state $j \in \mathcal{H}$ who (off equilibrium) is disclosed lying. A standard equilibrium refinement, which says that an off-equilibrium inference will attribute the message to the agent with the strongest incentive to deviate pins down off-equilibrium reputations after disclosure at $\mathcal{M}^-(0)$ and ensures uniqueness. Denote the investigator's policy by (π, γ). The variable γ denotes the probability reveals a liar's drawn state. If $\gamma = 0$, we are under the coarse disclosure regime. Denote the part of state K's reputation that is independent of $\mathcal{M}^-(\hat{t}_j)$ as

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

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

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

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

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

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

$$-k^* \frac{\mathrm{d}\varphi^{C*}}{\mathrm{d}\gamma} > \sum_{j\in\mathcal{K}} \pi\mathcal{M}^-(\hat{t}_j^*). \tag{17}$$

To derive $\mathrm{d}\varphi^C/\mathrm{d}\gamma$, we apply the implicit function theorem to the equilibrium condition

$$h(\varphi, \gamma) = K(D(\varphi, \gamma) - S(\varphi, \gamma)).$$

We find that

$$\frac{\mathrm{d}\varphi^c}{\mathrm{d}\gamma} = \frac{\left(\tilde{r}(\hat{\boldsymbol{t}}^*)\xi\frac{\mathrm{d}m}{\mathrm{d}x} - (1-\tilde{r}(\hat{\boldsymbol{t}}^*))\frac{\mathrm{d}n}{\mathrm{d}x}\right) - \frac{\mathrm{d}\tilde{r}}{\mathrm{d}x}(1-\pi)(\mathbb{E}(t) - \mathcal{L}(\varphi, x))}{\frac{\mathrm{d}\tilde{r}}{\mathrm{d}\varphi}(1-\pi)(\mathbb{E}(t) - \mathcal{L}(\hat{\boldsymbol{t}}^*)) + \tilde{r}(\hat{\boldsymbol{t}}^*)\xi + (1-\tilde{r}(\hat{\boldsymbol{t}}^*))(1-\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi})},$$

where $m(\varphi,\gamma) \equiv (1-\pi)\mathbb{E}(t) + (1-\pi\gamma)\mathcal{L}(\hat{\boldsymbol{t}}) - \varphi$, $n(\varphi,\gamma) \equiv \varphi^C - (1-\pi\gamma)\mathcal{L}(\hat{\boldsymbol{t}})$, $\xi \equiv \tilde{r}\sum_{j\in\mathcal{K}} \alpha_j^2(K-k^*) + (1-\tilde{r}))$, and $\tilde{r}$ is defined as in Proposition 3b. Plugging this into (17) and rearranging gives

$$((1-\tilde{r}(\hat{\boldsymbol{t}}^*))+\xi\tilde{r}(\hat{\boldsymbol{t}}^*))k^*\left(\mathcal{L}(\hat{\boldsymbol{t}}^*)-\frac{1}{k^*}\sum_{j\in\mathcal{K}}\mathcal{M}^-(\hat{t}_j^*)\right)>$$

$$>(1-\pi)(\mathbb{E}(t)-\mathcal{L}(\hat{\boldsymbol{t}}^*))\sum_{j\in\mathcal{K}}\frac{\mathrm{d}\tilde{r}}{\mathrm{d}\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}\left(\sum_{k\in\mathcal{K}}\mathcal{M}^-(\hat{t}_k^*)-k^*\mathcal{M}^-(\hat{t}_j^*)\right)+$$

$$+((1-\tilde{r}(\hat{\boldsymbol{t}}^*))(1-\pi\gamma)+\xi\tilde{r}(\hat{\boldsymbol{t}}^*)\pi(1-\gamma))\sum_{j\in\mathcal{K}}\left(\frac{\partial\mathcal{L}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi^C}(k^*\mathcal{M}^-(\hat{t}_j^*)-\sum_{k\in\mathcal{K}}\mathcal{M}^-(\hat{t}_k^*))\right).$$

The left hand side displays the direct effect of increasing γ, which is positive since there are relatively more large than small liars, implying that the weighted average of their types is higher than the unweighted average; $\mathcal{L}(\hat{\boldsymbol{t}}(\varphi^C,\pi,\gamma))>\frac{1}{k^*}\sum_{j\in\mathcal{K}}\mathcal{M}^-(\varphi^C,\pi,\gamma)$. The first term on the right hand side cancels out. To see this, note that $\frac{\partial\tilde{r}}{\partial\hat{t}_j}\frac{\partial\hat{t}_j}{\partial\varphi}=\frac{\partial\tilde{r}}{\partial\hat{t}_1}\frac{\partial\hat{t}_1}{\partial\varphi^C}$ for $j\leq k^*$ and zero otherwise, so that the derivative terms can be moved out of the sum and the remaining terms add up to zero. We move to the second term on the r.h.s. Rewrite

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

Plugging the last two results into the inequality,

$$((1-\tilde{r}(\hat{\boldsymbol{t}}^*))+\xi\tilde{r}(\hat{\boldsymbol{t}}^*))k^*\left(\mathcal{L}(\hat{\boldsymbol{t}}^*)-\frac{1}{k^*}\sum_{j\in\mathcal{K}}\mathcal{M}^-(\hat{t}_j^*)\right)>$$

$$>((1-\tilde{r}(\hat{\boldsymbol{t}}^*))(1-\pi\gamma)+\xi\tilde{r}(\hat{\boldsymbol{t}}^*)\pi(1-\gamma))\frac{\partial\hat{t}_1}{\partial\varphi^C}k^*\left(\mathcal{L}(\hat{\boldsymbol{t}}^*)-\frac{1}{k^*}\sum_{j\in\mathcal{K}}\mathcal{M}^-(\hat{t}_j^*)\right).$$

This inequality holds since $(1-\tilde{r}(\hat{\boldsymbol{t}}^*))+\xi\tilde{r}(\hat{\boldsymbol{t}}^*)\geq(1-\tilde{r}(\hat{\boldsymbol{t}}^*))(1-\pi\gamma)+\xi\tilde{r}(\hat{\boldsymbol{t}}^*)\pi(1-\gamma)$ and $\frac{\partial\hat{t}_1}{\partial\varphi^C}<1$ as long as μ is not too large. We conclude that the aggregate lying rate decreases. The average size of the lie increases because $\frac{\mathrm{d}\hat{t}_1}{\mathrm{d}\gamma}>\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\gamma}>0>\frac{\mathrm{d}\hat{t}_{j+1}}{\mathrm{d}\gamma}>\frac{\mathrm{d}\hat{t}_{k^*}}{\mathrm{d}\gamma}$ for some $j<k^*$, so that the proportion of lower states among liars increases, while the

proportion of higher states decreases.

A.13 Proof of Proposition 4d′

With deed-based image concerns, the equilibrium properties are that agents lie if and only if they draw $j \leq k^* < K$. 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 drawing j is equal to

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

where φ^D denotes the reputation of reporting K. In equilibrium, the reputation of K is equal to

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

It follows that

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

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

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

Similar arguments as those given in the proof of Proposition 2 imply that, in equilibrium,

$$h^D(\varphi, \pi) \equiv \sum_{j \leq k^*} F(\hat{t}_j(\varphi^D)) - \sum_{k > k^*} \frac{1 - r_j(\varphi^D, \pi)}{r_j(\varphi^D, \pi)} = 0.$$

This function uniquely defines the equilibrium $\varphi^{D*}(\pi)$. It increases in φ^D and π. Consider two values π and $\pi' > \pi$. It holds that

$$h(\varphi^{D*}(\pi'), \pi') = h(\varphi^{D*}(\pi), \pi) = 0 < h(\varphi^{D*}(\pi), \pi').$$

Therefore, $\varphi^{D*}(\pi') < \varphi^{D*}(\pi)$. We conclude (ii): lying is higher under π than under π'.

(i): The threshold state k^* is the largest integer $j < K$ for which $y(K) + \mu(1 - \pi) r_K(\varphi^D, \pi) \geq y(j) + \mu(1 - \pi)$. After increasing π, the reputation terms on the l.h.s. and r.h.s. will decrease. If the decrease in reputation on the r.h.s. is larger than the decrease on the l.h.s. the inequaltiy becomes more binding, suggesting that it may also hold for $k^* + 1$. We show that the derivative of

$$(1 - \pi) - (1 - \pi) r_K(\varphi^D, \pi) = (1 - \pi)(1 - r_K(\varphi^D, \pi))$$

with respect to π is negative. This derivative is

$$-(1 - r_K(\varphi^D, \pi)) - (1 - \pi) \underbrace{\frac{\mathrm{d}r_K}{\mathrm{d}\pi}}_{>0} < 0.$$

Therefore, the threshold state weakly increases.

Online Appendix

A Extensions

This section considers two extensions that change two assumptions on preferences that we maintained throughout the main text: fixed lying costs and a homogenous image concern.

A.1 Increasing lying costs

This part considers the robustness of the results to possible generalizations of the lying cost function. Consider a more general case of agents' utility function:

$$u(j, t, a) = y(a) - c(j, t, a) + \mu \mathcal{R}_a.$$

In the main part we considered fixed lying costs where $c(j, t, a) = t$. GK&S and K&S provide results for the case of a deed-based model with lying costs consisting of a fixed, moral type-dependent and a variable, moral type-independent component. For example, K&S study the case where $c(j, t, a) = t + \kappa|a - j|$. They show that all equilibrium features of the deed-based model remain qualitatively the same; introducing the variable cost changes how participants trade off full and partial lying on the margin (a higher variable cost parameter κ makes partial lying relatively more attractive) but does not lead to a qualitatively different equilibrium. It is relatively straightforward to show that the same results translate to our setting. As long as variable lying costs do not depend on the moral type, they will not fundamentally change equilibrium behavior.

To study another potentially interesting case, in this part I consider behavior under type-dependent increasing lying costs of the form $c(j, t, a) = t|a - j|$. Here, the moral type now determines the slope of the lying cost function, with higher types facing a steeper slope. There are two reasons why increasing lying costs can lead to partial lying. First, with increasing lying costs, agents might prefer to tell a partial lie for purely intrinsic reasons. For example, if the direct payoff function $y(a)$ is strictly concave, then there might be agents whose payoff gain outweighs the moral cost when telling a small lie (going from 1 to 3), but not when telling a large lie (going from 1 to 6). Second, increasing lying costs might interact with the image concern to motivate agents to lie partially. To be able to cleanly state how signaling motives change agents' lying behavior under increasing lying costs, we will in this part assume that $y(a) = a$. This establishes an easily comparable benchmark; if lying costs and direct payoffs are linear functions, then absent image concerns, agents either lie to report K or tell the truth. If we set $\mu = 0$, agents will lie if and only if $t \leq 1$ and $j < K$. If they lie, they will report K.

We can now ask how behavior is different with image concerns, i.e., when $\mu > 0$. Suppose that an equilibrium exists where every liar reports K. In such an equilibrium, a larger fraction must lie after drawing 1 than after drawing $K - 1$. This implies two things; first, the marginal liar from state $K - 1$ is indifferent between reporting K and $K - 1$. Second, the marginal liar from state 1 is indifferent between reporting K and 1. Taken together, both facts imply that the marginal liar form 1 is of a higher moral type than the marginal liar from $K - 1$. A consequence is that the marginal liar from 1 will prefer reporting $K - 1$ over K, a contradiction:

Proposition 5a. *With lying costs $c(t, j, a) = t|j - a|$, $y(a) = a$, and $\mu > 0$ and if $K > 2$, there is no equilibrium in which liars only report K.*

Proof. Suppose there is an equilibrium in which liars only report K. Then, there are $K-1$ indifference conditions, which, for every state $1, \ldots, K-1$ determine a threshold type $\hat{t}_j$. Agents of type (j, t) will lie if $t \leq \hat{t}_j$. The indfference condition is

$$K + \mu\mathcal{R}_K - \hat{t}_j(K - j) = j + \mu\mathcal{R}_j,$$

which can be rewritten to

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

Note that in equilibrium $\mathcal{R}_K < \mathcal{R}_j$. It follows that $\hat{t}_1 > \ldots > \hat{t}_{K-1}$. In equilibrium, no type can have an incentive to deviate and lie to a number different from K. Consider the type $\hat{t}_1$. The incentive constraint postulates that

$$K + \mu\mathcal{R}_K - \hat{t}_1(K - 1) \geq K - 1 + \mu\mathcal{R}_{K-1} - \hat{t}_1(K - 2).$$

Rearranging, this condition is equal to

$$1 + \mu(\mathcal{R}_K - \mathcal{R}_{K-1}) \geq \hat{t}_1$$

Note however that in equilibrium,

$$1 + \mu(\mathcal{R}_K - \mathcal{R}_{K-1}) = \hat{t}_{K-1},$$

which implies $\hat{t}_1 \leq \hat{t}_{K-1}$, a contradiction. Therefore, an equilibrium where every liar reports K does not exist. □

In contrast to the no-image benchmark, variable lying costs predict an equilibrium with a lot of partial lying. In this equilibrium, agents who draw j will report any number between $j+1$ and K if they lie–which number they exactly report depends on their moral type. The least moral types report K, followed by slightly more moral types reporting $K - 1$ and so on:

Proposition 5b. *When lying costs are of the form $c(t, j, a) = t|j - a|$, $y(a) = a$, and $\mu > 0$, there is an equilibrium which is characterized by the threshold types $1 > \hat{t}_1 > \ldots > \hat{t}_{K-1} > \hat{t}_K = 0$. In equilibrium, agents of type (j, t) will lie and report k if and only if $j < k$ and $t \in (\hat{t}_k, t_{k-1}]$.*

Proof. We construct the equilibrium stated in the proposition and then show that it exists. Denote by

$$\mathcal{M}^B(\hat{t}_a, \hat{t}_b) = \mathbb{E}(t|t \in (\hat{t}_a, \hat{t}_b])$$

the expected moral type if the moral type is between two thresholds $\hat{t}_a < \hat{t}_b$. Define a function

$$\mathcal{R}(\hat{t}_a, \hat{t}_b, j) \equiv \frac{(1 - F(\hat{t}_a))\mathcal{M}^+(\hat{t}_a) + (j - 1)(F(\hat{t}_b) - F(\hat{t}_a))\mathcal{M}^B(\hat{t}_a, \hat{t}_b)}{(1 - F(\hat{t}_a)) + (j - 1)(F(\hat{t}_b) - F(\hat{t}_a))}.$$

In the stated equilibrium, the reputations of the different states will be

$$\mathcal{R}_j(\hat{t}_j, \hat{t}_{j-1}) = \mathcal{R}(\hat{t}_j, \hat{t}_{j-1}, j) \text{ if } j > 1 \text{ and}$$
$$\mathcal{R}_1(\hat{t}_1) = \mathcal{M}^+(\hat{t}_1).$$

The equilibrium $\hat{t}$-thresholds are determined by a number of indifference conditions. For example, the type $\hat{t}_1$ must be indifferent between truthfully reporting 1 or lying to report 2;

$$2 + \mu\mathcal{R}(\hat{t}_2, \hat{t}_1, 2) - \hat{t}_1 = 1 + \mu\mathcal{M}^+(\hat{t}_1).$$

Rearranging this condition, we can define a function $\mathcal{T}_1(\hat{t}_1, \hat{t}_2) \equiv 1 + \mu(\mathcal{R}(\hat{t}_2, \hat{t}_1, 2) - \mathcal{M}^+(\hat{t}_1)) - \hat{t}_1$. In equilibrium, $\mathcal{T}_1(\hat{t}_1^*, \hat{t}_2^*) = 0$. In a similar fashion, we can define the functions

$$\begin{aligned}\mathcal{T}_j(\hat{t}_{j-1}, \hat{t}_j, \hat{t}_{j+1}) \equiv& 1 + \mu(\mathcal{R}(\hat{t}_{j+1}, \hat{t}_j, j+1) - \mathcal{R}(\hat{t}_j, \hat{t}_{j-1}, j)) - \hat{t}_j \text{ for } j \in \{2, \ldots, K-2\} \text{ and}\\ \mathcal{T}_{K-1}(\hat{t}_{K-2}, \hat{t}_{K-1}) \equiv& 1 + \mu(\mathcal{R}(0, \hat{t}_{K-1}, K) - \mathcal{R}(\hat{t}_{K-1}, \hat{t}_{K-2}, K-1)).\end{aligned}$$

All of them need to equal zero in equilibrium. We can solve them recursively. Begin with $\mathcal{T}_1(\hat{t}_1, \hat{t}_2)$ and fix $\hat{t}_2$ at any value between 0 and 1. Note that $\mathcal{R}(\hat{t}_j, \hat{t}_j, j) = \mathcal{M}^+(\hat{t}_j)$ and

$$\mathcal{R}(\hat{t}_j, 1, j) = \frac{(1 - F(\hat{t}_j))\mathcal{M}^+(\hat{t}_j) + (j-1)(F(1) - F(\hat{t}_j))\mathcal{M}^B(\hat{t}_j, 1)}{(1 - F(\hat{t}_j)) + (j-1)(F(1) - F(\hat{t}_j))} < \mathcal{M}^+(\hat{t}_j) \leq \mathcal{M}^+(1).$$

Therefore,

$$\begin{aligned}\mathcal{T}_1(\hat{t}_2, \hat{t}_2) =& 1 + \mu(\mathcal{M}^+(\hat{t}_2) - \mathcal{M}^+(\hat{t}_2)) - \hat{t}_2 = 1 - \hat{t}_2 > 0 \text{ and}\\ \mathcal{T}_1(1, \hat{t}_2) =& 1 + \mu(\mathcal{R}(\hat{t}_2, 1, 2) - M^+(1)) - 1 = \mu(\mathcal{R}(\hat{t}_2, 1, 2) - M^+(1)) \leq 0.\end{aligned}$$

By the intermediate value theorem, there exists a $\hat{t}_1$ for any $\hat{t}_2 \in (0, 1]$ which solves $\mathcal{T}_1$. We conclude that $\mathcal{T}_1$ implicitly defines a function $\hat{t}_1^*(\hat{t}_2)$ with the properties $\hat{t}_1^*(\hat{t}_2) > \hat{t}_2$ if $\hat{t}_2 < 1$ and $\hat{t}_1^*(1) = 1$.

We can use this implicitly defined function to replace $\hat{t}_1$ in equation $\mathcal{T}_2$;

$$\mathcal{T}_2(\hat{t}_1^*(\hat{t}_2), \hat{t}_2, \hat{t}_3) = 1 + \mu(\mathcal{R}(\hat{t}_3, \hat{t}_2, 3) - \mathcal{R}(\hat{t}_2, \hat{t}_1^*(\hat{t}_2), 2)) - \hat{t}_2.$$

Making use of former results, note that for any $\hat{t}_3 \in (0, 1]$,

$$\begin{aligned}\mathcal{T}_2(\hat{t}_1^*(\hat{t}_3), \hat{t}_3, \hat{t}_3) =& 1 + \mu(\mathcal{M}^+(\hat{t}_3) - \underbrace{\mathcal{R}(\hat{t}_1^*(\hat{t}_3), \hat{t}_3, 2)}_{<\mathcal{M}^+(\hat{t}_3)}) - \hat{t}_3 > 0 \text{ and}\\ \mathcal{T}_2(\hat{t}_1^*(1), 1, \hat{t}_3) =& 1 + \mu(\mathcal{R}(\hat{t}_3, 1, 3) - M^+(1)) - 1 \leq 0.\end{aligned}$$

Therefore, for any $\hat{t}_3 \in (0, 1]$, a $\hat{t}_2$ exists. We conclude that $\mathcal{T}_2$ implicitly defines a function $\hat{t}_2^*(\hat{t}_3)$ with the properties $\hat{t}_2^*(\hat{t}_3) > \hat{t}_3$ if $\hat{t}_3 < 1$ and $\hat{t}_2^*(1) = 1$.

Similar steps show that functions $\hat{t}_j^*(\hat{t}_{j+1})$ with the properties $\hat{t}_j^*(\hat{t}_{j+1}) > \hat{t}_{j+1}$ if $\hat{t}_{j+1} < 1$ and $\hat{t}_j^*(1) = 1$ exist for all $j \in \{3, \ldots, K-2\}$. In a last step, we plug the function $\hat{t}_{K-2}^*(\hat{t}_{K-1})$ into $\mathcal{T}_{K-1}$;

$$\mathcal{T}_{K-1}(\hat{t}_{K-2}^*(\hat{t}_{K-1}), \hat{t}_{K-1}) \equiv 1 + \mu(\mathcal{R}(0, \hat{t}_{K-1}, K) - \mathcal{R}(\hat{t}_{K-1}, \hat{t}_{K-2}^*(\hat{t}_{K-1}), K-1)).$$

Now note that $\mathcal{R}(0, 0, K) = \mathbb{E}(t)$, $\mathcal{R}(0, \hat{t}_{K-2}^*(0), K-1) < \mathbb{E}(t)$, and $\mathcal{R}(0, 1, K) < \mathbb{E}(t)$.

Therefore,

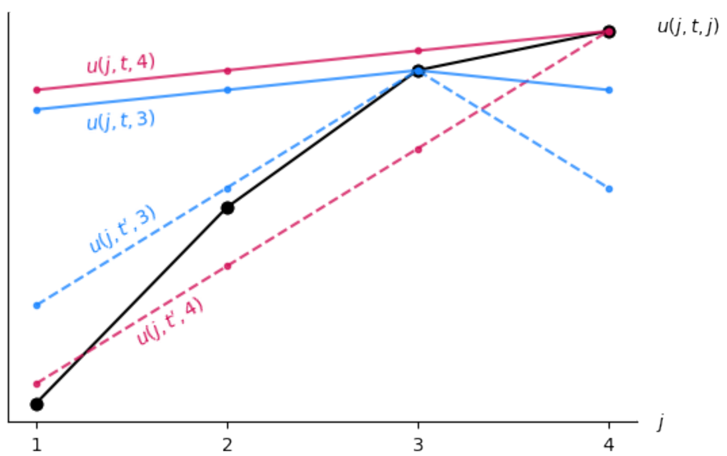
$$\mathcal{T}_{K-1}(\hat{t}^*_{K-2}(0), 0) = 1 + \mu(\mathbb{E}(t) - \mathcal{R}(0, \hat{t}^*_{K-2}(0), K-1)) - \hat{t}_3 > 0 \text{ and} \tag{18}$$

$$\mathcal{T}_{K-1}(\hat{t}^*_{K-2}(1), 1) = 1 + \mu(\mathcal{R}(0, 1, K) - M^+(1)) - 1 < 0. \tag{19}$$

This shows that a $\hat{t}_{K-1}$ exists for which $\mathcal{T}_{K-1}(\hat{t}^*_{K-2}(\hat{t}_{K-1}), \hat{t}_{K-1}) = 0$. The indifference conditions can thus be solved recursively: Find $\hat{t}^*_{K-1}$ for which $\mathcal{T}_{K-1}(\hat{t}^*_{K-2}(\hat{t}^*_{K-1}), \hat{t}^*_{K-1}) = 0$, plug this into $\hat{t}^*_{K-2}(\hat{t}_{K-1})$ to obtain $\hat{t}^*_{K-2}$, and so forth to finally obtain $\hat{t}^*_1(\hat{t}^*_2)$. The resulting $\hat{t}^*_j$ give the equilibrium vector of threshold types. To see that all of them are strictly smaller than one, note that a threshold $\hat{t}_j$ can only be equal to 1 if $\hat{t}_{j+1}$ is equal to 1. Since $\hat{t}^*_{K-1} < 1$ in equilibrium (by the strict inequality in Equation (18)), $\hat{t}^*_{K-2} < 1$ and therefore all remaining thresholds are also strictly smaller than 1. □

The equilibrium thus predicts that, in equilibrium, every state larger than 1 is reported by a liar with positive probability. The extreme prediction follows from the interaction between image concerns and increasing lying costs: With a nonzero image weight, the truthful reporting utility becomes strictly concave in the reported state. As a consequence, the marginal utility difference from reporting $j + 1$ over j becomes smaller as j increases. Therefore, it becomes optimal for some moral types to lie to report a smaller number than K. Figure 6 illustrates this dynamic by plotting the equilibrium utility received from truth-telling and lying for selected types in a game with $K = 4$. The black line plots the utility that agents receive from truthfully reporting j. The red lines plot the utility that agents of moral types t and t' receive when they report 4, with $t < t'$. Increasing t causes the red line to pivot downwards, as higher moral types face a steeper slope of the lying cost. The red lines also show that the type t prefers lying and reporting 4 to truth-telling after drawing a number smaller than 4 (the solid red line is above the black line). Type t' on the other hand only prefers lying to 4 over truth-telling after drawing 1. The blue lines in the figure plot the equilibrium utility that types t and t' receive after reporting 3. They show that type t prefers reporting 4 to reporting 3 after drawing any state (the solid red line is above the solid blue line) while t' prefers reporting 3 to reporting 4 after drawing a number smaller than 4 (the dashed blue line is above the dashed red line left of $j = 3$). There is also no downwards lying as no type will ever receive a utility value from lying downwards that is larger than the utility value from truth-telling.

Figure 6. Equilibrium with increasing lying costs and $K = 4$



Some experimental evidence exists that casts doubts on the equilibrium prediction of the increasing lying costs model. GK&S, for example, report that, in an observed lying game where participants can report numbers between 1 and 10, almost no individual dishonestly reports a number smaller than 9. In the observed game, only the composition effect of the character-based model is active but this alone is enough to predict an equilibrium in the observed game that has the same qualitative features as the one described in Proposition 5b. That is, it would predict that every state except for 1 is reported dishonestly with positive probability. This prediction, however, is not borne out in the data.

A.2 Heterogenous image concerns

While the paper assumed homogenous image concerns so far, papers such as Friedrichsen and Engelmann (2018) and Butera, Metcalfe, Morrison, and Taubinsky (2022) provide empirical evidence that different individuals care about their image to different extents. When this is the case, and when individuals anticipate heterogenous image concerns in others, behavior might change in meaningful ways. I briefly discuss the implications of heterogenous image concerns in the following.

Suppose agents hold an image concern which is drawn from a distribution $g(\mu)$ which has full support on $[0, \bar{\mu}]$ and which is independent of t. Partial lying will now arise as part of an equilibrium if there is a type with a sufficiently large image concern.

Proposition 6a. *When agents draw their image concern from a distribution $g(\mu)$ with full support between $[0, \bar{\mu}]$ and if $K > 2$, there is no equilibrium where liars only report K if $\bar{\mu}$ is sufficiently large.*

Proof. Suppose there is an equilibrium where liars only report K. Thus, agents must strictly prefer reporting K to reporting $K - 1$, conditional on lying. This implies an incentive constraint which is most binding for types with $\bar{\mu}$;

$$y(K) + \bar{\mu}\mathcal{R}_K \geq y(K-1) + \bar{\mu}\mathcal{R}_{K-1}.$$

Rearrange this to

$$\frac{\Delta(K, K-1)}{\bar{\mu}} \geq \mathcal{R}_{K-1} - \mathcal{R}_K.$$

Now, the equilibrium is such that all states smaller than K must have a reputation weakly larger than $\mathbb{E}(t)$, since agents reporting these states are in the right tail of the moral type distribution. By the martingale property of beliefs we know that, conversely, $\mathcal{R}_K < \mathbb{E}(t)$. Therefore, the right-hand side of the inequality above is strictly positive. There is a contradiction if $\bar{\mu}$ is sufficiently large. □

Partial lying thus still emerges under heterogenous image concerns but it will be of a slightly different kind. Remember how in the baseline analysis, liars are indifferent between 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 a high image payoff more than others, which leads them to strictly prefer partial to full lying. The resulting equilibrium is one where liars separate by their image type; as the following proposition shows, for an intermediate range of $\bar{\mu}$, the less image concerned liars report K while more image concerned liars report $K - 1$.

Proposition 6b. *Suppose that agents draw their image concern from a distribution $g(\mu)$ with full support between $[0, \bar{\mu}]$. For intermediate parameter values of $\bar{\mu}$ and for $K > 3$, there is an* *equilibrium which is characterized by threshold types $\hat{\mu} \in (0, \bar{\mu})$, $\hat{t}_{Kj}(\mu)$, and $\hat{t}_{K-1j}(\mu)$. Agents of type (j, t, μ) lie and report K if $\mu \leq \hat{\mu}$ and $t \leq \hat{t}_{Kj}(\mu)$. They lie and report $K-1$ if $\mu > \hat{\mu}$ and $t \leq \hat{t}_{K-1j}(\mu)$.*

Proof. In equilibrium, the moral type threshold of liars lying to K is given by

$$\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu) \equiv \Delta(K, j) + \mu(\mathcal{R}_K - \mathcal{R}_j).$$

Similarly, the threshold of lying to $K-1$ is equal to

$$\hat{t}_{K-1,j}(\mathcal{R}_K, \mathcal{R}_j, \mu) \equiv \Delta(K-1, j) + \mu(\mathcal{R}_{K-1} - \mathcal{R}_j).$$

Liars prefer reporting K over reporting $K-1$ if

$$y(K) + \mu\mathcal{R}_K \geq y(K-1) + \mu\mathcal{R}_{K-1}.$$

From this, we can derive the threshold image type $\hat{\mu}$ who is indifferent between reporting K or $K-1$;

$$\hat{\mu} = \frac{\Delta(K, K-1)}{\mathcal{R}_{K-1} - \mathcal{R}_K}. \tag{20}$$

The reputation of state j in equilibrium is then given by

$$\begin{aligned}
&\mathcal{R}_j = \mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}_j, \hat{\mu}, j) \equiv \\
&\frac{1}{\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu + \int_{\hat{\mu}}^{\bar{\mu}} \bar{F}(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu} \\
&\times \Big[\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu))\mathcal{M}^+(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu \\
&+ \int_{\hat{\mu}}^{\bar{\mu}} \bar{F}(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))\mathcal{M}^+(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu \Big].
\end{aligned}$$

Define a function

$$\rho_j(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}_j, \hat{\mu}) \equiv \mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}_j, \hat{\mu}, j) - \mathcal{R}_j.$$

In equilibrium $\rho_j(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}_j, \hat{\mu}) = 0$. Note that $\rho_j(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathbb{E}(t), \hat{\mu}) = \mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathbb{E}(t), \hat{\mu}, j) - \mathbb{E}(t) > 0$ and that $\rho_j(\mathcal{R}_K, \mathcal{R}_{K-1}, \hat{\mu}, \mathbb{E}(t)) = \mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \hat{\mu}, \bar{t}, j) - \bar{t} < 0$. Therefore, for any three parameters $\mathcal{R}_K, \mathcal{R}_{K-1}, \hat{\mu}$, we can always find a vector of equilibrium reputations $\mathcal{R}_1^*, \ldots, \mathcal{R}_{K-2}^*$ of the lower states consistent with it.

The equilibrium reputation of $K-1$ is equal to

$$\begin{aligned}
&\mathcal{R}_{K-1} = \mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}_{K-2}, \ldots, \mathcal{R}_1, \hat{\mu}) \equiv \\
&\frac{1}{\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu + \sum_{j \neq K-1} \int_{\hat{\mu}}^{\bar{\mu}} F(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu} \times \\
&\Big[\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,K-1}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mu))\mathcal{M}^+(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mu))g(\mu)\,\mathrm{d}\mu + \\
&\sum_{j \neq K-1} \int_{\hat{\mu}}^{\bar{\mu}} F(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))\mathcal{M}^-(\hat{t}_{K-1,j}(\mathcal{R}_{K-1}, \mathcal{R}_j, \mu))g(\mu)\,\mathrm{d}\mu \Big].
\end{aligned}$$

Note from Equation (20) that we can write the equilibrium $\hat{\mu}^*$ as a function of $\mathcal{R}_K$ and $\mathcal{R}_{K-1}$. Replacing $\hat{\mu}$ and the lower reputations yields

$$\rho_{K-1}(\mathcal{R}_K, \mathcal{R}_{K-1}) \equiv$$
$$\mathcal{R}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mathcal{R}^*_{K-2}(\mathcal{R}_K, \mathcal{R}_{K-1}, \hat{\mu}^*(\mathcal{R}_K, \mathcal{R}_{K-1})), \ldots, \mathcal{R}^*_1(\mathcal{R}_K, \mathcal{R}_{K-1}, \hat{\mu}^*(\mathcal{R}_K, \mathcal{R}_{K-1})), \hat{\mu}^*(\mathcal{R}_K, \mathcal{R}_{K-1}))$$
$$- \mathcal{R}_{K-1}.$$

In equilibrium $\rho_j(\mathcal{R}_K, \mathcal{R}_{K-1}) = 0$. Fix $\mathcal{R}_K$ at a value between 0 and $\mathbb{E}(t)$. Values that R_{K-1} can take on that are compatible with equilibrium are on $[\mathcal{R}_K + \Delta(K, K-1)/\bar{\mu}, \bar{t}]$. Evaluating ρ_{K-1} at these extreme values;

$$\rho_{K-1}(\mathcal{R}_K, \bar{t}) = \mathcal{R}(\mathcal{R}_K, \bar{t}, \mathcal{R}^*_{K-2}(\mathcal{R}_K, \bar{t}, \hat{\mu}^*(\mathcal{R}_K, \bar{t})), \ldots, \mathcal{R}^*_1(\mathcal{R}_K, \bar{t}, \hat{\mu}^*(\mathcal{R}_K, \bar{t})), \hat{\mu}^*(\mathcal{R}_K, \bar{t})) - \bar{t} < 0,$$
$$\rho_{K-1}(\mathcal{R}_K, \mathcal{R}_K + \Delta(K, K-1)/\bar{\mu}) = \frac{\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,K-1}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mu)\mathcal{M}^+(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_{K-1}, \mu))g(\mu)\,\mathrm{d}\mu}{\int_0^{\hat{\mu}} \bar{F}(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}_j, \mu)g(\mu)\,\mathrm{d}\mu} -$$
$$(\mathcal{R}_K + \frac{\Delta(K, K-1)}{\bar{\mu}}).$$

Since the first term on the r.h.s. in the equation above is always larger than $\mathbb{E}(t)$ and $\mathcal{R}_K < \mathbb{E}(t)$ in equilibrium, the r.h.s. is larger than zero as long as $\Delta(K, K-1)/\bar{\mu}$ is not too large. In this case, a solution exists such that $\rho_{K-1}(\mathcal{R}_K, \mathcal{R}^*_{K-1}(\mathcal{R}_K)) = 0$. Therefore, $\mathcal{R}_K$ pins down all remaining reputations $\mathcal{R}_{K-1}, \ldots, \mathcal{R}_1$ and also $\hat{\mu}$.

Lastly, we have to find a $\mathcal{R}^*_K$ so that $\rho_K(\mathcal{R}^*_K) = 0$, where

$$\rho_K(\mathcal{R}_K) \equiv$$
$$\frac{1}{\int_{\hat{\mu}^*}^{\bar{\mu}} \bar{F}(\hat{t}_{K-1,K}(\mathcal{R}^*_{K-1}, \mathcal{R}_K, \mu)g(\mu)\,\mathrm{d}\mu + \int_0^{\hat{\mu}^*} F(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}^*_j, \mu))g(\mu)\,\mathrm{d}\mu} \times$$
$$\left[\int_{\hat{\mu}}^{\bar{\mu}^*} \bar{F}(\hat{t}_{K-1,K}(\mathcal{R}^*_{K-1}, \mathcal{R}_K, \mu)\mathcal{M}^+(\hat{t}_{K-1,K}(\mathcal{R}^*_{K-1}, \mathcal{R}_K, \mu))g(\mu)\,\mathrm{d}\mu + \right.$$
$$\left.\sum_{j \neq K} \int_{\hat{\mu}^*}^{\bar{\mu}} F(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}^*_j, \mu))\mathcal{M}^-(\hat{t}_{K,j}(\mathcal{R}_K, \mathcal{R}^*_j, \mu))g(\mu)\,\mathrm{d}\mu\right] - \mathcal{R}_K.$$

Since this function is strictly larger than zero when evaluated at 0 and strictly smaller than $\mathbb{E}(t)$ when evaluated at $\mathbb{E}(t)$, a solution exists. □

Apart from predicting a separation by image type, the equilibrium above predicts downward lying: An highly image concerned agent will prefer honestly reporting $K-1$ over honestly reporting K. If their intrinsic lying cost is sufficiently low, they will thus also prefer dishonestly reporting $K-1$ after drawing K. More extreme versions of the downward lying prediction plausibly exist as $\bar{\mu}$ increases beyond the intermediate range. For example, in the most extreme case an agent with very high image concern (e.g. $\mu \to \infty$) has a strict incentive to report 1 after drawing K, even if K is large.

The heterogenous image concerns equilibrium in addition can rationalize report distributions where the modal report is smaller than K. This can happen if there is a positive correlation between an agent's moral type and image concern. Then, the liars who report K because they care little about their image are also those with the lowest moral types, and other liars will dislike pooling with them. If this motive is strong enough, more agents will report $K-1$ than K in order to avoid making the same report that the worst types make. Note that this prediction is exclusive to the character-based model: In such an equilibrium, reporting $K-1$ is more obviously a lie than reporting K, since more liars report $K-1$. Such an equilibrium would therefore be impossible in a deed based model. In the character-based model however, if liars reporting $K-1$ are of a higher moral type than those reporting K, such an equilibrium can be sustained through the composition effect.

There is evidence that the highest state is not always reported by most participants. For example, 8 out of 24 papers included in the AN&R meta-study that employ a one-shot die-roll lying game contain experiments where the highest state is not the modal report. Most of these experiments have been conducted outside of traditional lab environments in settings where the social distance between observer and participants is arguably lower and where the image motive thus might play a greater role. For example, **?** 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 at one weekday afternoon. They find that some soldiers lie to the army official and that most of them report the second-highest state.

B Example of a non-symmetric equilibrium

This section provides an example of an equilibrium where the reports of liars depends on their lying cost. Consider a setup with $K=3$ and the following strategy profile:

$$\begin{aligned} s(j|j,t) &= 1 \text{ if } j > 1, \\ s(3|1,t) &= 1 \text{ if } t \leq \hat{t}_a \\ s(2|1,t) &= 1 \text{ if } t \in (\hat{t}_a, \hat{t}_b] \\ s(1|1,t) &= 1 \text{ if } t \geq \hat{t}_b. \end{aligned}$$

That is, agents lie only if they draw 1. There are two quality segments of liars. Liars with the worst quality report the highest state and other liars report the middle state.

Assume preferences are uniformly distributed between zero and $T>0$. The equilibrium reputations are

$$\begin{aligned} \mathcal{R}_1^C(T,\hat{t}_b) &= \frac{T+\hat{t}_b}{2} \\ \mathcal{R}_2^C(T,\hat{t}_a,\hat{t}_b) &= \frac{1}{2} \times \frac{T^2+(\hat{t}_a+\hat{t}_b)(\hat{t}_b-\hat{t}_a)}{T+\hat{t}_b-\hat{t}_a} \\ \mathcal{R}_3^C(T,\hat{t}_a) &= \frac{1}{2} \times \frac{T^2+\hat{t}_a^2}{T+\hat{t}_a}. \end{aligned}$$

The equilibrium is characterized by two threshold values $(\hat{t}_a, \hat{t}_b)$ and two indifference conditions. The first is that the agent of type $(1, \hat{t}_b)$ must be indifferent between lying and truth-telling;

$$\begin{aligned} y(1) + \mu\mathcal{R}_1^C(T,\hat{t}_b) &= y(3) + \mu\mathcal{R}_3^C(t,\hat{t}_a) - \hat{t}_b \\ \Rightarrow \hat{t}_b &= \frac{1}{1+\mu/2}\left(\Delta(3,1) + \mu(R_3(T,\hat{t}_a) - T)\right). \end{aligned} \tag{21}$$

The second equilibrium condition is that liars must be indifferent between reporting states 2 and 3;

$$y(3) + \mu\mathcal{R}_3^C(T, \hat{t}_a) = y(2) + \mu\mathcal{R}_2^C(T, \hat{t}_a, \hat{t}_b). \tag{22}$$

Consider parameter values $\mu = 1$, $T = 7$, $y(1) = 0$, $y(2) = 4.9$, and $y(3) = 5$. We can plug Equation (21) into Equation (22) and solve for $\hat{t}_a$. The resulting parameter values are $\hat{t}_a \approx 1.12$ and $\hat{t}_b \approx 3.06$, which imply that each of the states are reported (from low to high) with frequencies 18.75%, 42.57%, and 38.68%. Note that in this example, the second-highest state is reported with a higher frequency than the highest state. In the symmetric equilibrium with homogeneous image concerns the reporting frequencies are monotonely increasing in j. Therefore, the example induces a different reporting frequency than the one induced by symmetric equilibrium.

C Remark: Upper bound on μ

The precise upper bound on μ will depend on the distribution function $F(t)$. Here we show that, if $F(t)$ is log-concave, $\frac{\mathrm{d}\hat{t}_j}{\mathrm{d}\varphi} < 1$ is sufficient (and therefore, e.g., $\mu \leq 1$ by Lemma 1). Suppose that this condition holds. A sufficient condition for $\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\varphi} < 1$ is that

$$\sum_{j \in \mathcal{K}} \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\hat{t}_j} \leq 1. \tag{23}$$

Taking derivatives, the sum term becomes

$$\begin{aligned}
\sum_{j \in \mathcal{K}} \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\hat{t}_j} =& \frac{\sum_{j \in \mathcal{K}}}{\left(\sum_{l \in \mathcal{K}} F(\hat{t}_l)\right)^2} \left[f(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) \sum_{l \in \mathcal{K}} F(\hat{t}_l) + F(\hat{t}_j)\mathcal{M}'(\hat{t}_j) \sum_{l \in \mathcal{K}} F(\hat{t}_l) - f(\hat{t}_j) \sum_{l \in \mathcal{K}} F(\hat{t}_l)\mathcal{M}^-(\hat{t}_l) \right] \\
=& \frac{\sum_{j \in \mathcal{K}}}{\sum_{l \in \mathcal{K}} F(\hat{t}_l)} \left[f(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) + F(\hat{t}_j)\mathcal{M}'(\hat{t}_j) - f(\hat{t}_j)\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) \right] \\
=& \frac{\sum_{j \in \mathcal{K}}}{\sum_{l \in \mathcal{K}} F(\hat{t}_l)} \left[f(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) + F(\hat{t}_j)\frac{f(\hat{t}_j)}{F(\hat{t}_j)} \left(\hat{t}_j - \mathcal{M}^-(\hat{t}_j)\right) - f(\hat{t}_j)\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) \right] \\
=& \frac{\sum_{j \in \mathcal{K}} f(\hat{t}_j)}{\sum_{l \in \mathcal{K}} F(\hat{t}_l)} \left[\hat{t}_j - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) \right].
\end{aligned}$$

We now show that condition (23) always holds if $F(t)$ is log-concave. By log-concavity of $f(t)$, $\mathcal{M}^{-\prime}(t) \in (0, 1)$. It follows that

$$\sum_{j \in \mathcal{K}} \frac{F(\hat{t}_j)}{\sum_{l \in \mathcal{K}} F(\hat{t}_l)} \mathcal{M}'(\hat{t}_j) < 1. \tag{24}$$

The inequality in (23) holds if it is smaller than the left hand size of (24);

$$
\begin{aligned}
\frac{\sum_{j\in\mathcal{K}} f(\hat{t}_j)}{\sum_{l\in\mathcal{K}} F(\hat{t}_l)}\left[\hat{t}_j - \mathcal{L}(\hat{\boldsymbol{t}}(\varphi))\right] &\le \sum_{j\in\mathcal{K}} \frac{F(\hat{t}_j)}{\sum_{l\in\mathcal{K}} F(\hat{t}_l)}\mathcal{M}'(\hat{t}_j) \\
&= \frac{\sum_{j\in\mathcal{K}} f(\hat{t}_j)}{\sum_{l\in\mathcal{K}} F(\hat{t}_l)}\left[\hat{t}_j - \mathcal{M}^-(\hat{t}_j)\right] \\
\Rightarrow \frac{\sum_{j\in\mathcal{K}} f(\hat{t}_j)}{\sum_{l\in\mathcal{K}} F(\hat{t}_l)}\mathcal{M}^-(\hat{t}_j) &\le \frac{\sum_{j\in\mathcal{K}} f(\hat{t}_j)}{\sum_{l\in\mathcal{K}} F(\hat{t}_l)}\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) \\
\Rightarrow \sum_{j\in\mathcal{K}} f(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) &\le \sum_{j\in\mathcal{K}} f(\hat{t}_j)\mathcal{L}(\hat{\boldsymbol{t}}(\varphi)) \\
&= \sum_{j\in\mathcal{K}} f(\hat{t}_j)\sum_{l\in\mathcal{K}} \frac{F(\hat{t}_l)}{\sum_{k\in\mathcal{K}} F(\hat{t}_k)}\mathcal{M}^-(\hat{t}_l) \\
\Rightarrow \sum_{j\in\mathcal{K}} f(\hat{t}_j)\sum_{l\in\mathcal{K}} F(\hat{t}_l)\mathcal{M}^-(\hat{t}_j) &\le \sum_{l\in\mathcal{K}} f(\hat{t}_l)\sum_{j\in\mathcal{K}} F(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) \\
\Rightarrow \sum_{j\in\mathcal{K}} f(\hat{t}_j)F(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) + \sum_{j\in\mathcal{K}}\sum_{l\neq j} f(\hat{t}_j)F(\hat{t}_l)\mathcal{M}^-(\hat{t}_j) &\le \sum_{j\in\mathcal{K}} f(\hat{t}_j)F(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) \\
&\quad + \sum_{j\in\mathcal{K}}\sum_{l\neq j} f(\hat{t}_l)F(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) \\
\Rightarrow \sum_{j\in\mathcal{K}}\sum_{l\neq j} f(\hat{t}_j)F(\hat{t}_l)\mathcal{M}^-(\hat{t}_j) &\le \sum_{j\in\mathcal{K}}\sum_{l\neq j} f(\hat{t}_l)F(\hat{t}_j)\mathcal{M}^-(\hat{t}_j) \\
\Rightarrow 0 &\le \sum_{j\in\mathcal{K}}\sum_{l\neq j} f(\hat{t}_l)F(\hat{t}_j)[\mathcal{M}^-(\hat{t}_j) - \mathcal{M}^-(\hat{t}_l)] \\
\Rightarrow 0 &\le \sum_{j\in\mathcal{K}}\sum_{l\in\mathcal{K}} f(\hat{t}_l)F(\hat{t}_j)[\mathcal{M}^-(\hat{t}_j) - \mathcal{M}^-(\hat{t}_l)].
\end{aligned}
$$

The inequality above holds if for any pair $j < l$

$$
f(\hat{t}_l)F(\hat{t}_j)[\mathcal{M}^-(\hat{t}_j) - \mathcal{M}^-(\hat{t}_l)] > f(\hat{t}_j)F(\hat{t}_l)[\mathcal{M}^-(\hat{t}_j) - \mathcal{M}^-(\hat{t}_l)].
$$

After rearranging, this inequality becomes

$$
\frac{f(\hat{t}_l)}{F(\hat{t}_l)} > \frac{f(\hat{t}_j)}{F(\hat{t}_j)},
$$

which holds by log concavity and because $\hat{t}_j > \hat{t}_l$.