International Journal of Game Theory manuscript No.
(will be inserted by the editor)

Lies and Consequences

The Effect of Lie Detection on Communication Outcomes

Ivan Balbuzanov

Received: date / Accepted: date

Abstract I study a strategic-communication game between an informed sender and an uninformed receiver with partially aligned preferences. The receiver is endowed with the ability to probabilistically detect if the sender is lying. Specifically, if the sender is making a false claim about her type, with some commonly known probability p the receiver additionally observes a private signal indicating that the sender is lying. The main result is that the receiver's stochastic lie-detection ability makes fully revealing equilibria—the best outcome for the receiver—possible, even for small p (less than $\frac{1}{2}$). Additionally, if the language consists of precise messages, fully revealing equilibria exist only for $p = 1$ and for a set of intermediate values of p that is bounded away from 0 and 1, making the maximal ex-ante expected equilibrium utility of the receiver non-monotone in p. If vague messages are allowed, full revelation can be supported for all large enough p, overturning the non-monotonicity and improving communication outcomes relative to the precise-language case.

Keywords cheap talk · persuasion game · lie detection
JEL Classification: D83

Department of Economics, The University of Melbourne. Email address: ivan.balbuzanov@unimelb.edu.au. Phone: +61(0)3 8344-9795. ORCID ID: 0000-0003-0093-9651. Postal address: L4 111 Barry St; Department of Economics; The University of Melbourne 3010 VIC; Australia.

1 Introduction

Lying convincingly is hard—crafting a believable false story is difficult, the party intent on misrepresenting their private information may fail, and, as a result, the receiver might become aware of the lie. This can happen if the sender allows inconsistencies to slip into her message, via verbal or non-verbal cues indicating deception (Vrij, 2008, Ekman, 2009), or via discrepancies discovered after an audit or investigation. Even though statements differ significantly depending on their truthfulness, a standard assumption in most models of strategic communication (starting with Crawford and Sobel, 1982; also see Sobel, 2013, for a recent survey) is that communication outcomes are independent of the content of the messages sent.

I show that endowing a receiver with a probabilistic ability to detect the sender's deception can lead to dramatic improvement in communication outcomes and to full information revelation in equilibrium. While it is natural to expect that a higher probability of detecting lies is always better for the receiver (e.g. see Duffy and Feltovich, 2006, Wang et al, 2010), I show that this intuition need not hold. A higher probability of detecting a lie may decrease the maximum ex-ante expected utility of the receiver while relatively low lie-detection probability may support full revelation in equilibrium.

I study a cheap-talk game with Crawford-Sobel (C-S) preferences and natural-language communication, so that all messages have intrinsic meanings and each message makes a precise statement about the sender's type (e.g. "I am type t"). I endow the receiver with the ability to probabilistically detect if the sender is lying. Specifically, if the sender's message does not match her type, with some commonly known probability p the receiver additionally observes a private signal indicating that the sender is lying. That signal carries no supplementary information. The signal is also not verifiable and therefore the receiver does not have access to any exogenous contractual punishment that he can invoke after detecting a lie. The only action that the receiver can take after detecting a lie must be sequentially rational for him.

The main result is that, in a large class of cases, stochastic lie detection is enough to guarantee the emergence of a fully revealing equilibrium (FRE) — the best outcome for the receiver. Thus, conditions for FRE existence are not as stringent as previously thought: in sender-receiver games, FRE are very rare and, conditional on existing at all, such equilibria are complex (e.g. see Golosov et al, 2014). Specifically, I show that a FRE exists for $p = 1$ and, if the preferences of the sender and the receiver are sufficiently aligned, also for a set of values of p that is bounded away from 0 and 1. In particular, this implies that the maximal ex-ante expected equilibrium utility of the receiver is non-monotone in p. Additionally, all FRE in my model are shown to be truthful: i.e. the sender sends the message corresponding to her type. If we allow the sender to make vague statements about her type (e.g. "My type is at least $\underline{t}$ and no greater than $\overline{t}$."), communication improves: the set of values of p that support a FRE is larger and is no longer bounded away from 1.

The intuition behind these results has to do with the potential "penalties" available to the receiver upon catching the sender in a lie. More specifically, the best way that the receiver can incentivize truth-telling is by maximizing the expected loss in utility that each lie carries. The expected loss faced by a sender falsely claiming to be type t is determined by three factors: the probability of detection p, the receiver's equilibrium action corresponding to that message (i.e. his most preferred action when the sender is indeed type t), and the receiver's "punishment" action in case he catches the sender in a lie. A low p would not be enough to dissuade a sender who wants to be mistaken for type t from lying as her lie would be caught too seldom. Symmetrically, a high p would provide incentives for a sender type to pretend to be type t *hoping to trigger the receiver's punishment action*. This is true even if the receiver chooses an extreme action after detecting a lie because in equilibrium that action must be sequentially rational and thus it is bounded for bounded type spaces, and so might be appealing to some sender types. Thus, only intermediate p can support a FRE. The described non-monotonicity disappears for vague messages since a vague message, by being true for more sender types, decreases the scope for this kind of perverse deviation.

Note again that the receiver does not have a private informative signal about the sender's type in the usual sense—the signal the receiver observes indicates only whether the sender's message matches her true type.[1] This is possible, for example, in settings where the state of the world is relatively complex to describe (e.g. it concerns the details of an event) and thus it is possible that a false account might include inconsistent details due to the sender's limited attention or memory. These inconsistencies could be used to catch the sender in a lie even if the receiver knows nothing about the sender's true type.[2]

As an example, we can think of the sender as a potential eyewitness (she) of a crime who is being interviewed by a member of the defense's legal team in a lawsuit (the receiver). The state of the world encodes how much of the crime the witness has knowledge of, with higher states meaning more relevant knowledge. The defense's action is choosing whether to call the witness to the stand and how much prominence to give to her testimony (by, for example, changing when she is called to testify, how many questions she is asked, and whether and how much of the closing argument is to be built on her testimony). The defense wants to give higher prominence to more knowledgeable witnesses (possibly because they are convinced of the defendant's innocence), while the witness prefers slightly more exposure than what the defense would prefer to

[1] For a treatment that features a privately informed receiver see, for example, Chen (2009), Lai (2014) or Ishida and Shimizu (2016).

[2] This justification was suggested in an early working-paper version of Dziuda and Salas (2018), which also provided an auxiliary model of "storytelling" with inconsistencies arising due to limited memory.

give her.[3] With some probability, a defense lawyer can determine that the witness is lying about what she knows by noticing inconsistencies between her narrative of the crime and facts known to the defense. The lawyer then chooses whether and how much to emphasize the witness' testimony. As there are no legal or reputational penalties for deception in this setting, the punishment for lying must be endogenous—the defense chooses an action that is best suited for the sender's expected true type.[4]

A potential objection to this model regards the inability of the sender to intentionally and with certainty trigger a lie detection, which might be beneficial in some of the equilibria discussed below. However, even if the sender crafts her message intending to be found out in a lie by, say, contradicting existing facts, she could not be certain that the receiver would know those facts or spot the discrepancy. In the motivating example above, the eyewitness could try to appear as if she is lying by providing false details about the crime. Still, she cannot be sure that those details would be known to the defense legal team and they might remain unaware of her deception. I discuss this further in Section 5.

The paper is organized as follows. Section 2 presents the main model. Section 3 presents the results. Section 4 considers some extensions of the basic model. Section 5 discusses potential interpretations and justifications of the main assumptions of the model, including the extensive psychological evidence that people are able to recognize lying in face-to-face interactions. Section 6 briefly surveys the related literature. Section 7 concludes. The proofs omitted from the main text are collected in the Appendix.

2 Model

A (female) sender privately observes her type t, which is drawn from a differentiable distribution F with full support on the type space $\mathcal{T} = [0, 1]$. She then sends a message $m \in \mathcal{M}$, where $\mathcal{M}$ is the message space, to a (male) receiver who takes an action $y \in \mathbb{R}$. I impose an intrinsic meaning to the available messages by setting $\mathcal{M} = \mathcal{T}$.[5,6] As in Crawford and Sobel (1982), the sender and the receiver are endowed with continuous von Neumann-Morgenstern utility

[3] This could be due to the *misinformation effect* (Loftus and Hoffman, 1989), which causes an eyewitness to believe that she has a better memory of the crime than she actually does.

[4] In some of the FRE constructed below, the off-equilibrium "punishment" action of the lawyer would be to call to the stand and highlight the testimony of a witness who is found out to be falsely pretending to have little knowledge, and to de-emphasize (or not even call to the stand) if she is caught falsely pretending to be very informed of the crime.

[5] Below, I consider an alternative version of the model which allows for vague messages as in Milgrom (1981). I also consider refinements which require equilibria to be robust to the creation of new messages with particular meaning.

[6] Thus, messages' meanings, unlike in standard cheap-talk games, are not determined in equilibrium. The meaning of message m is fixed to be "The true state of the world is $t = m$." Messages might, however, *lose* their meaning in equilibrium. See footnote 10. Another departure from the previous literature is the (realistic) assumption that all messages

functions that I denote by $U^S(y,t;b)$ and $U^R(y,t)$ respectively, where b can be viewed as a measure of the discrepancy between the preferences of the receiver and the sender.[7] Talk is cheap: i.e. utility does not depend on the message sent. I concentrate on the following parameterization of the C-S model with constant[8] additive upward bias:

$$U^S(y,t;b) = -\ell(t+b-y), \text{ and}$$
$$U^R(y,t) = -\ell^R(t-y),$$

where $\ell, \ell^R : \mathbb{R} \to \mathbb{R}$ are convex (and strictly so for ℓ^R) even functions (i.e. $\ell(x) = \ell(-x)$ for all $x \in \mathbb{R}$) that are strictly increasing on $\mathbb{R}_+$ (respectively, decreasing on $\mathbb{R}_-$). Without loss of generality, I assume that $\ell(0) = \ell^R(0) = 0$ and $b > 0$. Note that the optimal actions for the receiver and for the sender are $y = t$ and $y = t + b$ respectively.

If the sender is lying (i.e. if her message does not correspond to her type), the receiver may observe a private signal indicating that the sender's message is not truthful. The receiver observes that signal with a commonly known and exogenously given probability p. The sender does not know whether the receiver will be able to detect lying when sending the message. More precisely, if the sender chooses a message that doesn't correspond to the true state of the world (i.e. $m \neq t$), the receiver observes a private signal $v = -1$ with probability p. In all other cases, he observes $v = 0$.

Let $M : \mathcal{T} \to \Delta\mathcal{M}$ denote the message strategy of the sender, while $Y : \mathcal{M} \times \mathcal{V} \to \Delta\mathbb{R}$ denotes the action strategy of the receiver, where $\mathcal{V} = \{-1, 0\}$ is the space of possible truthfulness signals the receiver may observe. I consider only pure-strategy equilibria[9] so $M(t)$ will denote the message sent by type t in equilibrium and $Y(m,v)$ will denote the action taken by the receiver when he receives a message m and observes a signal v. I investigate (pure-strategy) perfect Bayesian equilibria $\{M, Y\}$ of the model.

Note that the sender and the receiver cannot write contracts that condition on v. Thus, the receiver has no recourse to an exogenous punishment if he catches the sender in a lie. Instead, the receiver must form an appropriate belief regarding the sender's type based on the false message and then take a sequentially rational action. Off the equilibrium path, the receiver chooses an action $Y(m,v) \in [0,1]$ for all $m \in \mathcal{M}, v \in \mathcal{V}$.

In most of the sequel, I focus on *fully revealing equilibria* (henceforth FRE). I define FRE as equilibria in which the receiver perfectly learns the sender's

are available to all agents, thus decoupling messages' meanings from the set of types they are available to.

[7] The argument b is often suppressed in what follows.

[8] In section 4.4, I consider the case of variable sender bias and show that the main result is preserved.

[9] The assumption that both the sender and the receiver play pure strategies is with little loss of generality. As ℓ^R is strictly convex, the receiver has a unique best response to any belief about the sender's type he might have. Also, in all fully revealing equilibria, which constitute the focus of this paper, all sender types (except possibly $t = 0$) play pure strategies.

type after observing any message-signal pair (m, v) that can be induced by a message sent by some sender type on the equilibrium path. In other words, any message-signal pair that can be observed in a FRE can possibly be generated by only one type of the sender.

3 Results

3.1 Preliminaries

The model reduces to a standard C-S cheap-talk game for $p = 0$. For the intermediate values of p, there is a plethora of equilibria. For example, it is easy to show that all the standard C-S equilibrium outcomes—including the uninformative babbling equilibria—remain as equilibrium outcomes with lie detection, as long as $p < 1$.[10,11] However, for $p = 1$ the model is equivalent to a persuasion game (Milgrom, 1981, Grossman, 1981) in the sense that the unique equilibrium outcome is full revelation. An implicit assumption in persuasion games is that lying is always detected and it carries a punishment that is harsh enough to dissuade any sender type from reporting falsely—loss of sales revenue or of reputation, for example, or the threat of a fine or imprisonment if misrepresentation is illegal.[12] This is also true in my model in the case of $p = 1$ since, due to the sender's upward bias, all sender types prefer to report truthfully rather than to face the action $y = 0$. In other words, whenever $p = 1$ there arises a punishment that is sufficient to deter lying.

Proposition 1 *When $p = 1$, all equilibria are FRE.*

Proof It is clear that $M(t) = t$, $Y(m, 0) = m$, and $Y(m, -1) = 0$ for all $m \in [0, 1]$ is an equilibrium strategy profile since any sender type t (weakly) prefers truth-telling and the associated receiver's action $y = t$ over being caught lying and the associated receiver's action $y = 0$.

To show that full revelation is the unique equilibrium outcome, assume that there is an equilibrium in which more than one sender type sends the same (false) message $\overline{m}$. This implies we can find two such types $t_0 < t_1$ such that $Y(\overline{m}, -1) \in (t_0, t_1)$. But then type t_1 does better by revealing her type by

[10] For example, consider a C-S equilibrium where two actions are induced on the equilibrium path: y_1 whenever the sender's type is $t < t^*$ and y_2 when $t \geq t^*$. For any $m^* \neq 0$, the messaging strategy $M(t) = 0$ if $t < t^*$ and $M(t) = m^*$ otherwise, together with the action strategy $Y(m^*, 0) = Y(m^*, -1) = y_2$ and $Y(m, -1) = Y(m, 0) = y_1$ for $m \neq m^*$ is an equilibrium for $p < 1$ and induces the same outcome.

[11] In another equilibrium that arises in this model, the high types (all t above some threshold $t^*(p)$) separate by sending truthful messages, while the low types $t < t^*(p)$ pool together. Incidentally, it can be shown that $\lim_{p \to 1} t^*(p) = 0$, so this class of equilibria converges to the unique equilibrium outcome at $p = 1$ as established by Proposition 1. The proof of this claim is available upon request.

[12] Following Milgrom (1981), persuasion-game models tend to allow the sender to be vague and send messages that correspond to subsets of the type space. I study this extension in 4.1.

sending the message $m = t_1$ and inducing the action $y = t_1$ (this action is the only possible sequentially rational response to observing $m = t_1$ and $v = 0$), which she prefers over $Y(\overline{m}, -1)$ due to her upward bias.

A corollary of the proposition is that there is an equilibrium discontinuity at $p = 1$: the large equilibrium-outcome set for any $p < 1$ collapsing to a singleton at $p = 1$ implies that the equilibrium-outcome correspondence is not upper hemicontinuous at $p = 1$. The main reason is that, while perfect Bayesian equilibrium imposes little discipline on the receiver's off-equilibrium actions when $p < 1$ allowing for the existence of a large range of equilibria, it severely restricts those actions as soon as $p = 1$. Namely, when $p = 1$, the solution concept necessitates $Y(m, 0) = m$ as only a sender of type $t = m$ can send the message m accompanied by $v = 0$. This now provides sufficient incentives for high types to deviate to their corresponding truthful message causing unraveling from the top (as in Milgrom, 1981) that only FRE survive.

Note that Proposition 1 would still hold if we allowed for vague messages—i.e. if we allowed the message space to be a subset of the power set $2^{[0,1]}$, such that it includes all singleton sets with the understanding that a sender of type t is truthful when sending a message $m \subseteq [0, 1]$ if and only if $t \in m$. I analyze one such message space in Section 4.1.

Additionally, note that the truthful messaging strategy $M(t) = t$ is not the only possible sender's strategy in a FRE for $p = 1$. Another possibility is $M(t) = t$ for all $t \in (0, 1]$ and $M(0) = m^* \neq 0$ with the receiver having the same action strategy as in the proof of the Proposition.[13] The following lemma shows that these are the only possible equilibrium messaging strategies for $p = 1$. More importantly, it establishes that any FRE for $p \in (0, 1)$ is necessarily truthful. Indeed, assume that all types on some interval separate in an equilibrium. If any one of them (other than the lowest type)—say type t^*—is sending a false message, then types "immediately below" t^* have an incentive to deviate and send $M(t^*)$ instead.

Lemma 1 *In any equilibrium that is fully separating on some interval I for some $p \in (0, 1]$, all types $\{t \in I : t > \inf I\}$ are truthful (i.e. $M(t) = t$). Furthermore, all types are truthful in all FRE for all $p \in (0, 1)$.*

Proof Say some interval (t_l, t_h) achieves full separation in equilibrium and assume that some type $t^* \in (t_l, t_h)$ sends a message different from $m = t^*$ in equilibrium. Then by full separation we must have

$$Y(M(t^*), -1) = t^* \text{ and, if } p < 1,\ Y(M(t^*), 0) = t^*.$$

Then there exists some $\tilde{t} \in (t_l, t^*)$ with $\tilde{t} \neq M(t^*)$ (implying that $\tilde{t}$ can perfectly mimic t^*) such that $\tilde{t} + b > t^* > \tilde{t}$. Therefore type $\tilde{t}$ can profitably deviate by sending $M(t^*)$ rather than $M\left(\tilde{t}\right)$.

[13] The use of the message m^* by both $t = 0$ and $t = m^*$ does not prevent full revelation. Whenever m^* is received, the receiver can distinguish perfectly between the two possible sender types based on the value of v.

This result extends in a straightforward manner to intervals that are closed on the right. So in a FRE for $p \in (0,1)$, all types in $(0,1]$ would be truthful, which leaves only the message $m = 0$ for type $t = 0$.

I end with a result which allows me to characterize the degree of bias that could allow the existence of a FRE.

Lemma 2 *Whenever $b > 1/4$ and $p < 1$, there does not exist a FRE.*

In fact, in what follows I show that $b \leq 1/4$ is not only necessary but a sufficient condition for the existence of a FRE for some values of $p < 1$.

3.2 Existence of FRE in the Quadratic Case

In this section, I establish the main result by showing the existence of a FRE in the leading quadratic-loss utility example of the C-S model, where I can also provide exact bounds on the interval of values of p that can support a FRE. For the following analysis, it is sufficient to have only the sender's utility be based on a quadratic-loss function (i.e. $\ell(x) = x^2$), while any strictly convex loss function is permitted for the receiver. The only requirements on the distribution F are full support and differentiability.

Proposition 2 *In the case of the sender having quadratic-loss preferences, a FRE for $p \in (0,1)$ exists if and only if $b \leq 1/4$ and*

$$p \in \left[\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}; \frac{1}{2} + \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2} \right].$$

Remarkably and counterintuitively, since full revelation is the best outcome from the point of view of the receiver, Proposition 2 implies that the maximum ex-ante expected equilibrium utility of the receiver is non-monotone in p. To see why this is the case, we can think of the receiver as a principal, designing appropriate incentives so that it is an equilibrium action for any sender type to report truthfully. The belief that the receiver forms when he detects lying have to rationalize an action that discourages the sender from deviating from the truth-telling strategies prescribed by the FRE. In other words, he has to "hurt" the sender sufficiently when a lie is detected. With p given, the best the receiver can do is to maximize the distance between the action $Y(m,0) = m$ and the "punishment" action $Y(m,-1)$.

The punishment action also has to be sequentially rational for the receiver—it has to be a best response to some belief that is consistent with receiving the off-equilibrium message-signal combination $(m,-1)$ so it has to be in the interval $[0,1]$. So if the receiver catches a sender claiming to be a low type in a lie, he assumes that she is the highest possible type (the worst belief if her bliss point is in fact low); and vice versa: if the receiver catches a sender claiming to be a high type in a lie, he assumes that she is the lowest possible

type (the worst belief if her bliss point is in fact high). Clearly, whenever p is small, this would not be sufficient to deter the sender type with bliss action $Y = m$ from deviating to m.

Because of sequential rationality, any punishment $Y(m, -1)$ might be a desirable action for some sender types and, whenever p is high, they would want to deviate to the false message m hoping to trigger the punishment. So if a "low" message m has a "high" penalty action $Y(m, -1)$, it would be a desirable lie for "high" types (e.g. the sender types around $t = Y(m, -1) - b$). If the penalty action is "low", however, that wouldn't provide enough incentives for "low" types (around $t = 0$) to report truthfully. Thus, it is only for intermediate values of p (including an open neighborhood of $p = \frac{1}{2}$ if $b < \frac{1}{4}$) that a receiver can "implement" off-equilibrium beliefs that deter the sender from lying.

The assumption that the sender has type-dependent preferences is crucial, as it allows the Receiver to "tailor" message-specific punishment actions that support FRE. To see that, consider a simple case with type-independent preferences, in which all Sender types prefer higher actions. The worst (sequentially rational) punishment for all types would be the action $Y = 0$ but that action would not suffice to deter the low types from deviating.[14]

Furthermore, due to the form of the receiver's preferences (strictly concave), he cannot provide extra deterrence by randomizing after catching the sender in a lie. However, without the requirement of sequential rationality, if the receiver can credibly commit to take a sufficiently high (or low) action in the face of a lie, then a FRE would be sustainable for all p. Alternatively, if p is too high to sustain a FRE but the receiver can commit to disregard and not act upon his private signal sufficiently often, this would decrease the effective value of p and thus make a FRE sustainable.

It is notable that there are FRE even when p is relatively small—smaller than 1/2. The minimum value of p sufficient for the existence of a FRE is smaller the smaller b is. The intuition behind this effect is that a smaller b is equivalent to a larger type space and a larger type space provides more extreme actions to serve as harsher penalties in case of lying and thus strengthens incentives for truth-telling. As the length of the type space diverges to infinity (or, equivalently, as the magnitude of the bias b converges to zero), the characteristic function of the set of values of p for which we have full revelation converges pointwise to the characteristic function of $(0, 1]$. To see that, it is relatively easy to modify the proof of Proposition 2 to show that if $Y(m, -1) > b/\sqrt{p(1-p)} + m$, no type would want to falsely send the message m. As the length of the type space increases, these messages become available for more and more values of p. At the limit, a type space that is unbounded from above permits FRE for all $p > 0$ (cf. Kartik et al 2007).

Conversely, the higher b is, the smaller the interval of values of p for which there exists a FRE is. In other words, the higher the mis-alignment in the

[14] I am grateful to an anonymous referee for urging me to clarify the importance of this point.

preferences of the two parties is, the harder it is to sustain a FRE. At $b = 1/4$—the largest value of b, for which there exists a FRE for an internal p—a FRE exists if and only if $p = 1/2$. Obviously, when $b = 0$, a FRE exists for all values of p: whenever the preferences of the sender and the receiver are fully aligned, there always is an equilibrium in which the sender truthfully reports her type.

3.3 Existence of FRE in the General Case

In this section, I relax the assumption of quadratic-loss preferences and consider the general case of an arbitrary convex loss function for the sender. I show that the gist of Proposition 2 holds for all possible utility functions of the sender of the form $U^S(y, t; b) = -\ell(t + b - y)$, as defined in Section 2.

Proposition 3 *For the general form of the sender's utility and for $b \leq 1/4$, the set of values of p in $(0, 1)$, for which there exists a FRE, includes $p = 1/2$ and is bounded away from 0 and 1. Additionally, if $b < 1/4$ the set of those values has positive Lebesgue measure.*

Remarkably, any degree of curvature in ℓ is sufficient to guarantee existence of a FRE. The following example suggests that, in order to support a FRE, having single-peaked preferences is more important than having preferences that are everywhere risk averse. The example assumes that ℓ is piecewise concave, rather than convex everywhere (so that the preferences of the sender on either side of her bliss point are piecewise risk-loving), and demonstrates that there still exists a FRE for an open neighborhood of values of p around $1/2$.

Example 1 Fix $b = 1/8$, $p = 1/2$, and $\ell(x) = x^{0.9}$. Assume that the receiver takes the following actions in equilibrium:

$$\begin{aligned} Y(m, -1) &= 1 \text{ if } m < 1/2; \\ Y(m, -1) &= 0 \text{ if } m \geq 1/2; \text{ and} \\ Y(m, 0) &= m. \end{aligned}$$

One can show that if no sender type prefers sending a message $m = 1/2$ over truth-telling, then all sender types prefer truth-telling over any false message (and so a FRE can be sustained). Since the preferences are piecewise convex, one can show that the type who would most benefit from deviating to a false $m = 1/2$ message is type $t = 1/2 - b = 3/8$. Her utility from sending that message is $\frac{1}{2}U^S(0, 3/8) = -0.27$, while the utility from being truthful is $U^S(3/8, 3/8) = -1/8^{0.9} = -0.15$. Thus, she prefers being truthful and these off-equilibrium actions of the receiver support a FRE.

Proposition 3 does not directly generalize to the case of piecewise strictly concave ℓ. By modifying its proof using the idea from Example 1, however, it is straightforward to show that for every piecewise strictly concave ℓ there exists some $\overline{b} \in (0, 1/4)$ such that the conclusions of Proposition 3 hold for all $b \leq \overline{b}$.

4 Extensions

4.1 Vague Messages

In the model above I restrict the meaning of all messages to be "I am exactly type t." It is natural to expect that there might be situations where the sender might refuse to send any message (Farrell and Rabin, 1996), removing the receiver's opportunity to observe an informative private signal. More generally, the sender might want to send a vague message as in Milgrom (1981) claiming to be within a set of types rather than a particular type. Both of these possibilities can be accommodated by expanding the message space to include all closed subintervals of $[0,1]$, including the entire type space $[0,1]$ and all singletons. I denote this new and expanded message space by $\mathcal{M}^V$. Formally:

$$\mathcal{M}^V = \{[a,b] | a,b \in [0,1], a \leq b\},$$

with the convention that $[a,a] = \{a\}$.[15]

Modifying the receiver's private signal regarding the sender's truthfulness is straightforward. With probability p, he observes the signal $v = -1$ if and only if the sender is lying (in this case this means $t \notin m \subseteq [0,1]$). In all other cases, he observes the signal $v = 0$.

In addition to the usual conditions on the off-equilibrium actions (namely, $Y(m,v) \in [0,1]$), sequential rationality also requires that the receiver correctly reasons about which sender types are possible given a particular (off-equilibrium) message-signal combination. More precisely, the receiver's belief cannot place positive probability on any subset of m if she observes $(m,-1)$. Formally:

$$Y(m,-1) \in [0, \inf m] \text{ for all } m \in \mathcal{M}^V \text{ with } 1 \in m; \text{ and}$$
$$Y(m,-1) \in [\sup m, 1] \text{ for all } m \in \mathcal{M}^V \text{ with } 0 \in m.$$

These conditions are relaxed slightly to allow the receiver, after observing a private signal indicating a lie, to take an action that is on the boundary of the set that his private signal indicates as impossible. For example, if the out-of-equilibrium message and signal are $m = [0,1/2]$ and $v = -1$, the only possible belief that the receiver can have after such an observation is some distribution with support in the set $(1/2,1]$. Thus, his best response to that belief must be in the set $(1/2,1]$. The conditions above allow the receiver to take the action $Y([0,1/2],-1) = 1/2$ as well. This is a technical assumption necessary to guarantee the existence of an equilibrium for $p = 1$ (Proposition 1) and plays no further role in the subsequent analysis.

Adding vague messages significantly increases the message space and, as outlined above, sequential rationality naturally restricts the actions that a receiver can take after observing these messages together with his v. It is then

[15] The space of vague messages can also be taken to be a larger (even possibly improper) subset of the power set of $[0,1]$. Expanding $\mathcal{M}^V$ in this way should not affect Proposition 4, the main result of this section.

conceivable that expanding the message space in this manner might decrease the incentives of sender types to play according to the strategy prescribed in some FRE. This intuition turns out to be incorrect: in fact, the values of p that support a FRE expand when we allow for vagueness. First, I prove a version of Lemma 1 in the case of a vague message space.

Lemma 3 *If the sender can send vague messages (i.e. if the available message space is $\mathcal{M}^V$ rather than $\mathcal{M}$), in any equilibrium that is fully separating on some interval I for some $p \in (0,1]$, all types $\{t \in I : t > \inf I\}$ are truthful and send a message $m \in \mathcal{M}^V$ with $\inf m = t$.*

Proof Say some type t^* in a fully separating interval (t_l, t_h) does not send a message m with $\inf m = t^*$. I consider the two possible cases separately: either the message $M(t^*)$ sent is false ($t^* \notin M(t^*)$) or the message is true but $\inf m < t^*$.

If the message is false we must have

$$Y(M(t^*), -1) = t^* \text{ and, if } p < 1,\ Y(M(t^*), 0) = t^*.$$

There exists some $\tilde{t} \in (t_l, t^*)$ with $\tilde{t} \notin M(t^*)$[16] we have $\tilde{t}+b > t^* > \tilde{t}$. Therefore type $\tilde{t}$ can profitably deviate by sending $M(t^*)$ rather than $M(\tilde{t})$.

If the message is true but $\inf m < t^*$, some type $t \in (\inf M(t^*), t^*)$ satisfies $t + b > t^* > t$ and therefore she can profitably deviate by sending the same message as t^*.

The lemma states that any sender type t in any fully separating interval sends a truthful message and, furthermore, t is the lowest possible sender type who can send that message truthfully. This result is analogous to the messaging strategy of the sender in a persuasion game. Note that the lemma does not put any constraints on the message that type 0 sends in FRE. I proceed by proving a version of Propositions 2 and 3 with vague messages.

Proposition 4 *In the general case of convex loss utility (i.e. $U^S(y,t;b) = -\ell(t+b-y)$) with $\mathcal{M}^V$ as the message space, a FRE exists for some p if either*

(i) $p \geq \frac{1}{2}$, or
(ii) a FRE exists for the same p in the case without vague messages (Proposition 3).

Furthermore, in the special case of the sender having quadratic-loss preferences (i.e. $U^S(y,t;b) = -(t+b-y)^2$), the disjunction of conditions (i) and (ii) is not only sufficient but also necessary for the existence of FRE. In other words, with quadratic utility a FRE exists if and only if $b \leq 1/4$ and

$$p \in \left[\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}; 1\right].$$

[16] Since $\mathcal{M}^V$ is comprised of closed subintervals of $[0,1]$ and $t^* \notin M(t^*)$, all types t sufficiently close to t^* also satisfy $t \notin M(t^*)$

It is remarkable that adding vague messages can improve communication outcomes. To see why, consider the incentives of some sender type t to deviate from her FRE-prescribed strategy and pretend to be type θ by sending the message $M(\theta)$. Without vagueness, this would induce a lottery in the actions of the receiver, where she would choose θ with probability $1-p$ and $Y(M(\theta),-1)$ with probability p. If $\theta < t < t+b < Y(M(\theta),-1)$, this might be preferable to type t over her equilibrium payoff. With vague messages though, such a deviation might be impossible if $t \in M(\theta)$. Then deviating to the message $M(\theta)$ would induce the action $y = \theta$ with probability 1, which is worse for type t than her equilibrium payoff. Thus, a given "punishment" action might induce a deviation by a higher type without vagueness, but not under the vague message space $\mathcal{M}^V$.

The proof of Proposition 4 relies on the explicit construction of a fully revealing messaging strategy that can be supported as part of a FRE for all $p \geq \frac{1}{2}$. The message that a type $t > 0$ sends under this strategy is $[t, 1-\varepsilon_t]$ for some $\varepsilon_t \geq 0$ with $1-\varepsilon_t \geq \max\{t, 4b\}$ and $\varepsilon_t > 0$ for $t < 2b$. Making the equilibrium messages vague decreases the set of sender types who would have a profitable deviation to lying in the hope of triggering a "punishment" action. For a simple illustration of that approach, consider the following example, which constructs a FRE for the case of quadratic preferences and for values of p higher than those given in Proposition 2 (albeit not all $p \geq \frac{1}{2}$).

Example 2 Assume quadratic sender's preferences. In the proof of Proposition 2, I show that the following off-equilibrium actions support full revelation for the highest p that permits the existence of a FRE:

$$Y(M(t),-1) = \begin{cases} 1 & \text{if } t < 2b, \\ 0 & \text{if } t \geq 2b. \end{cases}$$

The reason that we cannot do better is that under a higher p, some type t^* would have a profitable deviation to a message $M(t)$ for t close to but smaller than $2b$. If $t^*(t)$ is the sender type that maximizes the utility from deviating to the message $M(t)$, setting $M(t) = [t, t^*(t)+\varepsilon]$ for all t close to $2b$ would make such deviations impossible for values of p above the upper limit from Proposition 2.

Without vagueness, Lemma 1 states that all messages are used in a FRE. It might appear that the multitude of messages that are never sent in equilibrium under $\mathcal{M}^V$ might be troublesome as they offer more opportunities for deviation for the sender (and, unlike the usual case, the receiver's off-equilibrium actions can be severely restricted whenever $v = -1$). This is, however, a red herring. Continuing with the example, consider the following actions for all off-equilibrium messages $m \in \mathcal{M}^V \setminus M(\mathcal{T})$:

$$\begin{aligned} Y(m,0) &= 0; \\ Y(m,-1) &= 0 \text{ if } 0 \notin m; \text{ and} \\ Y(m,-1) &= 1 \text{ if } 0 \in m. \end{aligned}$$

No sender type t strictly prefers the certain action $y = 0$ (induced by deviating to an off-equilibrium truthful message or an off-equilibrium false message with $0 \notin m$) to the corresponding equilibrium action $Y(M(t), 0) = t$. Similarly, no sender type t would deviate to an off-equilibrium false message with $0 \in m$, which induces the lottery over the action $y = 0$ with probability $1-p$ and the action $y = 1$ with probability p, since that lottery is weakly worse than lotteries available to that type in the FRE without vagueness under the off-equilibrium actions above. Consulting the proof of Proposition 2, it is not hard to see that no agent would want to deviate to such an extreme lottery even for values of p above the upper limit from Proposition 2, thus guaranteeing the existence of a FRE with vagueness for those values.

4.2 Equilibrium Selection

Propositions 2, 3 and 4 provide sufficient (and necessary in the case of Propositions 2 and 4) conditions for FRE existence. Since, as noted above, the model supports a variety of equilibria for every value of $p < 1$, the natural next step is to ask whether it is reasonable to expect to see a FRE being played for the relevant values of p.

In this section, I start by considering whether FRE is consistent with the most well known cheap-talk equilibrium refinements. I show that all FRE survive the demanding *strong announcement-proofness* refinement of Matthews et al (1991) (regardless of the value of p), as well as Farrell's (1993) *neologism-proofness* and Chen et al's (2008) *NITS (no incentives to separate)* refinements, both of which are weaker than strong announcement-proofness. These refinements are based on the possibility of some sender types making a *credible* announcement about their private information and the receiver's best-response action induced by that announcement being preferred by the deviating types to the action they induce in equilibrium. An equilibrium fails the corresponding refinement if such a credible announcement exists.

The three refinements differ in what constitutes a credible announcement, how complex that announcement is allowed to be, and how large the set of deviating types can be. To be more precise, I first need to adapt Matthews et al's (1991) terminology and notation to my setting. Call any pair $\langle M', X \rangle$ an *announcement*, where X is a non-empty subset of the type space $\mathcal{T} = [0, 1]$ and M' is an alternative messaging strategy $M' : X \to \Delta\mathcal{M}'$ for some (sufficiently rich) message space $\mathcal{M}'$. The old and new message spaces are disjoint: $\mathcal{M} \cap \mathcal{M}' = \varnothing$. In other words, the messages in $\mathcal{M}'$ can be thought of as newly created statements (with meanings determined by the announcement).[17] Now let $Y' : \mathcal{M}' \to [0, 1]$ be the receiver's best response to the belief induced

[17] Farrell (1993) discusses at length when it can be reasonable to assume that neologisms can be created. His view is that this is particularly appropriate whenever the sender and the receiver have a natural language in common, which is true in the setting considered here.

by the announcement $\langle M', X\rangle$ and Bayes' rule.[18] Formally, Y' depends on the exact announcement but, for notational simplicity and with little risk of creating confusion, I suppress this in what follows.

An equilibrium (M, Y) is *strongly announcement-proof* if there does not exist an announcement $\langle M', X\rangle$ that satisfies the following three conditions:

C1. Every type in X prefers making the announcement over her equilibrium message: $U^S(Y'(m), t) \geq U^S(Y(M(t)), t)$ for all $t \in X$ and for all m in the support of $M'(t)$, with the inequality strict for some $t \in X$ and some m.
C2. Every type not in X prefers her equilibrium message over trying to imitate types in X by making the announcement: $U^S(Y'(m), t) \leq U^S(Y(M(t)), t)$ for all $t \in \mathcal{T} \setminus X$ and for all $m \in M'(X)$.
C3. Announcement messages are optimal for the types sending them: $U^S(Y'(m), t) \geq U^S(Y'(m'), t)$ for all $t \in X$, for all m in the support of $M'(t)$, and for all $m' \in M'(X)$.

An equilibrium (M, Y) is *neologism-proof* if there does not exist an announcement $\langle M', X\rangle$ in which M' is a constant function and which satisfies conditions C1–C3.[19] An equilibrium (M, Y) satisfies *NITS* if there does not exist an announcement $\langle M', \{0\}\rangle$ satisfying C1. Equivalently, an equilibrium (M, Y) satisfies *NITS* if $U^S(Y(M(0)), 0) \geq U^S(0, 0)$. It is clear that any equilibrium that is strongly announcement-proof is also neologism-proof and satisfies the NITS condition.

All three refinements are quite strict. No equilibrium in the standard C-S model is strongly announcement-proof or even neologism-proof and only the most informative equilibrium satisfies NITS (Chen et al, 2008). The same applies here: while the standard C-S equilibrium outcomes are also equilibrium outcomes in this setting for all $p < 1$, none of them survive strong announcement- or neologism-proofness and only the most informative among them satisfies NITS. However, any FRE satisfies strong announcement-proofness and, thus, neologism-proofness and NITS:[20]

Proposition 5 *Any FRE is strongly announcement-proof.*

While it is clear that a FRE is the best possible outcome for the receiver, in this section I also provide sufficient conditions for a FRE outcome to also be the (ex-ante) best possible equilibrium outcome for the sender among a suitably restricted class of equilibria. It turns out that either a uniform prior over $\mathcal{T}$ with any strictly convex loss function for the receiver, or a quadratic loss function for the receiver and any differentiable full-support prior F are sufficient for the FRE outcome to Pareto dominate all other possible equilibrium

[18] I assume unique best responses to simplify the exposition. The results in this section do not depend on this assumption.

[19] Strictly speaking, this is stronger than the definition as given by Farrell (1993): to conform with it, the inequalities in C1 would have to be strict.

[20] Matthews et al (1991) define two more weaker refinements related to strong announcement-proofness. Naturally, all FRE in this model satisfy both weak announcement-proofness and announcement-proofness.

outcomes in that class. This result is analogous to Crawford and Sobel's (1982) Theorems 3 and 5—the most informative equilibrium in their model is also the one that maximizes both sender's and receiver's expected utility. To eliminate measurability issues, I restrict attention to equilibria whose messaging strategy satisfies the following condition.

Definition 1 A messaging strategy M satisfies *message connectedness* if $M^{-1}(m)$ is a connected set for all $m \in \mathcal{M}$.

In other words, the set $M^{-1}(m)$ for all $m \in \mathcal{M}$ is either empty, a singleton, or an interval. Thus, any message-connected messaging strategy partitions the type space into a collection of connected sets such that all sender types in any such set send the same message. This is sufficient for the existence of a conditional probability distribution over $\mathcal{T}$ regardless of what message-signal combination (m, v) the receiver observes.[21] Let us now turn to the second result of this section.

Proposition 6 *If F is uniform or if the receiver's utility is based on a quadratic loss function (i.e. $\ell^R(x) = x^2$), the FRE outcome is better than any other message-connected equilibrium outcome from the sender's ex-ante point of view.*

4.3 Exogenous Punishment

In this section, I briefly consider what happens if the receiver has access to some exogenous punishment when catching the sender in a lie. This punishment can be either contractual or a proxy for the effect of being caught lying in a repeated-game setting. More precisely, let the receiver choose whether to impose an additional punishment of ε in utility to the sender. Additionally, assume that this punishment is costless to the receiver. More precisely, if caught in a lie after sending message m and penalized, let the utility of a sender of type t be

$$-\ell(t + b - Y(m, -1)) - \varepsilon.$$

Allowing exogenous punishment expands the set of values of p that support a FRE in two ways. While previously no high enough $p < 1$ supported a FRE, with exogenous punishment this result is overturned.

Proposition 7 *In the case of general convex loss utility, precise-message space $\mathcal{M}$, and exogenous punishment ε, there exists a FRE for all $p \geq \frac{\ell(b)}{\ell(b)+\varepsilon}$.*

Additionally, it is not hard to verify that the proofs of Propositions 2 and 3 would remain largely unchanged. The main difference would be that a larger ε enlarges the set of interior values of p, for which there exists a FRE. Two corollaries that follow from this observation combined with Proposition 7 are

[21] Message connectedness is a weaker form of message monotonicity, which is a common assumption in similar models (e.g. see Kartik, 2009 and Chen, 2011).

1) the non-monotonicity in p of the maximal expected equilibrium utility of the receiver is preserved for all small enough ε, while 2) the set of values of p supporting a FRE converges to $(0, 1]$ as $\varepsilon \to \infty$.

4.4 Variable Sender Bias

It is possible to relax the assumption of constant sender bias b. For this section only, assume that the bias parameter varies by type. Namely, let $b(t)$ denote the bias of sender type t so that sender utility is $U^S(y, t) = -\ell(t + b(t) - y)$. I assume that the function $b : \mathcal{T} \to (0, \infty)$ is differentiable and $t + b(t)$ is strictly increasing in t.

It is easy to verify that under these assumption, Proposition 1 and Lemma 1 hold as stated. My main result is also preserved under this more general parameterization of the C-S preferences.

Proposition 8 *Whenever $b(t) \leq 1/4$ for all t, the set of values of p in $(0, 1)$ for which there exists a FRE in the case of general convex loss utility and precise-message space $\mathcal{M}$ includes $p = 1/2$ and is bounded away from 0 and 1.*

Note that, unlike what came above, the condition $b(\cdot) \leq 1/4$ is only sufficient and not necessary for FRE existence. For example, the bias of types close to $t = 1$ can be made arbitrarily large without affecting the existence of a FRE.

5 Discussion of the Assumptions

An ability to identify lying would greatly impact all informal interactions that offer an opportunity to transmit valuable information. Due to the informal nature of such interactions, there often are no formal penalties for lying.[22] Even if formal penalties are allowed, their use could be infeasible if the receiver is facing some institutional or market constraints when it comes to punishing the sender.

I provided a motivating example in the introduction—the interaction between an eyewitness of a crime and the defense's legal team. As another example, we can think of the sender as a job applicant interviewing for a position with a potential employer (the receiver), who has a variety of open positions that differ in their difficulty and contractual terms, such as pay. The state of the world indicates the relevant skills of the job applicant. The potential employer interviews the applicant, who can choose to misrepresent her skill level. With some probability, the employer can determine that the applicant is lying by, say, identifying contradictions between details of the applicant's claimed duties at a previous company and facts known to the employer about that company. He then chooses his action, which represents the position he

[22] These settings stand in contrast to settings such as, for example, income reporting for tax purposes where the detection of fraud carries the threat of a fine or a prison sentence.

offers to the applicant, potentially including the possibility of not making an offer. The employer's preference is that the position offered matches the true skill of the applicant—he wants to appoint low-skilled applicants to low-paid undemanding positions and high-skilled applicants to lucrative positions of responsibility. The job applicant prefers a more demanding (and better paid) position than the one the company would prefer to assign to her.[23]

Another example would be a firm that hires an auditor to evaluate the claims of its employees regarding some unobservable aspect of their work—for example, their productivity, share of effort in a joint project, hours worked while telecommuting, or exact level of spending on a business trip. The firm and its employees are likely to have at least partially aligned objectives but the firm may face constraints of legal (e.g. incomplete labor contracts or inadmissibility of the auditor's findings in court) or labor-market (e.g. inability to replace dismissed workers due to their unique skill sets or current labor-market conditions) character that prevent it from using a contractual punishment whenever it determines an employee is lying.

Another application is in the self-reporting of economic status for the purposes of determining eligibility for social-assistance programs. Martinelli and Parker (2009) study the Mexican program *Oportunidades*, which disbursed cash transfers to families conditional on regular health-clinic visits and on children's consistent school attendance. For eligibility purposes, applicants initially reported their households' possessions and characteristics, which were subsequently verified with a formal household visit. Martinelli and Parker discover that, even though under-reporting of household goods was rampant, lying about one's eligibility never brought about a formal punishment (other than being excluded from the program if ineligible).

As suggested above, a receiver's ability to probabilistically detect lying also seems like a natural way to unite the two extremes of cheap talk and persuasion games. For example, Crawford and Sobel (1982) conclude with a suggestion for extending their seminal results in a similar manner by "allowing S to be uncertain about R's ability to check the accuracy of what he is told." In the rest of this section, I discuss possible interpretations of the main assumptions.

One justification for the assumption of lie detection comes from Dziuda and Salas (2018). They suggest that if the state of the world is relatively complex (e.g. it concerns the details of an event) then the sender's description of that state might contain inconsistent details. The reason for such inconsistencies might be that some people are bad liars, short on deliberation time when inventing their stories, or uninformed about factors that might be used to find inconsistencies in their story. The detection of inconsistencies is crucial in settings such as police interrogation of crime suspects and eyewitnesses, as well as courtroom testimony.

A possible objection to the model presented in this paper regards the inability of the sender to intentionally and with certainty trigger lie detection. Any message that does not match the sender's type is assumed to induce a

[23] This motivating example is borrowed from Farrell and Rabin (1996).

p-$(1-p)$ lottery but it might be in the sender's best interest to, if able, lie *and make sure she is caught lying*. For example, this would be the case if the "punishment" action in response to some message is a sender's type's bliss action. Such an ability would change the results presented in this paper. As noted in the introduction, however, even if her statement contains details that contradict facts known by the sender, she cannot be certain that the receiver has the same knowledge and, thus, an ability to find out the lie. [24]

5.1 Psychological and Experimental Evidence

The assumption that the receiver can detect lying is relatively novel in the economic literature. However, Sobel (2009) suggests that such models can help explain phenomena such as senders' excessive honesty in experimental settings. Frank et al (1993) find that experimental subjects who had a face-to-face interaction before playing a one-shot prisoner's dilemma game were significantly better than chance at predicting correctly the other player's strategy. Brosig (2002) conducts a similar experiment and finds analogous results. Ability to detect lying might explain findings such as the observed excessive (relative to the equilibrium prediction) truth-telling in face-to-face cheap-talk experiments (Holm and Kawagoe, 2010), the excessive cooperation in prisoner's dilemma games with face-to-face communication (e.g. Frohlich and Oppenheimer, 1998), and improvements over the theoretical efficiency upper bounds in bargaining double-auction experiments with face-to-face pre-play communication (Valley et al, 2002; see also Radner and Schotter, 1989) and in market-for-lemons experiments with unstructured face-to-face bargaining (Valley et al, 1998), even relative to other forms of communication.

Sobel (2009) goes on to note that there is "evidence from other disciplines that some agents are unwilling or unable to manipulate information for strategic advantage and that people may be well equipped to detect these manipulations in ways that are not captured in standard models." For the evidence Sobel cites, we can turn to the significant psychological literature suggesting that people are able to read verbal and non-verbal cues during face-to-face communication and use it to detect lying. Those cues stem from liars experiencing emotions such as guilt, fear, or excitement, from the cognitive effort that convincing lying requires, or from the liars' attempts to control their behavior to appear trustworthy. For more on these points, see Vrij (2008)'s excellent and exhaustive survey of the theory and methods of lie detection.

[24] It is also possible to expect that a message designed to be exposed as a lie (an "obvious lie") would manifest to the receiver as distinct from a false message intended to mislead. For example, an obvious lie may contradict commonly known facts, while a deceptive message would include at most subtle inconsistencies. As another example, the receiver might be able to distinguish between genuine nervousness on the part of the sender (as she is lying but does not want to get caught) and fake exaggerated nervousness intended to make the receiver think that she is lying. In such a case, the receiver can respond to obvious lies with the worst possible action ($y = 0$), deterring anyone but the lowest sender's type from deviating to an obvious lie and restoring the FRE-related results presented here.

Ekman (2009) distinguishes between two different ways of detecting lying. They are *leakages*—occurrences where the liar accidentally reveals part or all of the truth that she is trying to conceal (e.g. slips of the tongue, unchecked tirades)—and *deception clues*—occurrences where the liar's behavior suggests that she is lying but does not reveal the truth. Some deception clues noted by Ekman include tone of voice, speech pauses and errors, displaying incongruent emotions, gestures or facial expressions etc. In economic modeling, we can view leakages as corresponding to receivers who have an access to a private (e.g. Olszewski, 2004, Chen, 2009, Lai, 2014) or a public (e.g. Chen, 2012) informative signal, while deception clues correspond to the model presented here.

Vrij (2008, pp. 147-148) considers the evidence from 79 studies of lie detection by lay lie catchers (i.e. not police officers or customs agents, for example), in which they were shown a short video of other experimental participants making a statement. The receivers were not provided with any background information about the statement or the sender. The accuracy was better than chance: on average, the receiver correctly determined whether the sender was lying or telling the truth[25] in 54% of all cases. It is important to note that this accuracy increases when the sender had more to gain by lying, suggesting that the economic relevance of face-to-face interactions increases, rather than diminishes, the importance of lie detection. When concentrating on groups of verbal and non-verbal cues, trained psychology researchers can correctly identify truth-tellers and liars in between 67% and 86% of cases (Vrij, 2008, pp. 66, 108). Belot et al (2012), Belot and van de Ven (2016) and Chen and Houser (2016) have shown that receivers can detect deception in economically relevant situations.[26]

6 Related Literature

My framework unifies the two extremes of the literature on costless strategic communication—cheap talk (Crawford and Sobel, 1982) and persuasion games (Milgrom, 1981, Grossman, 1981). In cheap-talk games the receiver has no way of knowing if the sender is lying, while in persuasion games the receiver can always detect lying and punish it accordingly and, as a result, the sender never lies. Guided by this observation, probabilistic lie detection can be seen as an intermediate case. Indeed, I show that the model is equivalent to a cheap-talk game whenever p is zero and it incorporates a persuasion game with C-S preferences as the other limiting case at $p = 1$, which has a unique

[25] It is easy to check that all the main results of this paper are preserved if, in addition to lie detection, we also endow the receiver with an exogenous ability to stochastically detect truth telling.

[26] Professional poker is a colorful example of the ability of people to detect lies in face-to-face communication. The players go to great lengths to avoid "tells" in their eye movements or facial expressions by cultivating a "poker face" or by relying on hats and shades to hide their faces. Hayano (1980) provides an interesting description of strategic deception in poker.

FRE.[27] Thus, for interior values of p, we can view the game as occupying an intermediate point between the two extreme settings.

The proposed model is closest to Dziuda and Salas (2018), who also assume that the sender can detect lying with some probability.[28] The authors similarly assume precise-language communication augmented with the receiver's ability to detect inconsistent messages from the sender with probability p. The two key classes of differences between the two models are the differences in the modeling assumptions on the preferences and the communication technology, and the different equilibrium-selection strategies. Dziuda and Salas study a setting with non-aligned preferences—the sender's utility is linear and strictly increasing in the receiver's action. They also assume that the sender can, if she so chooses, send a false message that is guaranteed to be detected as a lie. Furthermore, while I focus only on FRE, they study only equilibria that satisfy two conditions. Translating them into the notation used here, the first condition is that any off-equilibrium message-signal combination $(m, 0)$ is "believed:" i.e. $Y(m, 0) = m$. (Note that this is trivially true for the FRE studied above.) The second condition is that all on- and off-equilibrium lies that are detected induce the same action: the receiver does not condition his action on the type that the sender was pretending to be. The two conditions lead Dziuda and Salas to an essentially unique equilibrium which converges to the unique babbling equilibrium of the associated cheap talk game as p converges to zero and to the unique FRE of the associated persuasion game as p converges to one. They find that increasing p always increases information transmission in equilibrium. Furthermore, a FRE in their model exists only for $p = 1$, even if the two equilibrium-refining conditions are not imposed. Dziuda and Salas do not consider vague language.

Another similar model is Holm (2010). He characterizes the equilibria of a simple game with lie detection, two sender types, two messages, and two actions. The preferences of the two players are unaligned—the sender wants to deceive the receiver about the true state of the world, while the receiver wants to learn it. Another notable paper is Lai (2014), who is mainly concerned with modeling a receiver with a private signal but, due to the form that private signal takes, the receiver is sometimes certain that the sender is lying. Finally, Hodler et al (2014) discover a non-monotonicity that has a similar flavor to the one I describe in Section 3: in a communication game, the persuasiveness of a sender is non-monotonic with respect to her lying costs. The two results are logically independent, however.

The literature on games with "partially verifiable" signaling can also be viewed as an intermediate point between the costless and perfectly convincing

[27] Seidmann and Winter (1997) show that a wide class of persuasion games, including those with C-S preferences, have a unique equilibrium that is a FRE. See Mathis (2008) for a further generalization. See also Hagenbach et al (2014).

[28] I commenced work on this idea before becoming aware of their work in progress.

lying of cheap talk[29] and the impossible lying in persuasion games. In models with partial verifiability, a sender can possibly send only a subset of all possible (false) messages. This limited ability to lie has been assumed in the analyses of both pure communication games (e.g. Lipman and Seppi, 1995, Forges and Koessler, 2005) and mechanism design problems (e.g. Green and Laffont, 1986, Celik, 2006, Deneckere and Severinov, 2008, Glazer and Rubinstein, 2012, Mylovanov and Zapechelnyuk, 2016).

6.1 Perturbed Cheap-Talk Games

Another strand of the cheap talk literature that is relevant to the proposed model includes a number of perturbed cheap-talk games (i.e. modifications of the canonical model which include it as a limiting case). They include lying costs (Kartik et al, 2007, Kartik, 2009; see also Banks, 1990, Callander and Wilkie, 2007), non-strategic players (i.e. naive receivers and/or honest senders as in Ottaviani, 2000, Ottaviani and Squintani, 2006, Chen, 2011) and noisy cheap talk (Blume et al, 2007).

Kartik (2009) is closest to the model here. In his model, the sender and receiver have partially aligned preferences as in the C-S model and, additionally, the sender suffers lying costs that are increasing in the distance between her announcement and the true state of the world and are further scaled by an exogenous parameter.[30] Thus, similarly to the model studied here, Kartik's model encompasses both the cheap-talk and persuasion-game settings whenever the scaling parameter is zero and at the limit when tending to infinity, respectively. Kartik's main result is the existence of a partially separating equilibrium: an equilibrium where low types separate but use inflated language (so in equilibrium they all pay some lying costs but the receiver correctly infers their true type), while the high types pool similarly to the equilibria of the canonical model. This equilibrium is (essentially) unique under the monotonic D1 refinement proposed by Bernheim and Severinov (2003). No FRE exists[31] and there is no evidence that the quality of information transmission is non-monotone in the scaling parameter.

Blume et al (2007) build on an idea by Myerson (1991) to show that adding a small probability of communication-jamming noise in a C-S-type game can improve welfare. Similarly, Ivanov (2010) and Ambrus et al (2013) show that biased strategic intermediators can also improve information transmission *if they use mixed strategies*.[32] These results are reminiscent of my findings here

[29] It is unusual to talk of lying in cheap-talk models, since messages acquire their meaning in equilibrium but even if the elements of the message space had exogenous meaning, that meaning would be replaced in equilibrium precisely because lying is easy.

[30] Lying costs are sometimes motivated by the punishment or reputational damage a sender might suffer if discovered to have been untruthful. From this point of view, the model of lie detection I present here can be seen as endogenizing lying costs.

[31] Kartik et al (2007) find that a FRE can exist with lying costs but only with an unbounded state space.

[32] See also Goltsman et al (2009).

since we can view the probabilistic lie detection as an addition of a source of randomness to the standard C-S game, similar to error-inducing noise or a mediator who uses a mixed messaging strategy. The main difference is that this randomness plays no role on the equilibrium path in the receiver-optimal equilibria that I concentrate on.

7 Conclusion

In this paper, I study a cheap-talk game with natural-language communication (both vague and precise) and exogenous stochastic lie detection. The main result is that, if the preferences of the sender and the receiver are sufficiently aligned, a fully revealing equilibrium exists for intermediate (even quite low) probabilities of lie detection, as well as for guaranteed lie detection. With precise language, we observe a non-monotonicity in the maximal ex-ante expected equilibrium utility of the receiver with respect to the probability of lie detection. Surprisingly, allowing for vague messages improves communication outcomes by making full-revelation equilibria supportable for all sufficiently large lie-detection probabilities.

There is a number of potential extensions of the analysis in this paper. A complete characterization of the equilibria of the model might not be tractable due to the severe equilibrium multiplicity inherent in the set-up but an appropriate equilibrium refinement, such as strong announcement-proofness, might restrict the equilibrium set sufficiently to permit such an analysis. An equilibrium analysis of the cases with intermediate ($b \in (1/4, 1)$) and high ($b \geq 1$) degrees of sender bias would also be interesting: in particular, can adding lie detection improve communication outcomes relative to the best equilibria of Crawford and Sobel (1982)?

Another possibly fruitful avenue of further research is extending the analysis to a game where the lie-detection probability is imperfect or is endogenized. For example, if the lie-detection technology permits false positives, no FRE would exist. However, the essence of the main result—the observed non-monotonicity in p—might carry through to that setting. Another example would be a game, in which the receiver makes an (observable or unobserved) investment in a costly auditing technology, after which the two agents play the sender-receiver game studied here.[33]

Acknowledgements I have greatly benefited from comments from the associate editor, the anonymous referees and from David Ahn, Wioletta Dziuda, Haluk Ergin, Nisvan Erkal, Joseph Farrell, Johannes Hörner, Yuichiro Kamada, Maciej Kotowski, Botond Kőszegi, Matthew Leister, Simon Loertscher, Cesar Martinelli, John Morgan, Takeshi Murooka, Omar Nayeem, Santiago Oliveros, In-Uck Park, Matthew Rabin, Roberto Raimondo, Antonio Rosato, Emilia Tjernström, Steven Williams, and participants at the 2016 APET Workshop on Democracy, Public Policy, and Information at Deakin University. Any remaining errors are my own.

[33] This approach is similar to the literature on optimal auditing (e.g. Townsend, 1979, Border and Sobel, 1987, Mookherjee and Png, 1989).

References

Ambrus A, Azevedo E, Kamada Y (2013) Hierarchical cheap talk. Theor Econ 8(1):233–261

Banks J (1990) A model of electoral competition with incomplete information. J Econ Theory 50(2):309–325

Belot M, van de Ven J (2016) How private is private information? The ability to spot deception in an economic game. Exp Econ

Belot M, Bhaskar V, van de Ven J (2012) Can observers predict trustworthiness? Rec Econ Stat 94(1):246–259

Bernheim B, Severinov S (2003) Bequests as signals: An explanation for the equal division puzzle. J Polit Econ 111(4):733–764

Blume A, Board O, Kawamura K (2007) Noisy talk. Theor Econ 2(4):395–440

Border K, Sobel J (1987) Samurai accountant: A theory of auditing and plunder. Rev Econ Stud 54(4):525–540

Brosig J (2002) Identifying cooperative behavior: Some experimental results in a prisoner's dilemma game. J Econ Behav Organ 47(3):275–290

Callander S, Wilkie S (2007) Lies, damned lies, and political campaigns. Games Econ Behav 60(2):262–286

Celik G (2006) Mechanism design with weaker incentive compatibility constraints. Games Econ Behav 56(1):37–44

Chen J, Houser D (2016) Promises and lies: Can observers detect deception in written messages. Exp Econ

Chen Y (2009) Communication with two-sided asymmetric information, mimeo

Chen Y (2011) Perturbed communication games with honest senders and naive receivers. J Econ Theory 146(2):401–424

Chen Y (2012) Value of public information in sender-receiver games. Econ Lett 114(3):343–345

Chen Y, Kartik N, Sobel J (2008) Selecting cheap-talk equilibria. Econometrica 76(1):117–136

Crawford V, Sobel J (1982) Strategic information transmission. Econometrica 50(6):1431–1451

Deneckere R, Severinov S (2008) Mechanism design with partial state verifiability. Games Econ Behav 64(2):487–513

Duffy J, Feltovich N (2006) Words, deeds, and lies: Strategic behaviour in games with multiple signals. Rev Econ Stud 73(3):669–688

Dziuda W, Salas C (2018) Communication with detectable deceit, mimeo

Ekman P (2009) Telling Lies: Clues to Deceit In the Marketplace, Politics, and Marriage. W. W. Norton

Farrell J (1993) Meaning and credibility in cheap-talk games. Games Econ Behav 5(4):514–531

Farrell J, Rabin M (1996) Cheap talk. J Econ Perspect 10(3):103–118

Forges F, Koessler F (2005) Communication equilibria with partially verifiable types. J Math Econ 41(7):793–811

Frank R, Gilovich T, Regan D (1993) The evolution of one-shot cooperation: An experiment. Ethol Sociobiol 14(4):247–256
Frohlich N, Oppenheimer J (1998) Some consequences of e-mail vs. face-to-face communication in experiments. J Econ Behav Organ 35(3):389–403
Glazer J, Rubinstein A (2012) A model of persuasion with boundedly rational agents. J Polit Econ 120(6):1057–1082
Golosov M, Skreta V, Tsyvinski A, Wilson A (2014) Dynamic strategic information transmission. J Econ Theory 151:304–341
Goltsman M, Hörner J, Pavlov G, Squintani F (2009) Mediation, arbitration and negotiation. J Econ Theory 144(4):1397–1420
Green J, Laffont J (1986) Partially verifiable information and mechanism design. Rev Econ Stud 53(3):447–456
Grossman S (1981) The informational role of warranties and private disclosure about product quality. J Law Econ 24(3):461–483
Hagenbach J, Koessler F, Perez-Richet E (2014) Certifiable pre-play communication: Full disclosure. Econometrica 82(3):1093–1131
Hayano D (1980) Communicative competency among poker players. J Commun 30(2):113–120
Hodler R, Loertscher S, Rohner D (2014) Persuasion, binary choice, and the costs of dishonesty. Econ Lett 124(2):195–198
Holm H (2010) Truth and lie detection in bluffing. J Econ Behav Organ 76(2):318–324
Holm H, Kawagoe T (2010) Face-to-face lying—An experimental study in Sweden and Japan. J Econ Psychol 31(3):310–321
Ishida J, Shimizu T (2016) Cheap talk with an informed receiver. Econ Theory Bull 4(1):61–72
Ivanov M (2010) Communication via a strategic mediator. J Econ Theory 145(2):869–884
Kartik N (2009) Strategic communication with lying costs. Rev Econ Stud 76(4):1359–1395
Kartik N, Ottaviani M, Squintani F (2007) Credulity, lies, and costly talk. J Econ Theory 134(1):93–116
Lai E (2014) Expert advice for amateurs. J Econ Behav Organ 103:1–16
Lipman B, Seppi D (1995) Robust inference in communication games with partial provability. J Econ Theory 66(2):370–405
Loftus E, Hoffman H (1989) Misinformation and memory: The creation of new memories. J Exp Psychol Gen 118(1):100–104
Martinelli C, Parker S (2009) Deception and misreporting in a social program. J Eur Econ Assoc 7(4):886–908
Mathis J (2008) Full revelation of information in sender-receiver games of persuasion. J Econ Theory 143(1):571–584
Matthews S, Okuno-Fujiwara M, Postlewaite A (1991) Refining cheap-talk equilibria. J Econ Theory 55(2):247–273
Milgrom P (1981) Good news and bad news: Representation theorems and applications. Bell J Econ 12(2):380–391

Mookherjee D, Png I (1989) Optimal auditing, insurance, and redistribution. Q J Econ 104(2):399–415
Myerson R (1991) Game Theory: Analysis of Conflict. Harvard University Press
Mylovanov T, Zapechelnyuk A (2016) Optimal allocation with ex-post verification and limited punishments, mimeo
Olszewski W (2004) Informal communication. J Econ Theory 117(2):180–200
Ottaviani M (2000) The economics of advice, mimeo
Ottaviani M, Squintani F (2006) Naive audience and communication bias. Int J Game Theory 35(1):129–150
Radner R, Schotter A (1989) The sealed-bid mechanism: An experimental study. J Econ Theory 48(1):179–220
Seidmann D, Winter E (1997) Strategic information transmission with verifiable messages. Econometrica 65(1):163–169
Sobel J (2009) Signaling games. Encyclopedia of Complexity and Systems Science pp 8125–8139
Sobel J (2013) Giving and receiving advice. In: Acemoglu D, Arellano M, Dekel E (eds) Advances in Economics and Econometrics: Tenth World Congress, Cambridge University Press
Townsend R (1979) Optimal contracts and competitive markets with costly state verification. J Econ Theory 21(2):265–93
Valley K, Moag J, Bazerman M (1998) 'A matter of trust': Effects of communication on the efficiency and distribution of outcomes. J Econ Behav Organ 34(2):211–238
Valley K, Thompson L, Gibbons R, Bazerman M (2002) How communication improves efficiency in bargaining games. Games Econ Behav 38(1):127–155
Vrij A (2008) Detecting Lies and Deceit. Wiley
Wang J, Spezio M, Camerer C (2010) Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. Am Econ Rev 100(3):984–1007

A Proofs

Proof (Lemma 2:) Assume there exists a FRE where the receiver's actions off the equilibrium path are given by $Y(\cdot,-1)$. In each of the following cases for the possible values of b, I show that a sender's type has a profitable deviation to lying, which is a contradiction.

First, if $b \geq 1$, if type $t=0$ deviates to the message $m=1$, her expected utility would be

$$\begin{aligned} pU^S(Y(1,-1),0)+(1-p)U^S(1,0) &= -p\ell(b-Y(1,-1))-(1-p)\ell(b-1) \\ &\geq -p\ell(b)-(1-p)\ell(b-1) \\ &> -p\ell(b)-(1-p)\ell(b) \\ &= -\ell(b)=U^S(0,0), \end{aligned}$$

which is the sender's equilibrium utility. Note that the non-strict inequality above follows from the fact that $Y(1,-1)\in[0,1]$.

Second, if $b \in [1/2, 1)$, if type $t = 0$ deviates to the message $m = b$, her expected utility would be

$$pU^S(Y(b,-1),0) + (1-p)U^S(b,0) = -p\ell(b - Y(b,-1)) \geq -p\ell(b) > -\ell(b),$$

where the first inequality follows from the fact that $|b - Y(b,-1)| \leq b$ because $Y(b,-1) \in [0,1]$.

Third, if $b \in (1/4, 1/2)$ and $Y(1/2,-1) \in [0, 2b]$, if type $t = 0$ deviates to the message $m = 1/2$, her expected utility would be

$$\begin{aligned} pU^S(Y(1/2,-1),0) + (1-p)U^S(1/2,0) &= -p\ell(Y(1/2,-1) - b) - (1-p)\ell(1/2 - b) \\ &> -p\ell(b) - (1-p)\ell(b) \\ &= -\ell(b) = U^S(0,0). \end{aligned}$$

Finally, if $b \in (1/4, 1/2)$ and $Y(1/2,-1) \in (2b, 1]$, if type $\hat{t} \equiv Y(1/2,-1) - 2b$ deviates to the message $m = 1/2$, her expected utility would be

$$\begin{aligned} pU^S(Y(1/2,-1),\hat{t}) + (1-p)U^S(1/2,\hat{t}) &= -p\ell(b) - (1-p)\ell(Y(1/2,-1) - b - 1/2) \\ &> -p\ell(b) - (1-p)\ell(b) \\ &= -\ell(b) = U^S(\hat{t},\hat{t}), \end{aligned}$$

where the inequality follows from the fact that $|Y(1/2,-1) - b - 1/2| < 1/4 < b$.

Proof (Proposition 2:) By Lemma 1, in all FRE we have $M(t) = t$ for all $t \in \mathcal{T}$ and $Y(m,0) = m$ for all $m \in \mathcal{M}$. In equilibrium, a sender of type t's utility is $U^S(t,t) = -b^2$. If she deviates to some false report $m \neq t$, it is

$$pU^S(Y(m,-1),t) + (1-p)U^S(m,t) = -p(t + b - Y(m,-1))^2 - (1-p)(t + b - m)^2.$$

Denote this value by $\mathbb{E}\left[U_m^S(t)\right]$. Notice that $\mathbb{E}\left[U_m^S(t)\right]$ is strictly concave in t.

Fix p and a family of off-equilibrium actions $\{Y(m,-1)\}_{m\in\mathcal{M}}$ forming a part of a FRE. By Lemma 2, we have $b \leq 1/4$. Also, we can assume that for each message m there is a unique type $t(m)$ for whom the expected utility of falsely reporting m is larger than the expected utility of any other type when falsely reporting m.[34] In other words

$$\{t(m)\} = \arg\max_{t\in[0,1]} \mathbb{E}\left[U_m^S(t)\right].$$

It is easy to check that $t(m) = \max\{pY(m,-1) + (1-p)m - b; 0\}$. Then, whenever $t(m) = pY(m,-1) + (1-p)m - b \geq 0$, $\mathbb{E}\left[U_m^S(t(m))\right]$ can be computed to satisfy

$$\mathbb{E}\left[U_m^S(t(m))\right] = -p(1-p)(Y(m,-1) - m)^2 \tag{1}$$

whenever $p \in (0,1)$.[35] Notice that in this case, $\mathbb{E}\left[U_m^S(t(m))\right]$ is decreasing in the distance $|Y(m,-1) - m|$. Thus, in order to check that there are no incentives for deviation, it would suffice to check incentive compatibility for the message m for which that distance is the smallest. If $pY(m,-1) + (1-p)m - b < 0$ (implying $t(m) = 0$), we have

$$\mathbb{E}\left[U_m^S(t(m))\right] \leq -p(1-p)(Y(m,-1) - m)^2.$$

[34] It is possible that $t(m) = m$, in which case there is no such unique type. However, due to the continuity of utility in the sender's type, the following analysis, in which I study the incentives of $t(m)$ to deviate to the message m, carries through.

[35] It is worth noting here that (1) does not hold whenever $p = 1$. This is the reason that $p = 1$ does not emerge as a value supporting a FRE from the rest of the arguments comprising the proof of Proposition 2.

It is clear that since $U^S(t,t)$ does not depend on t, to verify that there are no incentives for deviation from truthfulness/full revelation it suffices to check that it is incentive compatible for $t(m)$ to report truthfully rather than deviate to m for each $m \in \mathcal{M} = [0,1]$. In other words, we need to show

$$\sup_m \mathbb{E}\left[U_m^S(t(m))\right] \leq -b^2.$$

Consider the following off-equilibrium actions:

$$Y(m,-1) = \begin{cases} 1 & \text{if } m < 1/2, \\ 0 & \text{if } m \geq 1/2. \end{cases}$$

Clearly, these actions can be rationalized if the sender believes that she is certainly facing type $t = 1$ (or $t = 0$) after observing $(m,-1)$ for $m < 1/2$ (or $m \geq 1/2$).

Note that under these beliefs we have

$$\sup_m \mathbb{E}\left[U_m^S(t(m))\right] \leq \sup_m -p(1-p)(Y(m,-1)-m)^2 = -p(1-p)\frac{1}{4}.$$

Thus, if $-p(1-p)\frac{1}{4} \leq -b^2$, this would guarantee that the proposed off-equilibrium actions can support a FRE. In the case $b \leq 1/4$, the inequality is true if and only if

$$p \in \left[\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}, \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}\right]. \tag{2}$$

Other values of p could support a FRE only if the inequality

$$\sup_m \mathbb{E}\left[U_m^S(t(m))\right] \leq -p(1-p)\frac{1}{4}$$

holds strictly. This is possible only if $\mathbb{E}\left[U_m^S(t(m))\right] \leq -p(1-p)(Y(m,-1)-m)^2$ holds strictly for some open neighborhood of values around $m = 1/2$. In other words, we need $pY(m,-1) + (1-p)(m) - b < 0$ and $t(m) = 0$ for those values of m. Adding the inequality necessary for equilibrium compliance, the following system of inequalities is derived

$$\begin{aligned} pY(m,-1) + (1-p)m - b < 0; \\ \mathbb{E}\left[U_m^S(0)\right] = pU^S(Y(m,-1),0) + (1-p)U^S(m,0) \leq -b^2. \end{aligned} \tag{3}$$

The first inequality implies $Y(m,-1) < b$ because we are in the case $b \leq 1/4$ and the inequalities need to be satisfied for all m in some neighborhood of $1/2$. But the left-hand side of the second inequality is increasing over these values of $Y(m,-1)$ so, to maximize compliance, we may assume $Y(m,-1) = 0$. It is then easy to check that the first inequality is satisfied for $m < \frac{b}{1-p}$ and the second—for $m \geq 2b$. Notice that it is possible for both of these to be satisfied only if $p \geq \frac{1}{2}$.

Since we already know that $p \in \left[\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}, \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}\right]$ can support a FRE, I assume $p > \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}$ for the rest of the proof. This implies $\frac{b}{1-p} > \frac{1}{2}$. To see that, note that since $p > \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}$, it suffices to show that $\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2} \leq 2b$. Re-arranging, we can verify that this is equivalent to $b \leq \frac{1}{4}$. Therefore we have $2b \leq \frac{1}{2} < \frac{b}{1-p}$. So for any message $m \in \left[2b, \frac{b}{1-p}\right)$, the (off-equilibrium) action $Y(m,-1) = 0$ guarantees that no type would want to deviate to that message.

Now consider compliance for messages $m \geq \frac{b}{1-p}$. We want

$$\sup_{m \geq \frac{b}{1-p}} \mathbb{E}\left[U_m^S(t(m))\right] \leq -b^2.$$

It is clear that $t(m) = pY(m,-1) + (1-p)m - b \geq 0$ for all $m \geq \frac{b}{1-p}$ regardless of the value of $Y(m,-1)$. Then, by (1)

$$\sup_{m \geq \frac{b}{1-p}} \mathbb{E}\left[U_m^S(t(m))\right] = \sup_{m \geq \frac{b}{1-p}} -p(1-p)(Y(m,-1)-m)^2$$

for all $m \geq \frac{b}{1-p}$. This can be minimized by, for example, setting $Y(m,-1) = 0$ for all $m \geq \frac{b}{1-p}$ so that

$$\sup_{m \geq \frac{b}{1-p}} \mathbb{E}\left[U_m^S(t(m))\right] = -p(1-p)\left(\frac{b}{1-p}\right)^2.$$

In order to have equilibrium compliance, we must have $-p(1-p)\left(\frac{b}{1-p}\right)^2 \leq -b^2$. It is easy to check that this is satisfied for all $p \in [1/2, 1]$. Thus, in the case we are considering (i.e. $p > \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}$), no sender type would deviate to a message $m \geq \frac{b}{1-p}$ if $Y(m,-1) = 0$ for all such m.

Now consider the compliance for messages $m < 2b$. We want

$$\sup_{m<2b} \mathbb{E}\left[U_m^S(t(m))\right] \leq -b^2.$$

Fix any $m^* \in (b, 2b)$. If $t(m^*) = 0 > pY(m^*,-1) + (1-p)m^* - b$, which is possible only if $Y(m^*,-1) < b$, then

$$\begin{aligned}\mathbb{E}\left[U_{m^*}^S(t(m^*))\right] &= pU^S(Y(m^*,-1),0) + (1-p)U^S(m^*,0)\\ &\geq pU^S(0,0) + (1-p)U^S(m^*,0)\\ &> -b^2,\end{aligned}$$

where the second inequality follows from our analysis of the system (3) and the fact that $m < 2b$. So in order to ensure equilibrium compliance for all messages $m < 2b$, we must have $t(m) = pY(m,-1) + (1-p)m - b$ (at least for all $m \in (b, 2b)$)[36]. We noticed above that in such cases $\mathbb{E}\left[U_m^S(t(m))\right]$ is decreasing in the distance $|Y(m,-1) - m|$ so we can set $Y(m,-1) = 1$ for all $m < 2b$. Then, by (1), we have

$$\sup_{m<2b} \mathbb{E}\left[U_m^S(t(m))\right] = -p(1-p)(1-2b)^2.$$

We must have $-p(1-p)(1-2b)^2 \leq -b^2$. Solving, we get

$$p \in \left[\frac{1}{2} - \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2}; \frac{1}{2} + \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2}\right].$$

Since we are in the case $p > \frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}$, it can be checked that for $b \leq 1/4$ we have

$$\frac{1}{2} + \frac{\sqrt{1-16b^2}}{2} \leq \frac{1}{2} + \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2}.$$

[36] The system of inequalities (3) can be used to find appropriate off-equilibrium actions so that $t(m) = 0$ for some $m < b$. This would not increase the set of p supporting a FRE since for all small $\varepsilon > 0$

$$\sup_{m \in (2b-\varepsilon, 2b)} \mathbb{E}\left[U_m^S(t(m))\right] = -p(1-p)(1-2b)^2 > -p(1-p)(1-b)^2 \geq \sup_{m \in [0,b)} \mathbb{E}\left[U_m^S(t(m))\right].$$

Intuitively, some sender type would want to deviate to a message close to $m = 2b$ before any sender type would want to deviate to any message $m < b$.

So for all $p \in \left[\frac{1}{2} + \frac{\sqrt{1-16b^2}}{2}, \frac{1}{2} + \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2}\right]$, the following off-equilibrium actions guarantee FRE compliance:

$$Y(m,-1) = \begin{cases} 1 & \text{if } m < 2b, \\ 0 & \text{if } m \geq 2b. \end{cases}$$

We conclude that a full-revelation equilibrium can be supported if and only if

$$p \in \left[\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}; \frac{1}{2} + \frac{\sqrt{1-4\frac{b^2}{(1-2b)^2}}}{2}\right].$$

Proof (Proposition 3:) Let p be given. Let $Y(m,0) = m$ for all $m \in \mathcal{M}$ and also

$$Y(m,-1) = \begin{cases} 1 & \text{if } m < 1/2, \\ 0 & \text{if } m \geq 1/2. \end{cases}$$

I will show that given the receiver's actions above, whenever p is in some open neighborhood of $1/2$, it is optimal for the sender to choose $M(t) = t$ for all $t \in \mathcal{T}$. The pair of strategies would form a FRE.

For some message m, consider the maximum possible utility that can be achieved by a sender's type (falsely) sending that message:

$$\max_{t \in \mathbb{R}} \left[-p\ell(t+b-Y(m,-1)) - (1-p)\ell(t+b-m)\right].$$

The value of this optimization program is decreasing in $|m - Y(m,-1)|$. To see that, first note that if ℓ is strictly convex, the objective function is strictly concave so the maximizer t^* is unique and $t^* + b$ is strictly between m and $Y(m,-1)$. So a small decrease in the distance between m and $Y(m,-1)$ clearly increases the value of the objective function evaluated at t^*.

If ℓ is piecewise linear, the optimum is either a corner solution (i.e. $t^* + b$ equals either m or $Y(m,-1)$) for $p \neq 1/2$ or any t^* such that $t^* + b$ is in the closed interval defined by m and $Y(m,-1)$ for $p = 1/2$. Either way, the value of the objective function is $\min\{p, 1-p\}(-\ell(m - Y(m,-1)))$, which is decreasing in $|m - Y(m,-1)|$.

Denote the expected utility of type t sending a false message[37] m by $\mathbb{E}\left[U_m^S(t)\right]$. Note that

$$\sup_{t \in [0,1]} \mathbb{E}\left[U_m^S(t)\right] \leq \max_{t \in \mathbb{R}} \left[-p\ell(t+b-Y(m,-1)) - (1-p)\ell(t+b-m)\right]$$

and therefore

$$\begin{aligned} \sup_{m \in [0,1]} \sup_{t \in [0,1]} \mathbb{E}\left[U_m^S(t)\right] &\leq \sup_{m \in [0,1]} \max_{t \in \mathbb{R}} \left[-p\ell(t+b-Y(m,-1)) - (1-p)\ell(t+b-m)\right] \\ &= \max_{t \in \mathbb{R}} \left[-p\ell(t+b) - (1-p)\ell(t+b-1/2)\right], \end{aligned}$$

where the equality follows from the fact that the program's value is decreasing in $|m - Y(m,-1)|$.

It is optimal for the sender to always report truthfully if

$$\sup_{m \in [0,1]} \sup_{t \in [0,1]} \mathbb{E}\left[U_m^S(t)\right] \leq -\ell(b),$$

[37] In what follows, it would occasionally be useful to estimate $\mathbb{E}\left[U_m^S(t)\right]$ for $m = t$. I interpret $\mathbb{E}\left[U_t^S(t)\right]$ as the expected utility of type t from the same lottery as the one faced by a sender's type falsely reporting to be type t. Note that by continuity, $\mathbb{E}\left[U_t^S(t)\right]$ would be close to $\mathbb{E}\left[U_m^S(t)\right]$ for m close to t.

where $-\ell(b)$ is the sender's utility from reporting truthfully. So, in order to show what we need for $p = 1/2$, it suffices to show the following inequality:

$$\max_{t\in\mathbb{R}} \left\{ -\frac{1}{2}\left[\ell(t+b) + \ell(t+b-1/2)\right] \right\} \leq -\ell(b).$$

Notice that for the type t^* that maximizes the left-hand side's objective function, we again have that $t^* + b$ is between 0 and 1/2. So using Jensen's inequality, the following holds for the left-hand side

$$\frac{1}{2}(-\ell(t^*+b)) + \frac{1}{2}(-\ell(1/2-t^*-b)) \leq -\ell\left(\frac{1}{2}(t^*+b) + \frac{1}{2}(1/2-t^*-b)\right) = -\ell(1/4) \leq -\ell(b),$$

which is what we wanted to show. Additionally, notice that if this inequality is strict (i.e. if $b < 1/4$) by continuity of the left-hand side in p, there would be an open set of values around $p = 1/2$ that also satisfy the inequality, proving the positive measure of the parameters supporting a FRE.

To complete the proof, we need to show that there does not exist a FRE for all $p \in (0, \varepsilon) \cup (1-\varepsilon, 1)$ for some $\varepsilon > 0$. Let $m = b$ and consider the off-equilibrium actions $\{Y(m,-1)\}_{m\in\mathcal{M}}$ for some candidate FRE. If $Y(m,-1) \leq b$, it is easy to see that type $t = 0$ prefers sending the message m over truth-telling regardless of the values of p. If $Y(m,-1) > b$, we have

$$\mathbb{E}\left[U_m^S(Y(m,-1)-b)\right] = -(1-p)\ell(Y(m,-1)-m) \geq -(1-p)\ell(1-b).$$

We have $\lim_{p\to 1}\left[-(1-p)\ell(1-b)\right] = 0$. It is then clear that there exists $\varepsilon > 0$ such that for all $p > 1-\varepsilon$, we have

$$-(1-p)\ell(1-b) > -\ell(b)$$

and so the inequality $\mathbb{E}\left[U_m^S(Y(m,-1)-b)\right] > -\ell(b)$ holds for $p > 1-\varepsilon$. Therefore for all $p > 1-\varepsilon$, there does not exist $Y(m,-1)$ for which all sender types prefer truth-telling over sending the false message $m = b$. Thus, no off-equilibrium can be part of a FRE and there does not exist one.

For small values of p, we start by considering type $t = 0$'s possible deviation to the message $m = b$:

$$\mathbb{E}\left[U_b^S(0)\right] = -p\ell(Y(b,-1)-b).$$

The rest of the proof proceeds analogously.

Proof (Proposition 4:) I first show that if a FRE exists for p with $\mathcal{M}$, then it also exists for the same p with $\mathcal{M}^V$. Let P^* denote the set of values of p, for which there exists a FRE $\{M, Y\}$ in the case without vague messages. We need to show that for all $p \in P^*$ there exists a fully revealing equilibrium for $\mathcal{M}^V$. Fix such a p and consider the messaging profile $M^V(t) = \{t\}$ and the action function

$$\begin{aligned}
Y^V(\{t\}, 0) &= t,\\
Y^V(\{t\}, -1) &= Y(t,-1),\\
Y^V(m, 0) &= 0 \text{ if } m \in \mathcal{M}^V \setminus M(\mathcal{T}),\\
Y^V(m,-1) &= 0 \text{ if } 0 \notin m \in \mathcal{M}^V \setminus M(\mathcal{T}),\\
Y^V(m,-1) &= 1 \text{ if } 0 \in m \in \mathcal{M}^V \setminus M(\mathcal{T}).
\end{aligned}$$

I claim that $\{M^V, Y^V\}$ is a FRE. No sender type wants to deviate to a different equilibrium message since $\{M, Y\}$ is also a FRE. No sender type t strictly prefers the certain action $y = 0$ (induced by deviating to an off-equilibrium truthful message or an off-equilibrium false message $m \not\ni 0$) to the corresponding equilibrium action $y = t$. So it remains to be shown that no sender type t prefers the lottery formed by the actions $y = 0$ with probability $1-p$ and $y = 1$ with probability p (induced by deviating to an off-equilibrium false message

$m \ni 0$) to the equilibrium action $y = t$. Let $t^* \in [0,1]$ be a maximizer of the concave function $-p\ell(1-t-b)-(1-p)\ell(t+b)$, which is the utility that a type t derives from inducing that lottery.

Observe that since $\{M, Y\}$ is a FRE, no type wants to deviate to the message $\{0\}$ and therefore the expected utility of any type t from inducing the lottery over the actions 0 with probability $1-p$ and $Y^V(\{0\},-1) = Y(0,-1)$ with probability p (denoted by $\mathbb{E}[U^S_{\{0\}}(t)]$) is no greater than $-\ell(b)$. So it suffices to show that

$$-p\ell(1-t^*-b)-(1-p)\ell(t^*+b) \leq \mathbb{E}[U^S_{\{0\}}(t)]$$

for some $t \neq 0$. If $Y(0,-1) = 1$, we are done. Assume instead $Y(0,-1) < 1$. Note also that $Y(0,-1) \geq 2b$ because, otherwise, types close to $t = 0$ would have a profitable deviation to $m = \{0\}$.[38]

If $t^* > 0$, consider $t^{**} = \max\{0, t^* - (1 - Y(0,-1))\}$. If $t^{**} > 0$, we have

$$\begin{aligned}\mathbb{E}[U^S_{\{0\}}(t^{**})] &= -p\ell(Y(0,-1)-t^{**}-b)-(1-p)\ell(t^{**}+b)\\ &= -p\ell(1-t^*-b)-(1-p)\ell(t^{**}+b)\\ &> -p\ell(1-t^*-b)-(1-p)\ell(t^*+b),\end{aligned}$$

which is what we wanted to show, and where the inequality follows from the fact that $t^{**} < t^*$.

If $t^{**} = 0$, then $t^{**} \geq t^* - (1 - Y(0,-1))$ and so $Y(0,-1) - t^{**} - b \leq 1 - t^* - b$. Since $Y(0,-1) - t^{**} - b \geq 2b - b = b > 0$ and, by optimality of t^*, $t^* + b \in (0,1)$, we have $\ell(Y(0,-1) - t^{**} - b) \leq \ell(1 - t^* - b)$. Thus, we have

$$\begin{aligned}\mathbb{E}[U^S_{\{0\}}(t^{**})] &= -p\ell(Y(0,-1)-t^{**}-b)-(1-p)\ell(t^{**}+b)\\ &\geq -p\ell(1-t^*-b)-(1-p)\ell(t^*+b).\end{aligned}$$

Finally, assume $t^* = 0$. Since $b \leq 1/4$ and $Y(0,-1) \in [2b, 1)$ we have

$$-p\ell(1-t^*-b)-(1-p)\ell(t^*+b) < \mathbb{E}[U^S_{\{0\}}(0)].$$

Since $\mathbb{E}[U^S_{\{0\}}(0)]$ is continuous in type, there is some type t^{**} close to $t = 0$ which satisfies

$$-p\ell(1-t^*-b)-(1-p)\ell(t^*+b) < \mathbb{E}[U^S_{\{0\}}(t^{**})].$$

Next, I show that a FRE exists for all $p \geq \frac{1}{2}$. Assume $b < 1/4$. Consider the following fully revealing strategy profile (Y, M):

$$\begin{aligned}M(0) &= [0,1],\\ M(t) &= [t, 1-\varepsilon_t] \text{ for all } t \in \mathcal{T} \setminus \{0\},\\ Y(M(t),0) &= t \text{ for all } t \in \mathcal{T},\\ Y(M(t),-1) &= 1-\varepsilon_t \text{ if } t \leq 2b,\\ Y(M(t),-1) &= 0 \text{ if } t > 2b,\\ Y(m,0) &= 0 \text{ if } m \in \mathcal{M} \setminus M(\mathcal{T}),\\ Y(m,-1) &= 0 \text{ if } 0 \notin m \in \mathcal{M} \setminus M(\mathcal{T}),\\ Y(m,-1) &= \sup m \text{ if } 0 \in m \in \mathcal{M} \setminus M(\mathcal{T})\end{aligned}$$

for some $\varepsilon_t \geq 0$ such that $1 - \varepsilon_t \geq \max\{t, 4b\}$ and $\varepsilon_t > 0$ whenever $t \leq 2b$. I will show that this strategy profile is an equilibrium for all $p \geq \frac{1}{2}$.

No type has a profitable deviation to a true or false off-path message $m \notin M(\mathcal{T})$ with $0 \notin m$ since that induces the action 0 for sure. Similarly, no type has a profitable deviation

[38] I am discounting the possibility that $Y(0,-1) = 0$. If that is the case, by continuity we would be able to consider $\mathbb{E}[U^S_{\{\varepsilon\}}(t)]$ instead of $\mathbb{E}[U^S_{\{0\}}(t)]$.

to a true $m \notin M(\mathcal{T})$ with $0 \in m$. A sender type t deviating to a false message $m \notin M(\mathcal{T})$ with $0 \in m$ induces a lottery between actions 0 (with probability $1-p$) and $\sup m$ (with probability p). Note that $t \notin [0, \sup m]$ is possible only if $t > \sup m$. Thus, both 0 and $\sup m$ are worse for t than the action she induces in equilibrium.[39]

No type has a profitable deviation to a true on-path message $m \in M(\mathcal{T})$ since that would induce a lower action than truth-telling induces. What is left to verify is that no type has a profitable deviation to a false on-path message $m \in M(\mathcal{T})$. No type $t^* > 1 - \varepsilon_t$ has a profitable deviation to $M(t)$ since both induced actions are weakly worse than the one t^* induces in equilibrium.

We need to consider possible deviations only from types in $t^* \in [0, t)$. If $t \leq 2b$:

$$\mathbb{E}[U^S_{M(t)}(t^*)] = -(1-p)\ell(t^* + b - t) - p\ell(1 - \varepsilon_t - t^* - b).$$

If ℓ is linear, for all $p \geq \frac{1}{2}$ we have

$$\begin{aligned}\mathbb{E}[U^S_{M(t)}(t^*)] &\leq \mathbb{E}[U^S_{M(t)}(t)] \\ &= -(1-p)b - p(1 - \varepsilon_t - t - b) \\ &\leq -(1-p)b - pb \\ &= -b,\end{aligned}$$

which is the equilibrium payoff and where the second inequality follows from $1 - \varepsilon_t \geq 4b$ and $t \leq 2b$.

If ℓ is strictly convex, then

$$\arg\max_{\hat{t} \in \mathbb{R}} \mathbb{E}[U^S_{M(t)}(\hat{t})] + b$$

is in the interval $(t, 1 - \varepsilon_t)$, it is strictly increasing in p, and it is at the midpoint of the interval $[t, 1 - \varepsilon_t]$ when $p = \frac{1}{2}$. Therefore, since the distance of the interval is at least $2b$, for all $p \geq \frac{1}{2}$, we have

$$\arg\max_{\hat{t} \in \mathbb{R}} \mathbb{E}[U^S_{M(t)}(\hat{t})] + b \geq t + b \Leftrightarrow \arg\max_{\hat{t} \in \mathbb{R}} \mathbb{E}[U^S_{M(t)}(\hat{t})] \geq t.$$

Thus, for all $t^* < t$ we have

$$\begin{aligned}\mathbb{E}[U^S_{M(t)}(t^*)] &< -(1-p)\ell(t + b - t) - p\ell(1 - \varepsilon_t - t - b) \\ &\leq -(1-p)\ell(b) - p\ell(b) \\ &= -\ell(b),\end{aligned}$$

where the second inequality follows from $1 - \varepsilon_t \geq 4b$ and $t \leq 2b$. Thus, for $p \geq \frac{1}{2}$ no type $t^* < t$ has a profitable deviation to the message $M(t)$ whenever $t \leq 2b$.

Now let's consider the case $t > 2b$. As above, for $t^* < t$ we have

$$\mathbb{E}[U^S_{M(t)}(t^*)] = -(1-p)\ell(t^* + b - t) - p\ell(t^* + b).$$

If ℓ is linear, for all $p \geq \frac{1}{2}$ we have

$$\begin{aligned}\mathbb{E}[U^S_{M(t)}(t^*)] &\leq \mathbb{E}[U^S_{M(t)}(0)] \\ &= -(1-p)(t - b) - pb \\ &< -b,\end{aligned}$$

[39] Note that this conclusion wouldn't be changed if we required that $Y(m, -1) > \sup m$ holds in equilibrium whenever $0 \in m$. By continuity, depending on p, we can make $Y(m, -1)$ close enough to $\sup m$ to prevent types right above $\sup m$ from having a profitable deviation.

where the second inequality follows from the fact that $t > 2b$.

If ℓ is strictly convex instead, $\arg\max_{\hat{t}} \mathbb{E}[U^S_{M(t)}(\hat{t})] + b$ is in the set $(0, t)$, it is decreasing in p, and, for $p = \frac{1}{2}$, it is at the midpoint of $[0, t]$. Set $t^{**} = \max\left\{0, \arg\max_{\hat{t}} \mathbb{E}[U^S_{M(t)}(\hat{t})]\right\}$. Thus, for $p \geq \frac{1}{2}$ we have

$$(t - (t^{**} + b)) > b$$

since $t > 2b$. For all $t^* < t$ we have

$$\begin{aligned}\mathbb{E}[U^S_{M(t)}(t^*)] \leq \mathbb{E}[U^S_{M(t)}(t^{**})] &= -(1-p)\ell(t - (t^{**} + b)) - p\ell(t^{**} + b) \\ &< -(1-p)\ell(b) - p\ell(b) \\ &= -\ell(b).\end{aligned}$$

Thus, no type $t^* < t$ has a profitable deviation to the message $M(t)$ whenever $t > 2b$. This completes the first part of the proof.

To establish the statement pertaining to the case of quadratic utility, note that the sufficiency follows from Proposition 2 and the first half of this proof. To demonstrate necessity, it suffices to show that there does not exist a FRE for any $p < \frac{1}{2} - \frac{\sqrt{1-16b^2}}{2}$. Toward contradiction, fix a p satisfying the inequality and let (Y, M) be a FRE for that value of p. I will show that regardless of the values of $M(1/2)$ and $Y(M(1/2), -1)$ in this strategy profile, the sender type t^*, defined by

$$t^* := pY(M(1/2), -1) + (1-p)(1/2) - b,$$

satisfies $t^* \notin M(1/2)$ and would have an incentive to deviate to the message $M(1/2)$.

As we are in a case where $p < 1/2$ and $b \leq 1/4$, the inequality $t^* \geq 0$ holds, regardless of the value of $Y(M(1/2), -1)$. Note also that a sufficient condition for the inequality $t^* < 1/2$ to hold for all $Y(M(1/2), -1) \in [0, 1]$ is $p < 2b$. This is satisfied because $\frac{1}{2} - \frac{\sqrt{1-16b^2}}{2} \leq 2b$ as observed in the proof of Proposition 2. Thus, $t^* \in [0, 1/2)$. By Lemma 3, $\inf M(1/2) = 1/2$ and so $t^* \notin M(1/2)$. Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[U^S_{M(1/2)}(t^*)\right] =& -p(pY(M(1/2), -1) + (1-p)(1/2) - Y(M(1/2), -1))^2 \\ &- (1-p)(pY(M(1/2), -1) + (1-p)(1/2) - 1/2)^2 \\ =& -p(1-p)(Y(M(1/2), -1) - 1/2)^2 \\ \geq& -p(1-p)(1/4) \\ >& -b^2,\end{aligned}$$

where the first inequality follows from the fact that $Y(M(1/2), -1) \in [0, 1]$ and the second follows from how we established (2). Thus, t^* has a profitable deviation to $M(1/2)$, which is the contradiction we need to complete the proof.

Proof (Proposition 5:) Let (M, Y) be a FRE. The equilibrium payoff for all t is $-\ell(b)$. Consider any announcement $\langle M', X\rangle$. I will show that $\langle M', X\rangle$ cannot satisfy condition C1.

If $m \in \mathcal{M}'$ is a message sent as a part of the announcement, let $M'^{-1}(m)$ denote all types in X that can send m. If $M'^{-1}(m)$ is a singleton for all m, then if $M'^{-1}(m) = \{t\}$ it follows that $Y'(m) = t$ and so $U^S(Y'(m), t) = -\ell(b)$ for all m and $t \in X$. Therefore, the inequality in C1 cannot hold strictly.

If $M'^{-1}(m)$ is not a singleton for some m and $Y'(m)$ is the corresponding best-response action by the receiver, it has to be the case that there are types $t_0, t_1 \in M'^{-1}(m)$ such that $t_0 < Y'(m) < t_1$. But then

$$U^S(Y'(m), t_1) < U^S(Y(M(t_1)), t_1) = U^S(t_1, t_1)$$

and so the inequality in C1 does not hold for all $t \in X$.

Proof (Proposition 6:) Consider some message-connected equilibrium with pooling. Namely let all types between $\underline{t}$ and $\bar{t}$ (with $\bar{t} > \underline{t}$) pool on some message m. Notice that since F is atomless, the probability distributions induced by Bayes' rule conditional on $(m, 0)$ and on $(m, -1)$ differ only on a set of Lebesgue measure zero[40] and they, in turn, differ on a zero-measure set from the probability density function over the set of types pooling on m

$$g(t) = \frac{F'(t)}{F(\bar{t}) - F(\underline{t})}.$$

So we can treat $g(\cdot)$ as the pdf of the distributions that are induced by the observation of either $(m, 0)$ or $(m, -1)$. I will show that, under either of the proposition's premises, the receiver's equilibrium action $Y(m, v)$ is just $\mathbb{E}_g[t]$.

First, if the receiver's loss function is quadratic, the receiver chooses y to maximize

$$\begin{aligned}\mathbb{E}_g[-(y-t)^2] &= -\int_{\underline{t}}^{\bar{t}} g(t)(y-t)^2 dt \\ &= -\mathbb{E}_g[t^2] + 2y\mathbb{E}_g[t] - y^2.\end{aligned}$$

This function is strictly concave in y so we can maximize it using the first-order condition. It is $2\mathbb{E}_g[t] = 2y$. Thus, $y^{\max} = \mathbb{E}_g[t]$.

Second, let's consider the case of F being uniform. Note then that, in this case, g is (essentially) uniform. In other words, we can assume $g(t) = 1/(\bar{t} - \underline{t})$. In this case, the receiver chooses y to maximize

$$-\int_{\underline{t}}^{\bar{t}} \ell^R(|y-t|)dt.$$

Since ℓ^R is strictly convex, it is not hard to see that the unique maximizer here is

$$y^{\max} = \mathbb{E}_g[t] = \frac{\underline{t} + \bar{t}}{2}.$$

Now consider the expected value of the expected distance between the sender's bliss action and the equilibrium action conditional on m—i.e. $\mathbb{E}_g[|t + b - \mathbb{E}_g[t]|]$. Letting $t^* := \max\{\underline{t}, \mathbb{E}_g[t] - b\}$, we have

$$\begin{aligned}&\int_{\underline{t}}^{\bar{t}} g(t)tdt = \mathbb{E}_g[t] \\ \Leftrightarrow &b - b + \int_{\underline{t}}^{t^*} g(t)tdt + \int_{t^*}^{\bar{t}} g(t)tdt - \mathbb{E}_g[t] = 0 \\ \Leftrightarrow &\int_{t^*}^{\bar{t}} g(t)(t + b - \mathbb{E}_g[t])dt - b = \int_{\underline{t}}^{t^*} g(t)(\mathbb{E}_g[t] - t - b)dt. \qquad \text{(I)}\end{aligned}$$

Analogously

$$\begin{aligned}\mathbb{E}_g[t] &= \int_{\underline{t}}^{t^*} g(t)tdt + \int_{t^*}^{\bar{t}} g(t)tdt \\ &\leq t^* \int_{\underline{t}}^{t^*} g(t)dt + \int_{t^*}^{\bar{t}} g(t)tdt \\ \Rightarrow \int_{t^*}^{\bar{t}} g(t)tdt &\geq \mathbb{E}_g[t] - t^* \int_{\underline{t}}^{t^*} g(t)dt. \qquad \text{(II)}\end{aligned}$$

[40] In fact, they differ only on a set that is either empty or a singleton, depending on whether m is truthful for some of the types pooling on m.

Thus

$$\begin{aligned}\mathbb{E}_g[|t+b-\mathbb{E}_g[t]|] &= \int_{\underline{t}}^{\overline{t}} g(t)|t+b-\mathbb{E}_g[t]|dt \\ &= \int_{\underline{t}}^{t^*} g(t)(\mathbb{E}_g[t]-t-b)dt + \int_{t^*}^{\overline{t}} g(t)(t+b-\mathbb{E}_g[t])dt \\ &= 2\int_{t^*}^{\overline{t}} g(t)(t+b-\mathbb{E}_g[t])dt - b \\ &= 2\int_{t^*}^{\overline{t}} g(t)tdt + 2(b-\mathbb{E}_g[t])\int_{t^*}^{\overline{t}} g(t)dt - b, \end{aligned} \tag{III}$$

where the third equality follows from (I). If $t^* = \underline{t}$, (III) reduces to

$$\begin{aligned}\mathbb{E}_g[|t+b-\mathbb{E}_g[t]|] &= 2\int_{\underline{t}}^{\overline{t}} g(t)tdt + 2(b-\mathbb{E}_g[t])\int_{\underline{t}}^{\overline{t}} g(t)dt - b \\ &= 2\mathbb{E}_g[t] + 2(b-\mathbb{E}_g[t]) - b \\ &= b.\end{aligned}$$

If $t^* = \mathbb{E}_g[t] - b > \underline{t}$, we must have $\mathbb{E}_g[t] - b > \underline{t} \geq 0$ and hence $\mathbb{E}_g[t] > b$. Then (III) becomes

$$\begin{aligned}\mathbb{E}_g[|t+b-\mathbb{E}_g[t]|] =& 2\int_{t^*}^{\overline{t}} g(t)tdt + 2(b-\mathbb{E}_g[t])\int_{t^*}^{\overline{t}} g(t)dt - b \\ \geq& 2\left(\mathbb{E}_g[t] - t^*\int_{\underline{t}}^{t^*} g(t)dt\right) + 2(b-\mathbb{E}_g[t])\int_{t^*}^{\overline{t}} g(t)dt - b \\ =& 2\mathbb{E}_g[t] - 2(\mathbb{E}_g[t]-b)(1-P) + 2(b-\mathbb{E}_g[t])P - b \\ =& 2\mathbb{E}_g[t] + 2(b-\mathbb{E}_g[t]) - b \\ =& b,\end{aligned}$$

where the inequality follows from (II) and I have denoted $P := \int_{t^*}^{\overline{t}} g(t)dt$. Either way, we have $\mathbb{E}_g[|t+b-\mathbb{E}_g[t]|] \geq b$. In other words, the expected distance between the sender's bliss action and the receiver's equilibrium action conditional on m is no less than b. Thus, the expected utility of the sender conditional on sending the message m in the equilibrium is

$$\mathbb{E}\left[-\ell\left(|t+b-\mathbb{E}_g[t]|\right)|m\right] \leq -\ell\left(\mathbb{E}_g[|t+b-\mathbb{E}_g[t]|]\right) \leq -\ell(b),$$

where I use the fact that ℓ is convex and increasing in its argument's absolute value.

Notice that if a type t separates by being the only to send a certain message m in equilibrium, we must have $Y(m,0) = Y(m,-1) = t$, and hence that type's equilibrium utility is $U^S(y,t) = -\ell(b)$. By the Law of Iterated Expectations

$$\mathbb{E}[U^S(y,t)] = \mathbb{E}\left[\mathbb{E}[U^S|m]\right] \leq -\ell(b),$$

because $\mathbb{E}[U^S|m] \leq -\ell(b)$ for all equilibrium messages m. The sender's utility in all FRE is $-\ell(b)$ and this completes the proof.

Proof (Proposition 7:) Assume $p \geq \frac{\ell(b)}{\ell(b)+\varepsilon}$, and consider the following fully revealing strategy profile:

$$M(t) = t \text{ and } Y(m,v) = \begin{cases} m & \text{if } v = 0, \\ 0 & \text{if } v = -1. \end{cases}$$

It is clear Y is a best response to M. To establish that this is an equilibrium, we only need to verify that the sender does not have profitable deviation. As $Y(m,-1)=0$ regardless of m, a sender of type t maximizes her expected utility from a deviation by sending the message $m=t+b$. The expected utility then is:

$$-p(\ell(t+b)+\varepsilon) \leq -p(\ell(b)+\varepsilon) \leq -\frac{\ell(b)}{\ell(b)+\varepsilon}(\ell(b)+\varepsilon) = -\ell(b),$$

which is the utility from truth-telling.

Finally, note that the receiver need not choose $Y(0,-1)=0$ in FRE: it is easy to show that as long as $\ell(b-Y(0,-1))+\varepsilon > \ell(b)$, no agent would deviate to the message $m=0$.

Proof (Proposition 8:) Abusing notation, extend the domain of the function $b(\cdot)$ by setting $b(t)=b(0)$ for $t<0$ and $b(t)=b(1)$ for $t>1$. With this modification in place, showing that there exists a FRE for values of p around $1/2$ is identical to the first part of the proof of Proposition 3.

Now I show that there does not exist a FRE for all $p \in (0,\varepsilon)\cup(1-\varepsilon,1)$ for some $\varepsilon>0$. Let $m=b(0)$ and consider the off-equilibrium actions $\{Y(m,-1)\}_{m\in\mathcal{M}}$ for some candidate FRE. If $Y(m,-1)\leq b(0)$, it is easy to see that type $t=0$ prefers sending the message m over truth-telling. If $Y(m,-1)>b(0)$, by continuity of $t+b(t)$, which maps onto $[b(0),1+b(1)]$ for $t\in[0,1]$, we can find some t^* such that $t^*+b(t^*)=Y(m,-1)$. Then

$$\mathbb{E}\left[U_m^S(t^*)\right] = -(1-p)\ell(Y(m,-1)-m).$$

Since $Y(m,-1)$ is bounded, we have $\lim_{p\to 1}\left[-(1-p)\ell(Y(m,-1)-m)\right]=0$. Since $b(\cdot)$ is a positive continuous function on the compact domain $\mathcal{T}$, it is bounded away from zero. It is clear then that for all sufficiently high p, the inequality

$$\mathbb{E}\left[U_m^S(t^*)\right] > -\min_t \ell(b(t))$$

holds. Therefore, for sufficiently high p type t^* prefers lying over truth-telling. Either way, the off-equilibrium actions cannot be part of a FRE. Therefore, there does not exist one.

For small p, note that

$$\mathbb{E}\left[U_{b(0)}^S(0)\right] = -p\ell(b(0)-Y(b(0),-1)).$$

The proof then proceeds analogously.