Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 207 (2023) 105572

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Communication with endogenous deception costs [☆]

Ran Eilat [a,*], Zvika Neeman [b]

[a] *Department of Economics, Ben Gurion University of the Negev, Israel*
[b] *School of Economics, Tel Aviv University, Israel*

Received 26 August 2021; final version received 22 September 2022; accepted 30 October 2022
Available online 25 November 2022

Abstract

We study how the suspicion that communicated information might be deceptive affects the nature of what can be communicated in a sender-receiver game. Sender is said to *deceive* Receiver if she sends a message that induces a belief that is different from the belief that should have been induced in the realized state. Deception is costly to Sender and the cost is endogenous: it is increasing in the distance between the induced belief and the belief that should have been induced. A message function that induces Sender to engage in deception is not credible and cannot be part of an equilibrium. We study credible communication with state-dependent and state-independent Sender's preferences. The cost of deception parametrizes the sender's ability to commit to her strategy. Through varying this cost, our model spans the range from cheap talk, or no commitment (Crawford and Sobel, 1982) to full commitment (Kamenica and Gentzkow, 2011).

JEL classification: C72; D82; D83

Keywords: Communication games; Costly deception; Bayesian persuasion; Cheap talk

[☆] We are very grateful to four referees and an Associate Editor for constructive comments and suggestions. We have also benefited from comments by Daniel Bird, Ezra Einy, Kfir Eliaz, Alex Frug, David Levine, Elliot Lipnowski, Benny Moldovanu, Ady Pauzner, Doron Ravid, Joel Sobel, Eilon Solan and participants at numerous seminars. Neeman is grateful for support from The Israel Science Foundation (grant #1465/18) and from the Sapir Center for Economic Development.

[*] Corresponding author.
E-mail addresses: eilatr@bgu.ac.il (R. Eilat), zvika@tauex.tau.ac.il (Z. Neeman).

https://doi.org/10.1016/j.jet.2022.105572
0022-0531/© 2022 Elsevier Inc. All rights reserved.

1. Introduction

Communication is indispensable for almost every economic and social interaction. Oftentimes, an agent with superior information conveys this information strategically to others in order to influence their behavior. If the interests of the parties are not perfectly aligned, then the informed agent may benefit from being dishonest. Yet, research shows that in many cases, in addition to possibly producing material gains, dishonesty is also costly.[1] How does the cost of dishonesty affect the nature of what can be communicated between the parties? The answer to this important question depends, obviously, on the source and form of this cost.

What makes an untruth costly to communicate? If Alice and Bob both stand outdoors at midday, and Alice tells Bob that it is midnight, this is a lie. If, when exiting the dentist's office, Alice announces that she just had the best time of her life, this is also a lie. However, if Bob is a "reasonable person,"[2] these lies do not change his perception of reality. Many would argue that lies that are not believed are costless.

In this paper we take the position that dishonesty is costly only to the extent that it undermines beliefs. Indeed, a common distinction between lying and deception is that a lie is "a statement that the speaker believes is false" whereas deception is a "statement – or action – that induces the audience to have incorrect beliefs" (Sobel, 2020).[3] Accordingly, we assume that deception (rather than mere lying) is costly, and study how the suspicion that communicated information might be deceptive affects the nature of what can be communicated in equilibrium.

We study a standard model of communication between an informed Sender (she) and an uninformed Receiver (he) to which we add a cost of deception. Sender observes a certain variable and sends a message about it to Receiver who, upon receiving the message, updates his belief and takes an action. Receiver's beliefs on the relevant variable depend on the prior distribution, Sender's equilibrium strategy, and the actual message sent. Sender is said to deceive Receiver if she sends a message that is different from what she was supposed to send according to her equilibrium strategy, in a way that distorts Receiver's beliefs relative to his equilibrium expectations. We assume that the cost of deception to Sender is increasing in the "distance" between the belief induced by the message actually sent, and the belief that should have been induced under the message that was supposed to be sent in equilibrium.

The novelty in our approach is two-fold. First, we introduce a (belief-dependent) cost of deception into communication games. This allows us to investigate the effect that costly deception has on the information that is communicated in equilibrium. The form of deception cost employed reflects the emphasis we place on the fact that untruths are costly only to the extent that they affect Receiver's beliefs. Second, because the cost of deception in our model is measured relative to Receiver's equilibrium expectations, it is *endogenous to the model*. This stands in contrast to other papers such as Sobel (2020) and Kartik (2009), in which truthfulness is evaluated relative to an exogenous standard. This distinction is important for two different reasons: (i) conceptually, deception can arguably only be evaluated relative to equilibrium expectations. After all, if Receiver does not anyway believe Sender, then he cannot be deceived. And, (ii) methodologically, as explained below, endogenous deception costs imply that communication

[1] See, e.g., Abeler et al. 2019 for a survey of the extensive experimental literature that documents this cost.

[2] See, e.g., https://en.wikipedia.org/wiki/Reasonable_person.

[3] As emphasized by Sobel (2020), these definitions imply that a lie need not be deceptive, and deception need not involve lying.

need not be monotone, which requires the development of new proof techniques and generates new insights.

Our approach conforms with traditional equilibrium analysis. We say that Sender's strategy is credible if it does not induce Sender to engage in deception. Clearly, a strategy that is not credible cannot be sustained in equilibrium, and we are only interested in what can be communicated in equilibrium. However, although no deception occurs along the path of play, the mere possibility of deception has a large effect on the information that is conveyed by Sender. Indeed, equilibria are oftentimes interpreted as self-enforcing "social norms." If a communication norm is established, it prescribes certain beliefs for the receiver. A deviating Sender understands that her Receiver will update his belief according to the norm, and that she is therefore deceiving him in a particular way. A conscientious Sender would be disturbed by such deception. The cost of deception in our model captures the magnitude of this disturbance.[4]

The ability of Sender to deceive Receiver is closely related to Sender's ability to commit to her message strategy, in the sense of sending the specific message prescribed by the strategy and not a different message. A sufficiently large cost of deception implies "full commitment" of Sender to her message strategy. Such a commitment is obviously very valuable. It is a standard assumption in the literature on Bayesian persuasion (Kamenica and Gentzkow, 2011). By contrast, costless deception implies that Sender cannot commit to follow her message strategy and that, consequently, messages should be interpreted as mere "cheap talk" (Crawford and Sobel, 1982). Both of these extreme cases serve as useful benchmarks for us. They have both been extensively studied in the literature. Through varying the cost of deception in our model, it is possible to span the range from cheap talk, or no commitment, to full commitment.

A motivating example. There are two players: Sender (she) and Receiver (he), and a binary state $\omega \in \{0, 1\}$. The prior belief that $\omega = 1$ is one-third. Receiver has to choose an action $a \in \{0, 1\}$. His payoff is given by $-|a - \omega|$. Hence, Receiver prefers to take action $a = 1$ if and only if his posterior belief that the state is $\omega = 1$ is at least one-half. The payoff for Sender is a. She therefore prefers that Receiver chooses action $a = 1$ regardless of the state.

Suppose that Sender knows the state. If Sender has no qualms about deceiving Receiver, then it is impossible to sustain informative communication in equilibrium. This is because Sender would never send a message that would lead Receiver to choose $a = 0$. Expecting this, Receiver would ignore Sender's message, and choose the action $a = 0$, which is optimal given the prior belief. The payoff to Sender in this case is zero.

Suppose however that it is costly for Sender to deceive Receiver: if Sender deviates from the message that Receiver "anticipates" her to send given the state, then she incurs a cost. Suppose that this cost is equal to the difference between Receiver's posterior belief that the state is $\omega = 1$ that is induced by Sender's actual message, and the belief that should have been induced by Sender's anticipated message.

In this case, it is possible to sustain an equilibrium in which Sender reveals the state. In this equilibrium Sender sends message m_1 in state $\omega = 1$ and message m_0 in state $\omega = 0$. Receiver's posterior belief that the state is $\omega = 1$ following message m_1 (resp. m_0) is 1 (resp. 0), and so he takes action $a = 1$ (resp. $a = 0$). To verify that this is indeed an equilibrium, note that Sender cannot benefit from deceiving Receiver and sending message m_1 in state $\omega = 0$. Sender's material gain from this deviation is one (because Receiver would take action $a = 1$ instead of $a = 0$).

[4] We are grateful to an anonymous referee for this interpretation.

However, this gain is exactly offset by Sender's cost of deception (because Receiver's posterior belief that the state is $\omega = 1$ following this deception is one instead of zero). The (expected) payoff to Sender in this equilibrium is one-third.

At the extreme, if deception is infinitely costly for Sender, then she can obtain an even higher payoff. In this case, Sender has full commitment power and so, as shown by Kamenica and Gentzkow (2011), she can obtain an expected payoff of two-thirds by employing the following strategy: when the state is $\omega = 1$, send message m_1; and when the state is $\omega = 0$, send messages m_1 and m_0 with equal probabilities. Given this strategy, Receiver takes action $a = 1$ following message m_1 and action $a = 0$ following message m_0. Because deception is infinitely costly, Sender cannot benefit from deviating and sending message m_1 when she should send message m_0.

In Section 2 we present the model and formally define the notion of deception costs. In Section 3 we study the implications of the possibility of costly deception in an environment in which Sender's payoff depends on the state of the world. To that end, we introduce the possibility of costly deception into the uniform-quadratic specification of Crawford and Sobel's (1982) model of strategic communication. The key difference between the equilibria that emerge in Crawford and Sobel's (1982) model and in ours is that in Crawford and Sobel (1982) *every* equilibrium is a partition equilibrium in which each element of the partition is an interval of Sender's types.[5] By contrast, in our model Sender may induce a credible partition of the state space with *non-convex* elements. However, we show that the *optimal* partition for Sender consists only of intervals. Our proof technique, which we explain in detail in Section 3, is substantially different from that of Crawford and Sobel (1982). This result allows us to explicitly solve for the optimal partition for Sender and describe how it relates to the optimal partition in Crawford and Sobel (1982). The fact that deception is costly facilitates more informative communication between Sender and Receiver compared to the most informative equilibrium in Crawford and Sobel (1982). In fact, we show that increasing the cost of deception is akin to decreasing the value of the parameter that measures the sender's bias in Crawford and Sobel's uniform-quadratic model.

In Section 4 we apply our definition of costly deception to a model in which Sender's payoff is independent of the state of the world. For this purpose, we introduce the possibility of costly deception into the payoff environment of Kamenica and Gentzkow's (2011) leading example. We provide a geometric characterization of Sender's highest equilibrium payoff in this model. We show that this highest payoff is obtained on a partial concavification of Sender's indirect payoff function with a bounded slope. We show that Sender's value is continuous in the cost of deception parameter but may be discontinuous in prior beliefs. We describe environments where communication involves either more or less garbling compared to the case of full commitment. Finally, we show that a lower cost of deception always hurts Sender and discuss the circumstances under which it either benefits or hurts Receiver.

In Section 5 we discuss two additional issues: an alternative tie-breaking rule for Receiver and the number of messages employed by Sender in an optimal credible message function. Section 6 concludes.

[5] Specifically, every equilibrium in Crawford and Sobel (1982) induces a partition of the sender's type space into intervals such that all the sender's types that belong to the same interval send the same message.

1.1. Related literature

Sobel (2020) introduces game-theoretic definitions of lying and deception.[6] Our definition of deception is consistent with his in that, in our model too, deception involves inducing "incorrect beliefs." However, we also add a cost of deception that is not explicitly incorporated into Sobel's model. Another key distinction is that, according to our definition, deception is measured with respect to equilibrium beliefs and is therefore endogenous, whereas in Sobel's model, deception is measured with respect to the true state, and so is determined by an exogenous standard.[7]

Kartik (2009) is perhaps closest in spirit to the ideas introduced in the part of our analysis in which Sender's preferences do depend on the state of the world. Kartik (2009) extends the analysis of Crawford and Sobel (1982) by incorporating the possibility of costly lying into the communication game. The cost of lying in Kartik's (2009) model depends only on the state and the literal message used by the sender, which may be interpreted as an announcement about the state. Equilibria in Kartik's model involve lying, but no deception. The key distinction between Kartik's (2009) model and ours is that we measure the cost of deception in terms of the differences in the receiver's induced beliefs. Another distinction is that, unlike in this paper, Kartik (2009) restricts attention to monotone equilibria. Consideration of non-monotone equilibria is important because it is not a-priori known whether the Sender- (or Receiver-) optimal equilibrium is indeed monotone. In Section 3 we devote a large part of the analysis to showing that the Sender-optimal equilibrium is indeed monotone also when deception is costly.[8]

A key difference between the equilibrium predictions of Kartik's (2009) model (and other related models, e.g., Ottaviani and Squintani, 2006) and the model presented in our Section 3 is that while equilibria in Kartik's (2009) model have a hybrid form, such equilibria do not exist in our Section 3 model. Specifically, in Kartik's (2009) model, the sender may pool her type on an interval of low states, and fully separate elsewhere. By contrast, in our model, either there is full separation, or pooling within intervals. This difference is due to the fact that in Kartik's (2009) model, mimicking a higher type is associated with a cost that depends both on the message sent and on the true state, and therefore may be different for different types, whereas in our model deception cost depends only on differences in equilibrium beliefs.[9]

The paper that is closest to the analysis in Section 4 is Guo and Shmaya (2021).[10] They consider a setting in which a sender provides probabilistic forecasts to a receiver through messages that have literal meaning in the form of "asserted distributions over states." The sender

[6] Sobel employs Austin's (1975) classification of speech acts: locution is what the speaker says; illocution refers to the interpretation of what she says; and perlocution is the consequences of the statement. A lie is a statement that the speaker believes is false. It is therefore defined purely in terms of locution. Deception is a statement or action that induces the audience to have incorrect beliefs. It is therefore an illocutionary act.

[7] Sobel defines a message to be deceptive if it induces beliefs that are a mixture between the beliefs that are induced by some other message and some completely inaccurate beliefs that shift Receiver's beliefs "further from the truth."

[8] Other models of lying consider perturbed versions of games in which, with positive probability, the sender is a behavioral type who always reports honestly; or the receiver is a behavioral type who interprets messages literally, believing that the state is m after receiving the message m (Chen, 2011); or that players incur lying costs when they break their promises to each other (Heller and Sturrock, 2020).

[9] Specifically, if types ω and $\omega + \Delta$ both reveal themselves in equilibrium, then the cost of deception that type ω incurs when mimicking type $\omega + \Delta$ depends on Δ but not on ω. It follows that if separation is possible in some interval of the type space, then it is possible everywhere.

[10] However, while in Section 4 we consider the case in which Sender's preferences are independent of the state of the world, Guo and Shmaya (2021) consider general sender-receiver games in parts of their analysis.

in their model bears a miscalibration cost that depends on the discrepancy between the forecast and the truth. By contrast, in our model the meaning of the sender's messages is determined endogenously in equilibrium (i.e., messages have no literal meaning in our model) and the cost of deception depends on the distance between the beliefs that the sender actually induced and the beliefs that the sender should have induced in equilibrium. Guo and Shmaya's notion of a calibrated equilibrium is credible according to our definition, but not vice-versa.[11] Their main focus is on promotion games in which the receiver has two actions and the sender's preferences are independent of the state, and on the case in which the cost intensity parameter is large. In this latter case, they show that the sender attains her full-commitment payoff under any extensive-form rationalizable play. In our model, the sender also attains her full-commitment payoff when deception is sufficiently costly.[12]

Another sense in which the model of Guo and Shmaya (2021) is similar to ours is that by varying the cost intensity parameter, it bridges the gap between cheap-talk models (such as Crawford and Sobel, 1982, and Lipnowski and Ravid, 2020) and models in which the sender has full-commitment power (Kamenica and Gentzkow, 2011). Another paper that bridges this gap is Lipnowski et al. (2022). In their model a sender commissions a study to persuade a receiver, but may influence the report with some state-dependent probability. They show that increasing this probability can benefit the receiver and can lead to a discontinuous drop in the sender's payoffs.

When the sender's preferences are independent of the state, the solution to the sender's problem admits an elegant geometric characterization. Kamenica and Gentzkow (2011) famously characterize the sender's value in terms of the *concave* envelope of her indirect payoff function. Lipnowski and Ravid (2020) characterize it in terms of the quasi-concave envelope of the sender's indirect payoff function, and Lipnowski et al. (2022) characterize it in terms of a mixture of the concave and quasi-concave envelopes of the sender's indirect payoff function. By contrast, in our framework, it is possible to characterize the sender's value in terms of the concave envelope of her indirect payoff function, with a bounded slope.

Other papers that study communication with partial commitment include Perez-Richet and Skreta (2022) and Nguyen and Tan (2019). Perez-Richet and Skreta (2022) consider a model in which an agent can manipulate a Blackwell experiment's input at a cost. They characterize receiver-optimal tests under different constraints in this setting. In Nguyen and Tan (2019), a sender has the opportunity to privately change the publicly observed outcome of a previously announced experiment, at a cost that depends on the outcome. They describe conditions under which the sender does not alter the experiment's outcome in the sender-optimal equilibrium. In their model, the sender benefits from assigning her preferred beliefs to messages that are harder to mimic.

Possibly misleading communication has also been studied experimentally. Recently, Fréchette et al. (2022) studied the role of commitment in communication experimentally using a framework that nests models of cheap talk, information disclosure, and Bayesian persuasion. Gneezy (2005) and Fischbacher and Föllmi-Heusi (2008) are examples of experimental papers on com-

[11] Guo and Shmaya (2021) study also equilibria in which costly miscalibration occurs. In contrast, in our model deception does not occur in equilibrium.

[12] The insight that changes in beliefs constrain the principal appears also in the context of the economics of privacy. For example, in Eilat et al. (2021) a mechanism designer is constrained by how much she is allowed to learn about players where learning is measured in terms of the difference in the designer's beliefs. See also Krähmer and Strausz (forthcoming) and Gradwohl (2018). In contrast, in our paper the cost due to the change in beliefs enters Sender's payoff function rather than the constraints she faces.

munication that associate the message to the state and treat messages as lies if they are not equal to the state.

Finally, the fact that the sender's payoff depends directly on the receiver's endogenous beliefs implies that the game we consider is a psychological game (Geanakoplos et al., 1989, Battigalli and Dufwenberg, 2009).[13] This literature explains the distaste for lying as an aversion to guilt (Battigalli and Dufwenberg, 2007).[14] Other papers consider communication between an informed sender and an uninformed receiver within the framework of psychological games, as we do, but with a very different focus from ours. See for example Caplin and Leahy (2004), Ottaviani and Sørensen (2006), Ely et al. (2015), and more recently Hagenbach and Koessler (2022).

2. Model

Consider a communication game with two players: Sender (S, she) and Receiver (R, he). Players' material payoffs depend on a state of the world and on Receiver's action. The state of the world is drawn from a set $\Omega \times \Theta$. The set Ω represents the "payoff relevant" part of the state.[15] The set $\Theta = [0, 1]$ is used to incorporate lotteries into Sender's choice of messages, as described below. The prior probability of the payoff relevant part of the state is denoted by $\pi \in \Delta(\Omega)$. Without loss of generality, we assume that the prior distribution over Θ is uniform. These two prior distributions are stochastically independent.

Sender observes the state of the world and sends a message to Receiver. We denote the set of possible messages by M, which for simplicity we assume to be an interval of real numbers.[16] Upon observing the message sent by Sender, Receiver chooses an action from a compact set $A \subset \mathbb{R}$.

Both Receiver and Sender are expected utility maximizers. The payoff for Receiver is given by $u_R(a, \omega)$, where $a \in A$ is Receiver's action and $\omega \in \Omega$ is the payoff relevant part of the state of the world. For any belief about the payoff relevant state $p \in \Delta(\Omega)$, denote the set of optimal actions for Receiver by

$$A^*(p) = \arg\max_{a \in A} \quad \mathbb{E}_\omega \left[u_R(a, \omega)\right], \tag{1}$$

where the expectation is computed according to the belief p. Sender's material payoff is denoted $u_S(a, \omega)$. If Sender deceives Receiver then she also incurs a cost of deception as described below. To simplify notation, if Receiver randomizes over actions using a distribution $\hat{a} \in \Delta(A)$, then we interpret $u_S(\hat{a}, \omega)$ as the expected value $\mathbb{E}_a\left[u_S(a, \omega)\right]$ where the expectation is computed according to the distribution $\hat{a}$. The functions $u_R(a, \omega)$ and $u_S(a, \omega)$ are assumed to be continuous.

Receiver's strategy is a measurable function $\tilde{a} : M \to \Delta(A)$. Sender's strategy is a measurable function $\sigma : \Omega \times \Theta \to M$.[17] We refer to Sender's strategy as her *message function*. Each message

[13] For a recent survey of the literature on psychological games see Battigalli and Dufwenberg (2022).

[14] Loginova (2012) studies guilt aversion in the setup of Crawford and Sobel (1982) and characterizes the partitional equilibria in the uniform-quadratic version of this model.

[15] We assume that Ω is a compact metrizable space. We let $\Delta(\Omega)$ denote the set of all Borel probability measures over Ω, endowed with the weak* topology.

[16] More generally, the set of messages can be given by any sufficiently rich measurable standard space, where for richness it suffices to assume that the cardinality of M is at least the cardinality of $\Omega \times \Theta$.

[17] The incorporation of the set Θ into the state space ensures that the state space is sufficiently rich so that the assumption that Sender employs a pure strategy involves no loss of generality.

that is sent under σ induces a posterior distribution over the payoff relevant states $\omega \in \Omega$, which we denote by $p_m^\sigma \in \Delta(\Omega)$.[18]

Fix a message function σ for Sender. According to the message function σ, Sender is supposed to send message $m = \sigma(\omega, \theta)$ in state (ω, θ). If, instead, Sender sends message $m' \neq \sigma(\omega, \theta)$, then she is said to *deceive* Receiver because message m' induces the "wrong" posterior belief $p_{m'}^\sigma$ instead of p_m^σ. We assume that the cost to Sender from sending message m' instead of message $m = \sigma(\omega, \theta)$ is

$$c\left(m' \,|m, \sigma\right) = \alpha \cdot d\left(p_{m'}^\sigma, p_m^\sigma\right),$$

where $d : \Delta(\Omega) \times \Delta(\Omega) \to \mathbb{R}_+$ is a distance function between beliefs over Ω, and $\alpha \geq 0$ is a parameter that scales the cost of deception.[19] That is, the cost of sending a message m' when the state of the world is (ω, θ) is proportional to the distance between the posterior belief p_m^σ that should have been induced by the message $m = \sigma(\omega, \theta)$ and the posterior belief that is actually induced by the message m', which is $p_{m'}^\sigma$.

Formally, an equilibrium of the communication game is a message function σ for Sender, a response strategy $\tilde{a}$ for Receiver, and Receiver's beliefs $\{p_m^\sigma\}_{m \in M}$ such that:

1. For every message m that is sent by σ, Receiver's posterior belief about the state p_m^σ is consistent with Sender's message function.[20]
2. Receiver's response to any message m is optimal given his belief.[21]
3. Sender's message function is optimal given Receiver's response strategy, the state, and the cost of deception. That is, for any state $(\omega, \theta) \in \Omega \times \Theta$ and message $m = \sigma(\omega, \theta)$,

$$u_S(\tilde{a}(m), \omega) \geq u_S(\tilde{a}(m'), \omega) - c\left(m' \,|m, \sigma\right), \tag{2}$$

for every message $m' \in M$.

A message function σ is said to be *credible* if it is part of an equilibrium of the communication game. We refer to condition (2) as Sender's *credibility constraint*.

Like other communication games, the communication game with costly deception studied here may also have multiple equilibria. One equilibrium that always exists is an "uninformative equilibrium," in which Sender employs a trivial message function that sends the same message in all states, Receiver's belief coincides with the prior, and Receiver best responds to this belief.

We focus on Sender-optimal equilibria of the communication game. A Sender-optimal equilibrium is a solution to the following problem:

[18] More rigorously, since M is a standard space, a regular conditional probability distribution exists (Shiryaev (1996), Theorem 4, p. 227). And, any two such distributions yield the same beliefs for all but a measure zero set of messages. The posterior distribution $p_{m \in M}^\sigma$ mentioned above is one version of such a distribution.

[19] Formally, we assume that $d(x, y)$ is a pseudometric. That is, it satisfies the following four properties: it is nonnegative, symmetric, $d(x, x) = 0$ for every x (but, possibly, $d(x, y) = 0$ for some $x \neq y$), and it satisfies the triangle inequality. Note that the triangle inequality is not satisfied by the quadratic lying cost in Kartik (2009) or the Kullback-Leibler distance in Guo and Shmaya (2021).

[20] If Receiver observes a message that was not supposed to be sent by σ, then we assume that Receiver's belief coincides with his belief after some arbitrary message that is sent by Sender in equilibrium. The exact identity of the on-path message in this case is not important. This assumption ensures that off-path messages never serve as a tempting deviation for Sender.

[21] If Receiver has a unique best response, then he chooses it; otherwise, Receiver can mix among his best responses. Namely, the support of the distribution $\tilde{a}(m)$ is a subset of $A^*(p_m^\sigma)$.

$$\max_{\langle \sigma, \tilde{a}, \{p_m^\sigma\} \rangle} \quad \mathbb{E}_\omega \left[\mathbb{E}_m \left[u_S \left(\tilde{a}(m), \omega \right) \mid \omega \right] \right] \tag{SP}$$

where $\langle \sigma, \tilde{a}, \{p_m^\sigma\} \rangle$ is an equilibrium of the communication game. In problem (SP), the inner expectation $\mathbb{E}_m$ is evaluated according to the probability distribution over messages induced by σ and $\theta \in \Theta$, conditional on the state $\omega \in \Omega$. The outer expectation $\mathbb{E}_\omega$ is evaluated according to the prior probability distribution over states π.

On the equilibrium path, Sender's message function σ determines Receiver's beliefs following every message that is sent by Sender. Receiver's beliefs determine Receiver's response. Notice that if for any belief Receiver has a unique best response, then the equilibrium is completely pinned down by Sender's message function σ (provided that it satisfies the credibility constraint, (2)). We analyze such an environment in Section 3 below.

Increasing the value of the deception cost parameter α relaxes Sender's credibility constraint. This expands the set of possible equilibria. It immediately follows that:

Observation 1. Sender's expected payoff in the solution to Problem (SP) is weakly increasing in the cost parameter α.

It is noteworthy that Sender's credibility is closely related to Sender's ability to commit to her message function. Specifically, credibility imposes an incentive compatibility constraint on Sender. If the cost of deception $c(\cdot|\cdot, \cdot)$ is infinite, then Sender has full-commitment power. That is, she would never deviate from any message function she employs, and so this incentive compatibility constraint is never binding. Otherwise, Sender may benefit from deviating from certain messages. In this case, Sender has only partial-commitment power because she can commit only to those message functions from which she would not want to deviate. Partial commitment limits Sender's ability to communicate. However, as famously shown by Crawford and Sobel (1982), nontrivial communication is possible even when the cost of deception is zero and Sender has no commitment power whatsoever.

Finally, as mentioned above, what distinguishes our approach is that in our model the cost of deception is endogenous: it depends on the distance between beliefs induced by Sender's message function σ in equilibrium.

3. Credibility with state-dependent utilities

We now turn to apply our notion of credibility to the payoff specification of Crawford and Sobel (1982). We focus our attention on the classic uniform-quadratic version of the model.

Suppose that $\Omega = [0, 1]$, with a uniform prior probability distribution. The set of Receiver's actions is given by $A = [0, 1]$. Sender's and Receiver's payoff functions are given by $u_S(a, \omega) = -(a - (\omega + b))^2$ and $u_R(a, \omega) = -(a - \omega)^2$, respectively, for some $b \geq 0$. It is easy to verify that for any belief p Receiver's optimal action is unique. As explained above, in this case Sender's message function pins down the equilibrium.

For simplicity, we restrict attention to the case where Sender chooses a measurable message function $\sigma : \Omega \to M$.[22] This allows us to identify a message function with the partition it induces over Ω. We also identify each message m with the set of states where message m is sent, $m \equiv \{\omega : \sigma(\omega) = m\}$.

[22] The restriction to $\sigma : \Omega \to M$ instead of $\sigma : \Omega \times [0, 1] \to M$ implies that we assume that Sender employs a pure strategy in each state $\omega \in \Omega$.

Denote the Lebesgue measure of message m by $|m|$. This is also the probability that message m is sent in equilibrium. We denote the expected state conditional on message m (mean of m) by μ_m and denote the infimum and supremum of m by $\underline{m}$ and $\overline{m}$, respectively.[23]

We assume that the distance between any two beliefs is given by the difference between their means. That is, the cost of sending message m', in a state that belongs to message m, is given by

$$c(m'|m,\sigma) = \alpha \cdot |\mu_m - \mu_{m'}|.$$

Notice that when $\alpha = 0$, our model coincides with that of Crawford and Sobel (1982).

Receiver observes the message sent by Sender and chooses an action $a \in A$. A simple calculation shows that the action that maximizes Receiver's payoff, following message m, is given by the mean of m. I.e., (with a slight abuse of notation),

$$\tilde{a}(m) = \mu_m.$$

The expected payoff to Sender from a given partition of the state space Ω into messages is given by the expected variance of the messages in this partition:

$$-\mathbb{E}_m \left[\mathrm{Var}\left[\omega | \omega \in m\right]\right] \tag{3}$$

up to a constant, where the expectation is taken according to the distribution over messages induced by the partition.[24]

The credibility condition (2) implies that type $\omega \in m$ of Sender prefers sending message m to sending any other message m':

$$-(\omega - \mu_m + b)^2 \geq -(\omega - \mu_{m'} + b)^2 - \alpha \left|\mu_{m'} - \mu_m\right|. \tag{4}$$

The left-hand side of (4) is type ω's payoff from sending the message m, after which Receiver takes the action μ_m. The right-hand side is type ω's payoff from sending the message m', inducing Receiver's action $\mu_{m'}$ but incurring a deception cost of $\alpha \left|\mu_{m'} - \mu_m\right|$. A partition that satisfies condition (4) is said to be a *credible partition*.

The objective of Sender is to find a credible partition that maximizes the objective function (3). Since any two messages that induce an identical expectation can be merged into one message without affecting credibility or the objective function's value, there is no loss of generality in assuming that each message in the partition chosen by Sender has a different mean.

The next lemma establishes a connection between the bias parameter b, the deception cost parameter α, and the number of messages included in a credible partition:

Lemma 1. *If $\alpha \geq 2b$ then revelation of the state (the "truthful partition") is credible. If $\alpha < 2b$ then the number of messages in any credible partition is bounded from above by $1/(2b - \alpha)$.*

Thus, when $\alpha \geq 2b$ the solution to Sender's problem is immediate. In this case the truthful partition in which Sender reveals the state is credible, and the value of Sender's objective function (3) is zero (and therefore Sender's payoff is $-b^2$). This implies that the truthful partition is also optimal.

[23] The mean μ_m is computed with respect to the version of the posterior distribution p_m^σ that we fixed in footnote 18.

[24] Given a message function σ, Sender's payoff is given by $-\mathbb{E}_m \mathbb{E}_{\omega \in m}\left[(\tilde{a}(m) - \omega - b)^2\right]$, which is equal to $-\mathbb{E}_m \mathbb{E}_{\omega \in m}\left[(\tilde{a}(m) - \omega)^2\right]$ up to a constant that is independent of σ. This last expression is equal to minus the expected induced variance (3) and also (by definition) to Receiver's expected payoff.

We henceforth restrict our attention to the case where $\alpha < 2b$. Since, by Lemma 1, we may assume that Sender employs a finite number of messages, no loss of optimality is implied by assuming that all messages included in a credible partition have a positive measure. Given a credible partition, we order the messages according to their conditional means, and denote the k^{th} message by m_k and its mean by μ_k.

The credibility constraint (4) is equivalent to the following two constraints holding for every $k < J$, where J denotes the number of messages in the partition:

$$\frac{\mu_k + \mu_{k+1}}{2} - \overline{m}_k \geq b - \frac{\alpha}{2}. \tag{ICup(k)}$$

$$\frac{\mu_{k-1} + \mu_k}{2} - \underline{m}_k \leq b + \frac{\alpha}{2}. \tag{ICdown(k)}$$

Constraints ICup(k) and ICdown(k) ensure that when the state ω belongs to message m_k Sender prefers to send m_k over sending messages with higher and lower means, respectively.[25]

We proceed with the following definition:

Definition 1. A partition of Ω into convex messages (intervals) is said to be a convex partition.

Crawford and Sobel (1982) famously showed that *any* equilibrium of the cheap-talk model induces a convex partition (in which the first element determines the structure of the entire partition). When deception is costly this result no longer holds. Specifically,

(1) In Crawford and Sobel (1982) equilibrium partitions are monotone: if type ω is indifferent between two messages m, m' with $\mu_m < \mu_{m'}$, then every type $\omega' > \omega$ strictly prefers m' to m and every type $\omega'' < \omega$ strictly prefers m to m' (this is a consequence of the assumption that Sender's preferences satisfy the single-crossing property). By contrast, in our setting, because the cost of switching to a different message is endogenous and depends on the type's equilibrium message, it is possible to have two types $\omega < \omega'$ such that ω' prefers m to m' but ω prefers m' to m. Moreover, *any* partition is credible for values of α that are sufficiently high.

(2) In Crawford and Sobel (1982) the first element of the partition determines the entire partition structure. This is because the structure of the partition is determined by a set of types who are indifferent between pairs of contiguous elements in the partition. By contrast, in our case, fixing the first element of a partition (even a convex partition) does not pin down the next elements of the partition. It is noteworthy that indifference conditions are not a necessary feature of the partition. Namely, it is possible to have convex partitions in which no type is indifferent between any pair of messages.

We proceed by showing that, although incentive compatibility does not imply convexity of the induced partition, the *optimal* partition (i.e., the one that maximizes the objective function (3)) is in fact convex. The main challenge is that, given a credible partition, it is difficult to find a credible "global" improvement for it. And, "local" improvements may violate credibility. Our approach is to perform a sequence of local improvements that converge to a convex partition while correcting for violations of credibility along the way.

[25] To see this, note that ICup(k) is tightest for messages with "adjacent means." And if the constraint is satisfied for $\omega = \overline{m}_k$ then it is satisfied for all $\omega \in m$. On the other hand, if the constraint is not satisfied for $\omega = \overline{m}_k$, then there exists a state $\omega \in m$ for which it is violated. An analogous argument applies for the ICdown constraints.

The next definition formalizes a notion of a partially convex partition. It is instrumental in describing the way in which a given partition is iteratively transformed through a sequence of steps, parametrized by k, into a fully convex partition.

Definition 2. A partition of $\Omega = [0, 1]$ into messages is said to be "tightly packed with k messages" on the interval $[0, l]$ if:

1. The union of the first k messages is equal to $[0, l]$;
2. Each message $m_1, \dots, m_k$ is convex; and
3. The incentive compatibility constraints ICup$(1), \dots,$ ICup$(k-1)$ are all binding.

The next lemma characterizes the maximal number of messages that can be tightly packed on the interval $[0, l]$.

Lemma 2. *For any $l \in [0, 1]$, the maximal number of messages that can be tightly packed on the interval $[0, l]$ is given by $I(l) \equiv \left\lceil \sqrt{\frac{1}{4} + \frac{l}{2b-\alpha}} - \frac{1}{2} \right\rceil$.*[26]

As expected, if $\alpha = 0$ then $I(1)$ is also the number of intervals in the most informative equilibrium identified in the uniform quadratic example in Crawford and Sobel (1982).

The next proposition describes the optimal partition for Sender.

Proposition 1. *The optimal partition of $\Omega = [0, 1]$ consists of $I(1) = \left\lceil \sqrt{\frac{1}{4} + \frac{1}{2b-\alpha}} - \frac{1}{2} \right\rceil$ tightly packed messages on Ω.*

To prove Proposition 1 we provide an iterative convergent algorithm that improves upon any credible partition that does not partition the set Ω into $I(1)$ tightly packed messages. We now describe the algorithm and defer the detailed proof to the appendix.

Start with a credible partition that does not consist of $I(1)$ tightly packed messages on Ω. Let k be the highest index for which the messages $m_1, \dots, m_{k-1}$ are tightly packed on the interval $[0, l_{k-1}]$, where $l_j \equiv (|m_1| + \cdots + |m_j|)$ for any $j > 0$. Fig. 1(a) illustrates such a partition (notice that messages m_k, m_{k+1}, m_{k+2} are not convex). If no such collection of messages exists, then $k = 1$. If all the messages are already tightly packed, but the number of messages is less than $I(1)$, then k is equal to the number of messages in that partition.

The algorithm described above "packs message m_k" and produces a new partition in which $I(l_k)$ messages are tightly packed on the interval $[0, l_k]$, all the ICup constraints are satisfied, and the modified partition yields a higher value of the objective function (3) to Sender, compared to the original partition.

The algorithm consists of two parts. In Part I message m_k is "convexified to the left" through a series of shifts of states across messages until message m_k is convex and placed immediately to the right of message m_{k-1}. This change preserves the probability measure of all the messages. At the end of Part I of the algorithm, the partition takes the form depicted in Fig. 1(b). Intuitively, convexification to the left improves the value of the objective function (3) because it decreases the variance of some messages while not affecting the variance of others and not affecting the

[26] The function $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

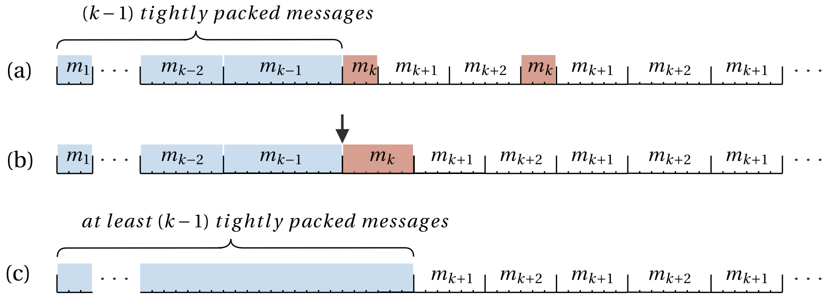

(k − 1) tightly packed messages
(a)
m1 . . . mk−2 mk−1 mk mk+1 mk+2 mk mk+1 mk+2 mk+1 . . .
(b)
m1 . . . mk−2 mk−1 mk mk+1 mk+2 mk+1 mk+2 mk+1 . . .
at least (k − 1) tightly packed messages
(c)
. . . mk+1 mk+2 mk+1 mk+2 mk+1 . . .

Fig. 1. The Convexification of Message m_k.

Algorithm Convexify and repack.

Require: Messages $m_1, ..., m_{k-1}$ are convex and tightly packed on the interval $\left[0, l_{k-1}\right]$

Part I - Convexify message m_k to the left

```
if message $m_k$ is not convex and adjacent to message $m_{k-1}$ then
    set $m_k = \left[\overline{m}_{k-1}, \overline{m}_{k-1} + |m_k|\right]$;
    for every $j > k$ such that $\underline{m}_j < \overline{m}_k$ do:
        replace every state $\omega \in m_j$ by the state $\omega + |[\omega, \overline{m}_k] \cap m_k|$;
end if
```

Part II - Repack

```
Repartition the interval $\left[0, l_k\right]$ into $I\left(l_k\right)$ tightly packed messages.
```

probabilities with which messages are sent. However, notice that after the change ICup$(k - 1)$ may no longer hold. This is because the convexification to the left of m_k decreases the mean μ_k, making it more attractive for higher types in m_{k-1} to deviate and report m_k. To restore incentive compatibility we proceed to the second part of the algorithm.

Part II of the algorithm repartitions the interval $[0, l_k]$ into $I\left(l_k\right)$ tightly packed messages. In the proof we show that this suffices to ensure that all the other ICup constraints are also satisfied, and in particular ICup$(I\left(l_k\right))$. The result of this part is depicted in Fig. 1(c). Note that it could be the case that $I\left(l_k\right) = k - 1$, so that repartitioning may in fact *decrease* the total number of messages. Nevertheless, in the proof we show that the overall effect of convexifying m_k to the left and repartitioning the interval $[0, l_k]$ improves the objective function's value.

If the partition generated by the algorithm does not consist of $I(1)$ messages that are tightly packed on Ω, then we apply the algorithm again on that partition. By Lemma 1, the number of messages in the original partition is bounded, therefore the process converges in a finite number of iterations.

In the proof we also show that whenever all the ICup constraints are binding, which is the case in any partition that consists of only tightly packed messages, then all the ICdown constraints are satisfied as well. Thus, the obtained partition, in which $I(1)$ messages are tightly packed on Ω is credible.

We conclude this section with a characterization of the messages that are induced by the optimal partition.

Corollary 1. *The optimal partition consists of* $I(1)$ *messages. Message* m_k, $k \in \{1, \ldots, I(1)\}$, *is given by the interval* $[\overline{m}_{k-1}, \overline{m}_k)$, *where*

$$\overline{m}_k = \frac{k}{I(1)} + 2\left(b - \frac{\alpha}{2}\right) k(k - I(1)).$$

Notably, the value of $\overline{m}_k$ that is described in the corollary is identical to the value characterized by Crawford and Sobel (1982), except that in the expression here, Sender's bias is offset by the cost parameter, so that instead of b in Crawford and Sobel's result, we have $b - \alpha/2$. Indeed, a higher value of the cost parameter α allows Sender to communicate her information about the state of the world more effectively, which increases her expected payoff.

4. Credibility with sender state-independent utility

In this section we incorporate our notion of credibility into a specification in which Sender's payoff is independent of the state. We impose the following three assumptions. First, we assume that the payoff-relevant part of the state ω is a real-valued random variable with a finite support. Next, we assume that the set of Receiver's optimal actions depends only on the expected state. That is, given a posterior belief p, the set $A^*(p)$ defined in (1) depends only on the mean of the distribution p, denoted by μ_p. Finally, we assume that Sender's preferences over Receiver's actions do not depend on the state, and to simplify notation, we henceforth omit the state from Sender's material payoff function.[27]

To facilitate a comparison between our model and that of Kamenica and Gentzkow (2011) (henceforth, KG) we start by writing Sender's problem as a constrained maximization problem over distributions of posterior beliefs.

We define the correspondence $\hat{u}_S : \Delta(\Omega) \twoheadrightarrow \mathbb{R}$ to be the *indirect payoff* of Sender. Thus, $\hat{u}_S(p)$ is the (set of) payoff(s) that Sender can achieve by inducing a posterior belief p to which Receiver best responds. For a belief p for which Receiver has a unique best response (i.e., $A^*(p)$ is a singleton), we have (with a slight abuse of notation),

$$\hat{u}_S(p) = u_S(A^*(p)). \tag{5}$$

Otherwise, if Receiver's best response at p is not unique, then $\hat{u}_S(p) = \{u_S(\tilde{a}) \; : \; \tilde{a} \in \Delta(A^*(p))\}$. The fact that u_R and u_S are continuous implies that $\hat{u}_S$ is a continuous correspondence.

Recall that every message m that is sent under a message function σ induces a posterior belief p_m^σ on the payoff-relevant part of the state ω. Accordingly, a message function σ induces a distribution of posterior beliefs. We denote such a distribution of posterior beliefs by $\tau \in \Delta(\Delta(\Omega))$, and the probability that τ induces a posterior $p \in \Delta(\Omega)$ by $\tau(p)$.

A distribution of posterior beliefs τ is said to be *Bayes plausible* if the expected posterior belief it induces is equal to the prior. As famously shown by (KG) and Aumann and Maschler (1995), a distribution of posterior beliefs τ can be induced by some message function σ if and only if τ is Bayes plausible.

Thus, we can rewrite Sender's problem (SP) as follows:

$$\max_{\tau} \quad \max_{\left\{v_p : v_p \in \hat{u}_S(p)\right\}_{p \in \text{Supp}(\tau)}} \quad \sum_{p \in \text{Supp}(\tau)} v_p \cdot \tau(p) \tag{SP1}$$

[27] Kamenica and Gentzkow (2011) refer to this case as one in which the "sender's payoff depends only on the expected state." This holds, for example, if $u_R(a, \omega) = -(a - \omega)^2$ and $u_S(a, \omega) = a$. It is easy to verify that in this case, for any belief p, Receiver's optimal action is given by μ_p. Thus, Sender's payoff from inducing the belief p is also μ_p. Other papers that employ state-independent Sender's preferences in large parts of their analysis include Guo and Shmaya (2021), Lipnowski and Ravid (2020), and Lipnowski et al. (2022).

$$\text{s.t.} \qquad \sum_{p\in \operatorname{Supp}(\tau)} p\cdot \tau(p) = \pi \qquad \text{(Bayes Plausibility)}$$

$$v_p \geq v_{p'} - \alpha \cdot d\left(p, p'\right), \quad \forall p, p' \in \operatorname{Supp}(\tau), \qquad \text{(Credibility)}$$

where $\operatorname{Supp}(\tau)$ denotes the support of τ, and v_p denotes Sender's equilibrium payoff under posterior belief p. When Receiver has a unique best response for any belief p, the indirect payoff $\hat{u}_S(\cdot)$ is a function, and so the inner maximum operator is degenerate because $v_p = \hat{u}_S(p)$ for all $p \in \operatorname{Supp}(\tau)$. In this case Sender's payoffs are pinned down by the induced beliefs. When Receiver has multiple best responses for some beliefs, the indirect payoff $\hat{u}_S(\cdot)$ is a correspondence. In this case, every payoff in $\hat{u}_S(p)$ can be obtained by some randomization performed by Receiver via an appropriately chosen tie-breaking rule. The inner maximum operator captures the fact that the equilibrium that is best for Sender employs the Sender-optimal tie-breaking rule.[28]

A distribution of posterior beliefs τ is said to be *feasible* if it is Bayes plausible and there exists a set of Sender's payoffs $\left\{v_p : v_p \in \hat{u}_S(p)\right\}_{p\in\operatorname{Supp}(\tau)}$ such that credibility is satisfied. Note that the "standard" problem of Bayesian persuasion involves maximizing the same objective function, under the same Bayes plausibility constraint (and with a degenerate inner maximum operator because in the Sender-optimal equilibrium, Receiver always breaks ties in favor of Sender, ex-post). The new component that is introduced in our costly deception framework is the credibility constraint.

We now proceed to characterize the solution to Sender's problem. Given Sender's indirect payoff $\hat{u}_S$, the convex hull of the graph of $\hat{u}_S$, denoted by $\operatorname{co}\left(\hat{u}_S\right)$, consists of all the convex combinations of elements in the graph of $\hat{u}_S$. That is,

$$\operatorname{co}\left(\hat{u}_S\right) = \left\{ \begin{array}{ll} (p,y): & \exists p_1,\ldots,p_k,\, p_i \in \Delta(\Omega) \text{ for all } i, \text{ and } \exists \lambda_1,\ldots,\lambda_k \geq 0, \sum_{i=1}^k \lambda_i = 1 \\ & \text{such that } p = \sum_{i=1}^k \lambda_i p_i \text{ and } y = \sum_{i=1}^k \lambda_i v_{p_i} \text{ where } v_{p_i} \in \hat{u}_S(p_i) \\ & \text{for all } i \end{array} \right\}.$$

Given $\alpha \geq 0$, we define the set $\operatorname{co}_\alpha\left(\hat{u}_S\right)$ similarly to $\operatorname{co}\left(\hat{u}_S\right)$, with one difference: it consists of all the convex combinations of elements in the graph of $\hat{u}_S$ that satisfy an additional set of pairwise restrictions that are parametrized by α:

$$\operatorname{co}_\alpha\left(\hat{u}_S\right) = \left\{ \begin{array}{ll} (p,y): & \exists p_1,\ldots,p_k,\, p_i \in \Delta(\Omega) \text{ for all } i, \text{ and } \exists \lambda_1,\ldots,\lambda_k \geq 0, \sum_{i=1}^k \lambda_i = 1 \\ & \text{such that } p = \sum_{i=1}^k \lambda_i p_i \text{ and } y = \sum_{i=1}^k \lambda_i v_{p_i} \text{ where } v_{p_i} \in \hat{u}_S(p_i) \\ & \text{for all } i \text{ and } \frac{|v_{p_j} - v_{p_i}|}{d(p_j,p_i)} \leq \alpha \text{ for every } i, j \end{array} \right\},$$

with the convention that $\frac{0}{0} = 0$, so that if $y \in \hat{u}_S(p)$ then $(p, y) \in \operatorname{co}_\alpha\left(\hat{u}_S\right)$ for all $\alpha \geq 0$.

If $y \in \hat{u}_S(p)$ then we say that p is the underlying posterior belief that induces y. The set $\operatorname{co}_\alpha\left(\hat{u}_S\right)$ contains all the pairs (p, y) for which the value y can be achieved by randomization over payoffs $\{v_{p_i}\}$ that are in the graph of $\hat{u}_S$, provided that: (i) the weights of the randomization are

[28] Note that as shown in Example 2 below, the tie-breaking rule employed by Receiver in Sender's ex-ante optimal equilibrium does not necessarily select the Sender's ex-post optimal action for every induced posterior belief. See also Lipnowski and Ravid (2020) and Lipnowski (2020) for how tie-breaking that does not benefit the sender ex-post, can be beneficial to the sender ex-ante.

such that the associated underlying posteriors average to p, and (ii) the randomization does not involve payoffs whose difference, divided by the distance between their underlying posteriors, is "too large" (i.e., exceeds α), which would make deception attractive to Sender.

Given $\alpha \geq 0$, define the *value* of belief p as follows[29]:

$$V(p, \alpha) \equiv \max \left\{ y : (p, y) \in \mathrm{co}_\alpha \left(\hat{u}_S \right) \right\}.$$

Given a prior belief π and a payoff $y \in \mathbb{R}$, if $(\pi, y) \in \mathrm{co}_\alpha \left(\hat{u}_S \right)$ then, by definition, there exists a Bayes plausible distribution τ that is credible and induces the expected payoff y. And, conversely, given π, if y can be induced by some Bayes plausible and credible distribution τ then $(\pi, y) \in \mathrm{co}_\alpha \left(\hat{u}_S \right)$. The next result follows immediately.

Proposition 2. *For every $\alpha \geq 0$, the highest value that Sender can achieve in the problem* (SP1) *is given by $V(\pi, \alpha)$.*

Higher deception costs expand the domain of message functions that are deemed credible, from which Sender can pick her preferred one. Indeed, as pointed out in Observation 1 in Section 2, higher deception costs are always *beneficial* for Sender: formally, if $\alpha < \alpha'$ then $\mathrm{co}_\alpha \left(\hat{u}_S \right) \subseteq \mathrm{co}_{\alpha'} \left(\hat{u}_S \right)$, which implies that $V(\pi, \alpha) \leq V(\pi, \alpha')$ for any prior π.

The structure of the set $\mathrm{co}_\alpha \left(\hat{u}_S \right)$ depends on the distance function d. As before, we assume that the distance between any two beliefs $p, p' \in \Delta(\Omega)$ is measured by the difference between the means induced by these distributions, i.e.,

$$d\left(p, p'\right) = \left|\mu_p - \mu_{p'}\right|. \tag{6}$$

Therefore, the cost of inducing the belief p', when the belief p should have been induced, is given by $\alpha \cdot \left|\mu_p - \mu_{p'}\right|$. Notice that if α is sufficiently large, then the credibility constraint is non-binding and Sender's problem becomes identical to that of the Bayesian persuader of KG that has full commitment power.

Under the specification of the distance function in Equation (6), the credibility constraint in Sender's problem (SP1) can be rewritten as follows:

$$\left| \frac{v_p - v_{p'}}{\mu_p - \mu_{p'}} \right| \leq \alpha \qquad \text{for all } p, p' \in \mathrm{Supp}(\tau) \tag{7}$$

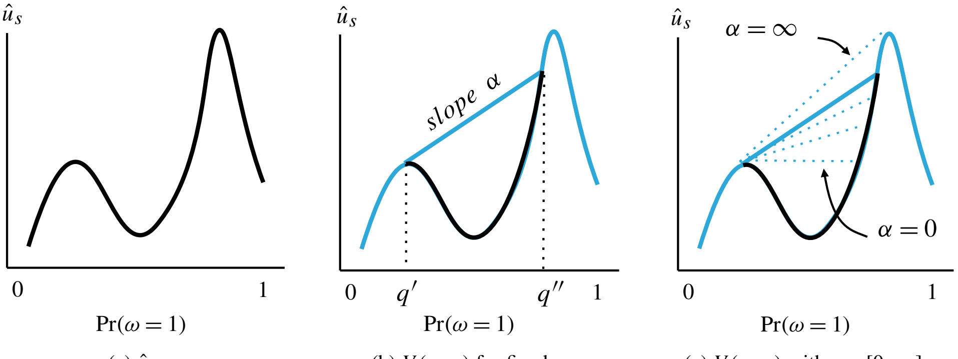
where v_p and $v_{p'}$ are Sender's equilibrium payoffs under the posterior beliefs p and p', respectively. It follows that for any two posterior beliefs that Sender induces in equilibrium, it must be the case that the material gain from deviating from one posterior belief to the other, divided by the distance between the means of the two posteriors, does not exceed α.

The following example provides a graphical illustration of the credibility constraint.

Example 1. Suppose that the payoff-relevant part of the state space is binary, with $\Omega = \{0, 1\}$. In this case, a distribution p over Ω can be represented by the probability q that the state is $\omega = 1$, and the mean of p is given by $\mu_p = q$.

Condition (7) has a geometric interpretation. To see it, consider the indirect payoff function $\hat{u}_S$ that is depicted in Fig. 2(a). Suppose that the prior distribution is given by some $\pi \in [0, 1]$. In Bayesian persuasion with full commitment ($\alpha = \infty$) Sender optimizes by "splitting" π into two

[29] The fact that $\hat{u}_S$ is continuous implies the existence of a maximum.



(a) $\hat{u}_S$ (b) $V(\pi, \alpha)$ for fixed α (c) $V(\pi, \alpha)$ with $\alpha \in [0, \infty]$

Fig. 2. A Geometric Illustration of the Credibility Constraint. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

probabilities, q' and q'', that are such that $\lambda \cdot q' + (1 - \lambda) \cdot q'' = \pi$ for some $\lambda \in [0, 1]$ (Bayes plausibility) so as to maximize the value of the objective function $\lambda \cdot \hat{u}_S \left(q'\right) + (1 - \lambda) \cdot \hat{u}_S \left(q''\right)$. The credibility constraint (7) implies that the *slope* of the line that connects the payoffs associated with these two probabilities, $\hat{u}_S \left(q'\right)$ and $\hat{u}_S \left(q''\right)$, cannot exceed α.

Fig. 2(b) depicts Sender's value $V(\pi, \alpha)$ (in blue) for a fixed deception cost α and different values of π (the horizontal axis depicts the probability $\Pr(\omega = 1)$ associated with each π). Note that the graph of this function consists of parts that coincide with the graph of $\hat{u}_S$ (when sending one uninformative message is optimal for Sender), and a line segment with slope α that connects points on the graph of $\hat{u}_S$ (when it is optimal for Sender to send messages that induce the posterior beliefs q' and q'' with probabilities that depend on π).

Fig. 2(c) illustrates what happens to $V(\pi, \alpha)$ when α is varied between zero and infinity. The uppermost dotted line in the figure corresponds to the graph of V when deception costs are infinite, i.e., $\alpha = \infty$ (or are just high enough to be non-binding). At the other extreme, the flat dotted line corresponds to the case where deception is costless. In that case Sender can only induce posterior beliefs that have identical indirect payoffs. This is the case that is analyzed by Lipnowski and Ravid (2020). The dotted lines in between correspond to different values of α, where higher lines correspond to higher values of α. ■

4.1. Continuity and discontinuity of Sender's value function

We now turn to discuss the continuity of the value function $V(\pi, \alpha)$. When V is discontinuous, Sender's expected payoff is highly sensitive to small changes in the parameters of the environment.

Our first observation is that Sender's value function may be discontinuous in the prior π. This is illustrated in the next example, which exploits a "jump" in Sender's payoff $\hat{u}_S$. Such jumps are typical when the set of Receiver's actions is finite.

Example 2. Suppose that there are two states, $\Omega = \{0, 1\}$, with prior probability π on state $\omega = 1$. Receiver chooses one of two actions $a \in \{0, 1\}$ and his payoff function is given by $u_R(a, \omega) = -|a - \omega|$. Sender's payoff function is given by $u_S(a, \omega) = a$. Thus, Sender's indirect payoff $\hat{u}_S$ is given by the bold line depicted in Fig. 3(a), where the values on the horizontal axis represent the belief that $\omega = 1$.

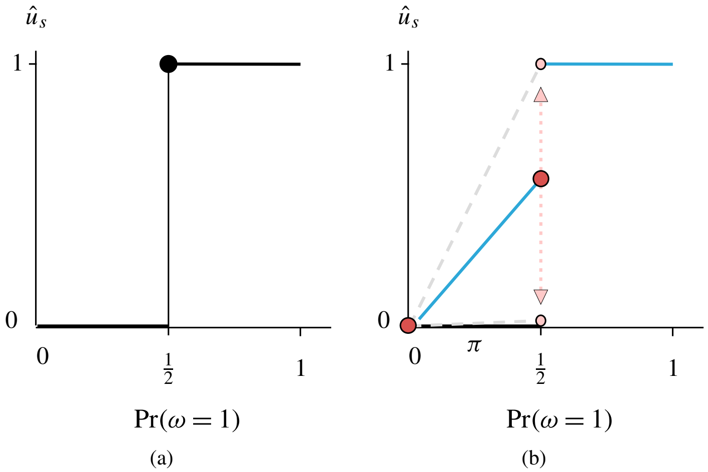


Fig. 3. α-dependent tie breaking.

If $\pi > \frac{1}{2}$, then Sender-optimal equilibrium is uninformative. In this equilibrium, Sender sends the same message in both states, and Receiver takes the action $a = 1$, which produces a payoff 1 for Sender. Otherwise, if $\pi \leq \frac{1}{2}$, then in the optimal equilibrium Sender induces the posterior beliefs 0 and $\frac{1}{2}$ with probabilities $1 - 2\pi$ and 2π, respectively. When his posterior belief is 0, Receiver takes the action $a = 0$. When his posterior belief is $\frac{1}{2}$, Receiver is indifferent. In this case, he takes the action $a = 1$ with probability $\min\{\frac{\alpha}{2}, 1\}$ and the action $a = 0$ with the complementary probability. It is not difficult to verify that this is the Sender-optimal equilibrium, and that Sender's expected payoff in this case is given by

$$V(\pi, \alpha) = \begin{cases} \pi \cdot \min\{\alpha, 2\} & \text{if } \pi < \frac{1}{2} \\ 1 & \text{if } \pi \geq \frac{1}{2} \end{cases} .$$

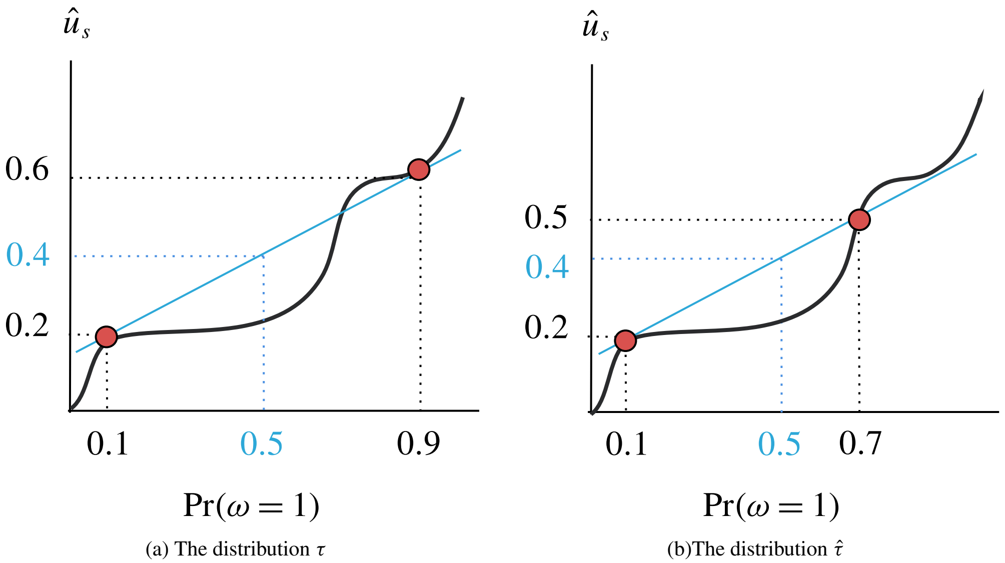
This function is discontinuous in π at the point $\pi = \frac{1}{2}$ for values of $\alpha < 2$. ■

We thus obtain,

Observation 2. Sender's value $V(\pi, \alpha)$ may be discontinuous in the prior π (when the value of α is not too large).

Inspection of Sender's value $V(\pi, \alpha)$ in Example 2 reveals that it is continuous in the cost parameter α. This continuity is facilitated by a judicious use of the tie-breaking rule employed by Receiver when the posterior belief on $\omega = 1$ is $\frac{1}{2}$. Fig. 3(b) depicts the way in which Sender's equilibrium payoff from inducing posterior belief $p = \frac{1}{2}$ varies continuously with α in a way that preserves the continuity of $V(\pi, \alpha)$ in α for every π.

Continuity in α holds more generally. The main challenge in proving this result stems from the fact that the correspondence that maps the parameters (π, α) into the set of feasible distributions τ is not lower hemi-continuous in α (and therefore Berge's maximum theorem does not apply in our case). This implies that for a given distribution of posterior beliefs τ, a small change in α may imply that there is no feasible distribution of posterior beliefs in the neighborhood of τ. This is illustrated in the next example.



(a) The distribution τ

(b)The distribution $\hat{\tau}$

Fig. 4. Continuity of the Distribution of Posteriors.

Example 3. Suppose that there are two states, $\Omega = \{0, 1\}$, with equal prior probabilities. As before, we represent a distribution over Ω by the probability that the state is $\omega = 1$. Assume that the prior probability is given by $\pi = 0.5$, that $\alpha = 0.5$, and that the function $\hat{u}_S$ is given by the bold curved line depicted in Fig. 4(a). The optimal credible distribution of posterior beliefs in this example is $\tau = (0.1, 0.9; \frac{1}{2}, \frac{1}{2})$ and it gives Sender an expected value that is equal to 0.4. To verify that credibility is satisfied, notice that the slope of the line that connects the two points $(0.1, 0.2)$ and $(0.9, 0.6)$ (depicted in light blue) is 0.5, which is equal to α.

Suppose now that α is slightly decreased. It is impossible to find two posterior beliefs close to 0.1 and 0.9, respectively, that satisfy the credibility constraint (i.e., such that the line that connects the two points associated with these posteriors has a slope less than or equal to the new value of α, which is less than 0.5). Thus, the correspondence that maps the parameters (π, α) into the set of feasible distributions is not lower hemi-continuous at $(\pi, \alpha) = (0.5, 0.5)$. ■

To overcome this difficulty, we show that given π, even if a feasible distribution τ is such that for some small change in α there is no feasible distribution that is close to τ, then there must exist another feasible distribution $\hat{\tau}$, that achieves the same expected value for Sender as τ, and $\hat{\tau}$ is such that for small changes in α there is a feasible distribution that is close to $\hat{\tau}$.

Example 3 (continued). As illustrated in Fig. 4(b), there exists a feasible distribution $\hat{\tau} = (0.1, 0.7; \frac{1}{3}, \frac{2}{3})$ that generates the same expected value for Sender of 0.4 as $\tau = (0.1, 0.9; \frac{1}{2}, \frac{1}{2})$. Note that for any parameters α' that is close to $\alpha = 0.5$, there exists a distribution of posterior beliefs that is feasible with respect to $(\pi = 0.5, \alpha')$ and is close to $\hat{\tau}$. For example, it is possible to pick a binary distribution of posterior beliefs that is supported on 0.1 and $0.7 - \varepsilon$ for some small $\varepsilon > 0$ that depends on α' and the curvature of the function $\hat{u}_S$. ■

We thus obtain the following result.

Proposition 3. *For any prior π, Sender's value $V(\pi, \alpha)$ is continuous in α.*

Two remarks are in order. First, the possible discontinuity of Sender's value $V(\pi, \alpha)$ in π stands in contrast to the continuity of the value function in standard Bayesian persuasion (which is equivalent to the case where $\alpha = \infty$). This is because $V(\pi, \infty)$ is the concave closure of the continuous correspondence $\hat{u}_S$, and is therefore continuous.

Second, recall that the cost of deception parameter α captures Sender's ability to commit to her message function. Proposition 3 thus asserts that Sender's value is continuous in Sender's commitment power. This result stands in contrast to the possible discontinuity of Sender's value in Sender's credibility in the model of Lipnowski et al. (2022).[30]

4.2. The effect of the cost of deception

As the cost of deception α decreases, the credibility constraint becomes tighter, and so the set of message functions that Sender can employ in equilibrium shrinks. Sender can restore her credibility by either adopting a message function in which deception is more costly, or by adopting a message function in which the gain from deception is smaller. In this subsection we discuss these two alternatives.

The next example shows how Sender can increase the cost of deception in response to a lower value of α by moving the means of the induced posterior beliefs farther apart.[31]

Example 4. Suppose that the payoff structure is the same as in Example 2 and that $\pi = \frac{1}{4}$. For values of $\alpha \in [1, 2]$ consider the following equilibria: Sender induces posterior beliefs 0 and $1/\alpha$ that the state is $\omega = 1$, with probabilities $1 - \frac{\alpha}{4}$ and $\frac{\alpha}{4}$, respectively. Receiver takes the action $a = 0$ when his posterior belief is 0, and the action $a = 1$ when his posterior belief is $1/\alpha$. It is easy to verify that this equilibrium is Sender-optimal, and that as α decreases (but is still above 1), the posterior beliefs that support the optimal distribution τ move farther apart. See Fig. 5(a). Intuitively, this movement increases the cost of deception and preserves the credibility of Sender's message function. This movement has the effect of "ungarbling" Sender's communicated information. This ungarbling allows Receiver to make a more informed choice and so increases Receiver's ex-ante expected payoff. ■

The ungarbling of Sender's communication in Example 4 is in the same spirit of what Lipnowski et al. (2022) refer to as "productive mistrust" (but for a different reason). Namely, a decrease in Sender's ability to commit gives rise to a more informative equilibrium, which makes Receiver better off.

[30] The difference in our results is due to the fact that in Lipnowski et al.'s model, Sender compensates for a weakening of (their notion of) credibility by sending a more informative signal. When credibility falls below a certain threshold, this is no longer sustainable, leading to a discontinuous jump in Sender's payoff. By contrast, in our model, Receiver's tie-breaking rule directly affects Sender's gain from deception. When α decreases, Sender's equilibrium gain from deception can be adjusted continuously, which preserves the continuity of Sender's payoff.

[31] Example 4 depicts a family of Sender-optimal equilibria (one for each value of $\alpha \in [1, 2]$) in which the induced posterior beliefs are 0 and $1/\alpha$. These are not the only Sender-optimal equilibria for the given payoff structure. Indeed, Example 2 depicts another family of Sender-optimal equilibria in which the induced posterior beliefs are 0 and $1/2$ for these values of α. Notice however that for any such value of α, the former equilibrium Pareto dominates the latter because it is strictly better for Receiver, and provides the same expected payoff to Sender.

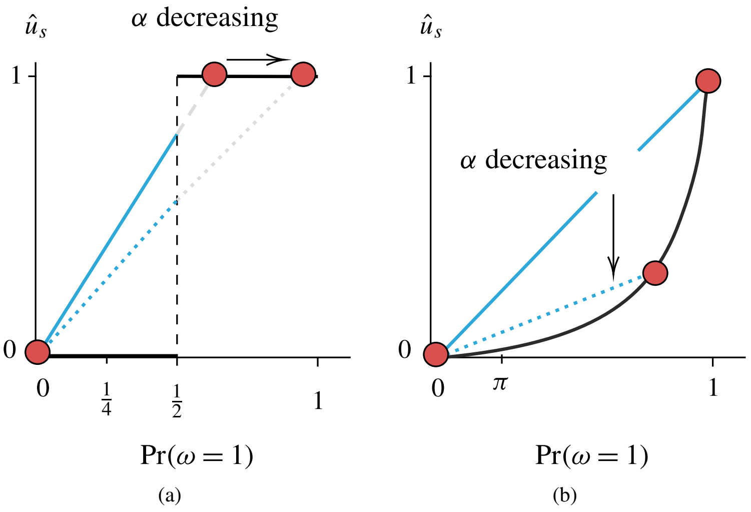


Fig. 5. Effects of changing α: (a) ungarbling; (b) garbling.

The other way in which Sender can respond to a decrease in the value of α is by decreasing the gain from deception. Sometimes, the only way to achieve this is by *additional* garbling of Sender's messages that moves the means of the induced posteriors closer together. Whether or not garbling or ungarbling is better for Sender depends on the specific context. The next proposition describes a sufficient condition that ensures that Sender responds to a lower value of α by garbling her message to Receiver.

Proposition 4. *Suppose that the state space Ω is binary, Receiver has a unique best response for every belief, and Sender's indirect payoff $\hat{u}_S$ is strictly convex.*[32] *If $\alpha' > \alpha$, then Sender's optimal distribution of posterior beliefs under α is a garbling of the optimal distribution under α'. Consequently, lower deception costs are weakly harmful for both Sender and Receiver.*

Fig. 5(b) depicts the case of a convex indirect payoff function $\hat{u}_S$ and illustrates that a lower α results in a distribution τ that is supported on posterior beliefs that are closer together.

Broadly speaking, a lower cost of deception implies that it is more difficult for Sender to commit and so is accompanied by a higher level of mistrust. To appreciate the effect of mistrust it is useful to observe that Sender faces a tension between his incentive to reveal and conceal information to Receiver. Receiver *always* prefers all information to be revealed. Example 4 depicts a situation where Sender's and Receiver's interests are sufficiently opposed for Receiver to benefit from Sender's difficulty to commit. Proposition 4 depicts a situation where Sender's and Receiver's interests are sufficiently aligned for both Sender and Receiver to suffer from Sender's difficulty to commit. In the former case, Sender does not disclose all the available information and, in order to preserve her credibility, she discloses more information; in the latter case, Sender prefers to fully disclose all the available information and, in order to preserve her credibility, she has to disclose less than she would want to if she was trusted by Receiver. Notice however that in the two extreme cases in which Sender's and Receiver's interests are perfectly aligned and

[32] These conditions imply that $\hat{u}_S$ is a strictly convex *function*.

diametrically opposed with respect to the revelation of information, a change in the value of α makes no difference. In the case of opposed interests, silence on Sender's part is always credible and optimal. And, when Sender and Receiver have perfectly aligned interests, Sender would not want to mislead Receiver anyway.

5. Discussion

5.1. Tie-breaking by receiver

Our focus in this paper is on Sender's (ex-ante) optimal equilibrium. It is noteworthy that in order to support this equilibrium, Receiver may sometimes have to break ties in a way that "helps" Sender to sustain her credibility, but is suboptimal for Sender ex-post. This is illustrated in Example 2, where for values of $\pi < \frac{1}{2}$ and $\alpha < 1$, when his posterior belief on $\omega = 1$ is $\frac{1}{2}$, Receiver randomizes between the two actions in a way that allows Sender to credibly communicate information. This is sub-optimal for Sender ex-post because Sender prefers action $a = 1$ to $a = 0$. Furthermore, in the Sender-optimal equilibrium, Receiver may need to randomize differently for different posterior beliefs.

The fact that Receiver's action is not optimal for Sender *ex-post*, may open the door for Sender to ask an indifferent Receiver to change his action. This raises the question of what would be the effect on equilibrium if, when indifferent, Receiver always chose the action that is optimal for Sender ex-post.

Analysis of this case reveals two notable differences from the analysis so far. First, Sender's value in this case may be discontinuous in α. For instance, in the setup of Example 2, suppose that the prior belief is $\pi = \frac{1}{4}$ and that Receiver always chooses the action $a = 1$ when his posterior belief is $\frac{1}{2}$. Then, Sender's value is discontinuous in α at $\alpha = 1$. Second, in some settings the Sender's optimal value may be approximated arbitrarily closely, but not exactly achieved.[33] These two observations would hold for any other tie-breaking rule in which Receiver randomizes with probabilities that are independent of the value of α.

5.2. An upper bound on the number of messages

How many messages does Sender need to employ in equilibrium? Suppose that the number of states in Ω is finite. Previous work (see, e.g., Le Treust and Tomala, 2019, Doval and Skreta, 2018, and Salamanca, 2021) has shown that, in general, the number of messages optimally employed by the sender in constrained communication problems is possibly larger than the number of states and is increasing in the number of constraints. In our setting, the number of constraints is *endogenous*: if the number of messages sent by Sender is K, then the number of (credibility) constraints is $K(K-1)$.

Hence, it is not a priori clear what is the number of messages that are required to support a Sender-optimal equilibrium. On the one hand, employing a small number of messages decreases the number of credibility constraints. On the other hand, it may be the case that the way to achieve the Sender-optimal payoff is to employ a large number of messages such that the gain from deviating from one message to another is small.

[33] For instance, suppose that in Example 2 Receiver has an additional action $a = 2$ with a payoff $-|0.5 - \omega| - 0.5$ that gives Sender a payoff of 2. Receiver's action in this case is identical to that in Example 2, except that when his belief that $\omega = 1$ is 0.5 he chooses the action $a = 2$. In this case, for $\alpha = 2$ Sender's value can be approximated but not achieved.

The next proposition shows that no loss of optimality is implied by restricting attention to equilibria that employ no more messages than the number of states $|\Omega|$.[34]

Proposition 5. *For any equilibrium that employs more than $|\Omega|$ messages, there exists an equilibrium that generates a weakly higher ex-ante expected payoff to Sender and employs no more than $|\Omega|$ messages.*

The proof of the proposition starts with the well-known observation that, by Carathéodory's theorem (Rockafellar, 1997), for any message function there exists another (possibly non-credible) message function that generates an identical ex-ante expected payoff with no more than $|\Omega| + 1$ messages.[35] The challenge is to show that given a *credible* message function that employs more messages, it is possible to reduce the number of messages in such a way that preserves credibility. To prove this, we show that in the process of reducing the number of messages, it is never the case that a message that was not sent in state ω under the original message function is sent in ω under the message function with the smaller number of messages. Finally, we rely on the observation that the expected payoff that is generated by an equilibrium that employs $|\Omega| + 1$ messages can be written as an average of the expected payoffs generated by two equilibria that each employs no more than $|\Omega|$ messages.

We thus obtain,

Corollary 2. *There exists a Sender-optimal equilibrium that employs no more than $|\Omega|$ messages.*

6. Conclusion

Credible communication establishes trust, which is crucial for social interaction and economic activity. In this paper we propose a theoretical foundation for understanding the link between deception and credible communication. We introduce the possibility of costly deception into communication games. The novelty in our approach is that deception costs depend on the players' beliefs and are therefore endogenous. We show how costly deception affects Sender's ability to commit to her strategy in two classical environments. We illustrate how imperfect commitment is mitigated by the fact that deception is costly. The model is tractable and can be applied to many economic environments. These environments can be either ones in which the informed party incurs a direct cost for being dishonest (e.g., because of ethical concerns, as in a doctor-patient or familial relationships) or ones in which the cost is indirect, and is due to reputational concerns in a reduced form of a dynamic interaction model (e.g., as in the case of central bankers, politicians, and the media). Pursuing these directions for future research would hopefully enhance the understanding of the determinants of credible communication and its effect on social trust.

Data availability

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

[34] Bester and Strausz (2001) and Heumann (2020) obtain a similar result for the case with no commitment. In our model, this is the case of $\alpha = 0$.

[35] The use of Carathéodory's theorem in such problems is standard (see, e.g., Bester and Strausz, 2001, and Kamenica and Gentzkow, 2011). In our case, the credibility constraint implies that a subtler argument is required.

Appendix A. Proofs

A.1. Proof of Lemma 1

Suppose that $\alpha < 2b$. Fix a credible partition. Recall that Receiver's optimal response to any message m is given by μ_m. For any two messages m and m' such that $\mu_{m'} > \mu_m$, condition (4) can be rewritten as:

$$\frac{\mu_m + \mu_{m'}}{2} - \omega \geq b - \frac{\alpha}{2} \tag{8}$$

for all $\omega \in m$. Suppose that there are N or more messages. There exist two messages m and m' such that $\mu_{m'} - \mu_m < \frac{1}{N}$. It follows that when applied to messages m and m', the left-hand side of (8) is smaller than $\mu_m - \omega + \frac{1}{2N}$. There exists a state $\omega \in m$ that is such that $\omega \geq \mu_m$. Therefore, for $N > \frac{1}{2b-\alpha}$ inequality (8) is violated. It follows that the number of messages is no larger than $\frac{1}{2b-\alpha}$.

Suppose now that $\alpha \geq 2b$. If Sender reveals the state then each message $m = \{\omega_m\}$ is a singleton state and $\mu_m = \omega_m$. Inspection of the constraint (4) reveals that it is satisfied. Moreover, since the variance of each message is zero, the objective function given by Equation (3) is also equal to zero, which is the highest possible value it can attain.

A.2. Proof of Lemma 2

Suppose that m_k and m_{k+1} are two adjacent convex messages. Convexity implies that $\mu_k = \frac{\underline{m}_k+\overline{m}_k}{2}$ and $\mu_{k+1} = \frac{\underline{m}_{k+1}+\overline{m}_{k+1}}{2}$. The incentive constraint ICup(k) is then given by:

$$\frac{\underline{m}_k + \overline{m}_k}{2} + \frac{\underline{m}_{k+1} + \overline{m}_{k+1}}{2} - 2\overline{m}_k \geq 2b - \alpha.$$

Convexity also implies that $\overline{m}_k - \underline{m}_k = |m_k|$ and $\overline{m}_{k+1} - \underline{m}_{k+1} = |m_{k+1}|$. Since the messages are adjacent, it follows that $\underline{m}_{k+1} = \overline{m}_k$. We can therefore equivalently write ICup(k) as follows:

$$|m_{k+1}| - |m_k| \geq 4b - 2\alpha. \tag{9}$$

Equation (9) is a necessary and sufficient condition for ICup(k) when messages are convex. If, in addition, ICup(k) is binding then Equation (9) holds in equality.

Set $x = |m_1| > 0$. Then, the fact that the messages are tightly packed implies that $|m_2| = x + 4b - 2\alpha$, $|m_3| = x + 8b - 4\alpha, \ldots, |m_k| = x + (k-1)(4b - 2\alpha)$, and so on. Thus,

$$|m_1| + \cdots + |m_k| = k(k-1)\,(2b - \alpha) + kx. \tag{10}$$

Since x can be set arbitrarily small, the maximal number of messages that can be tightly packed on the interval $[0, l]$ is given by

$$I\,(l) = \left\lceil \sqrt{\frac{1}{4} + \frac{l}{2b - \alpha}} - \frac{1}{2} \right\rceil.$$

A.3. Proof of Proposition 1

Consider a credible partition that does not consist of $I\,(1)$ tightly packed messages on Ω. Suppose that the number of messages in this partition is J. The algorithm described in the text

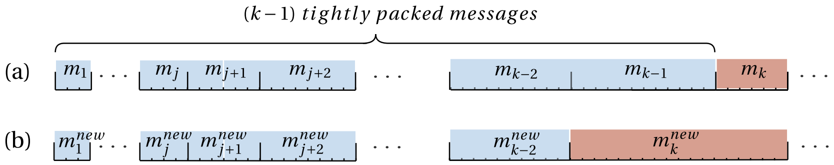


Fig. 6. Step1, Case I.

"packs message m_k" and produces a new partition in which $I(l_k)$ messages are tightly packed on the interval $[0, l_k]$. We show that in *each iteration* of the algorithm the value of the objective function improves and all ICup constraints are preserved.

Our proof proceeds in two steps. In step 1, we show that performing Part I of the algorithm improves the value of the objective function. Furthermore, after Part I is performed all but perhaps one of the ICup constraints are satisfied. If this one constraint is indeed violated, we perform a modification of the partition after which: (i) *all* the ICup constraints are satisfied, and (ii) the objective function's value is higher than that of the original partition (before the execution of Part I of the algorithm).

In step 2, we show that the partition produced in step 1 is in fact suboptimal relative to a partition in which messages are tightly "repacked" in a maximal manner, and in which all ICup constraints are satisfied. Steps 1 and 2 can be repeated until the resulting partition is one that consists of $l(1)$ tightly packed messages on Ω.

To conclude the proof, we show that this final partition satisfies all the ICdown constraints, and is therefore credible.

Step 1: Fix all ICup constraints and improve the objective function's value

Part I of the algorithm "convexifies" message m_k to the left. The outcome of this process is illustrated in Fig. 6(a). We refer to the partition before the convexification as the "original partition" and to the partition after the convexification as the "convexified partition." Since in the convexified partition the variance of each message m_j is weakly less than the variance of m_j under the original partition, it follows that the convexified partition attains a higher value for the objective function (3).

Notice that the convexification (as described in the algorithm in the text) shifts probability mass of message m_k towards lower states ("to the left") and shifts probability masses of messages m_j with $j > k$ towards higher states ("to the right"). Define ϕ_j to be the increase in the mean of message m_j as a consequence of the shift, multiplied by the probability of the message $|m_j|$ (i.e., the increase in the mean of message m_j is equal to $\frac{\phi_j}{|m_j|}$). Define Φ_k to be the decrease in the mean of message m_k as a consequence of the shift, multiplied by $|m_k|$ (i.e., the decrease in the mean of message m_k is equal to $\frac{\Phi_k}{|m_k|}$). By the Law of Iterated Expectation, $\Phi_k = \sum_{j>k} \phi_j$. Finally, denote by L the decrease in the value of $\overline{m}_k$ as a consequence of the shift (thus, L is equal to the difference between the measures $[\overline{m}_{k-1}, \overline{m}_k]$ and $|m_k|$). Because the decrease in the value of $\overline{m}_k$ is at least as large as the decrease in the mean of message m_k, it follows that:

$$\frac{\Phi_k}{|m_k|} \leq L. \tag{11}$$

In the convexified partition, the constraints ICup(1), ..., ICup($k-2$) are satisfied because the convexification does not affect them. The constraints ICup($k+1$), ..., ICup($J-1$) are also satisfied. To see this, note that credibility of the original partition implies that $\mu_j \leq \overline{m}_j \leq \mu_{j+1} \leq$

$\overline{m}_{j+1}$ for all $j \le J-1$ and therefore $\overline{m}_{k+1}, \ldots, \overline{m}_J$ are all larger than $\overline{m}_k$. Hence, the supremum of each message m_j with $j > k$ is unchanged between the original and the convexified partition, i.e., the values of $\overline{m}_{k+1}, \ldots, \overline{m}_J$ are unaffected by the convexification. Moreover, the values of $\mu_{k+1}, \ldots, \mu_J$ are all weakly greater in the convexified partition relative to the original one. Therefore, the fact that ICup$(k+1)$, ..., ICup$(J-1)$ are satisfied in the original partition implies that they are satisfied in the convexified partition as well.

In the convexified partition, the constraint ICup(k) is satisfied with a slack. To see this, notice first that the convexification weakly increases μ_{k+1} relative to its value in the original partition. Next, note that although the convexification decreases μ_k by $\frac{\Phi_k}{|m_k|}$ relative to the original partition, it also decreases $\overline{m}_k$ by L. By (11) it follows that the sum $\mu_k + \mu_{k+1}$ decreases by no more than $\frac{\Phi_k}{|m_k|}$ while $\overline{m}_k$ decreases by at least $\frac{\Phi_k}{|m_k|}$. Thus, the fact that ICup(k) is satisfied in the original partition implies that it is satisfied also in the convexified partition. In fact, the convexification creates a slack of at least $\frac{\Phi_k}{2|m_k|}$ in the ICup(k) constraint. We make use of this observation below.

If ICup$(k-1)$ is satisfied in the convexified partition, then all ICup constraints are satisfied. In this case, jump directly to step 2 below. Otherwise, we distinguish between two cases.

Case I. Suppose that $|m_{k-1}| \le \frac{2\Phi_k}{|m_k|}$. Merge message m_{k-1} and message m_k (which is now a convex message) into a new message called m_k^{new} with mean μ_k^{new}. For ease of notation we relabel all the other messages from m_j as m_j^{new}. We refer to the resulting partition as the "merged partition." This partition, which is illustrated in Fig. 6(b), is composed of the messages $m_1^{new}, \ldots, m_{k-2}^{new}, m_k^{new}, m_{k+1}^{new}, \ldots, m_J^{new}$. Notice that:

$$\mu_k^{new} = \mu_k - \frac{\Phi_k}{|m_k|} - \frac{|m_{k-1}|}{2} \tag{12}$$

$$\mu_k^{new} = \mu_{k-1} + \frac{|m_k|}{2} \tag{13}$$

$$\mu_j^{new} = \mu_j + \frac{\phi_j}{|m_j|} \qquad \text{for all } j \ge k+1 \tag{14}$$

$$\mu_j^{new} = \mu_j \qquad \text{for all } j \le k-2 \tag{15}$$

$$\overline{m}_k^{new} = \overline{m}_k - L, \tag{16}$$

where μ_j is the mean of message m_j in the original partition, for all j. To see why Equation (12) holds, notice that μ_k^{new} is equal to the original value of μ_k, minus $\frac{\Phi_k}{|m_k|}$ (due to the convexification of m_k), minus $\frac{|m_{k-1}|}{2}$ (due to the merging of m_k with m_{k-1}). Equation (13) holds because the mean of the merged message m_k^{new} is larger than that of the original m_{k-1} by $\frac{|m_k|}{2}$. Equations (14), (15), and (16) are all direct implications of the convexification of m_k.

In the merged partition, all the ICup constraints are satisfied:

1. ICup$((k-2)^{new})$ is satisfied because $\mu_k^{new} > \mu_{k-1}$, whereas $\mu_{k-2}^{new} = \mu_{k-2}$ and $\overline{m}_{k-2}^{new} = \overline{m}_{k-2}$. Therefore, the fact that ICup$(k-2)$ was satisfied in the original partition, i.e. $\frac{\mu_{k-2}+\mu_{k-1}}{2} - \overline{m}_{k-2} \ge b - \frac{\alpha}{2}$, implies that $\frac{\mu_{k-2}^{new}+\mu_k^{new}}{2} - \overline{m}_{k-2}^{new} \ge b - \frac{\alpha}{2}$.
2. ICup(k^{new}) is satisfied because, by Equation (12) and since $|m_{k-1}| \le \frac{2\Phi_k}{|m_k|}$, we have that $\mu_k^{new} \ge \mu_k - \frac{2\Phi_k}{|m_k|}$. Thus, the facts that ICup(k) was satisfied in the original partition, i.e., $\frac{\mu_k+\mu_{k+1}}{2} - \overline{m}_k \ge b - \frac{\alpha}{2}$, along with Equations (11), (14), and (16), imply that $\frac{\mu_k^{new}+\mu_{k+1}^{new}}{2} - \overline{m}_k^{new} \ge b - \frac{\alpha}{2}$.

3. All the other ICup constraints are unaffected by the merge. The fact that they are satisfied in the convexified partition implies that they are satisfied in the merged partition.

We now show that the merged partition yields a higher value of the objective function (3) compared to the original partition. Algebraic manipulation shows that the objective function (3) is equal to the weighted sum of square means of the partition elements

$$\sum_{j=1}^{J} |m_j| \left(\mu_j\right)^2 \tag{17}$$

up to a constant. We therefore have to show that:

$$\begin{aligned} &\sum_{j\le k-2} |m_j|(\mu_j^{new})^2 + |m_k^{new}| \left(\mu_k^{new}\right)^2 + \sum_{j\ge k+1} |m_j|(\mu_j^{new})^2 \\ \ge &\sum_{j\le k-2} |m_j|\mu_j^2 + |m_{k-1}|\mu_{k-1}^2 + |m_k|\mu_k^2 + \sum_{j\ge k+1} |m_j|\mu_j^2 \end{aligned}$$

where the left-hand side of the inequality is the value of (17) computed for the merged partition, and the right-hand side is the value of (17) computed for the original partition. Using Equations (14) and (15) above, we rewrite the inequality as follows:

$$2\sum_{j\ge k+1} \mu_j \phi_j + \sum_{j\ge k+1} |m_j| \left(\frac{\phi_j}{|m_j|}\right)^2 \ge |m_{k-1}|\mu_{k-1}^2 + |m_k|\mu_k^2 - |m_k^{new}| \left(\mu_k^{new}\right)^2.$$

Since $\sum_{j\ge k+1} |m_j| \left(\frac{\phi_i}{|m_j|}\right)^2 \ge 0$ and $\mu_j > \mu_{k+1}$ for any $j > k+1$, it suffices to show that:

$$2\Phi_k \mu_{k+1} \ge |m_{k-1}|\mu_{k-1}^2 + |m_k|\mu_k^2 - |m_k^{new}| \left(\mu_k^{new}\right)^2.$$

Substituting $|m_k^{new}| = |m_k| + |m_{k-1}|$ and rearranging yields:

$$2\Phi_k \mu_{k+1} \ge -|m_{k-1}| \left(\mu_k^{new} - \mu_{k-1}\right)(\mu_{k-1} + \mu_k^{new}) + |m_k| \left(\mu_k - \mu_k^{new}\right)(\mu_k + \mu_k^{new}).$$

Using Equations (12) and (13), we rewrite the inequality as follows:

$$2\left(\mu_{k+1} - \mu_k\right)\Phi_k \ge \frac{1}{2}|m_k|\,|m_{k-1}| \left(\frac{\Phi_k}{|m_k|} + \frac{|m_{k-1}|}{2} + \frac{|m_k|}{2}\right) - \Phi_k \left(\frac{\Phi_k}{|m_k|} + \frac{|m_{k-1}|}{2}\right). \tag{18}$$

Finally, we use the fact that ICup(k) is satisfied in the original partition to find a lower bound on $\mu_{k+1} - \mu_k$. To do that, we write ICup(k) equivalently as follows:

$$\mu_{k+1} - \mu_k \ge 2\left(\left(\overline{m}_k^{new} - \mu_k^{new}\right) - \left(\mu_k - \mu_k^{new}\right) + \left(\overline{m}_k - \overline{m}_k^{new}\right)\right) + 2\left(b - \frac{\alpha}{2}\right).$$

The fact that m_k^{new} is a convex message with measure $|m_{k-1}| + |m_k|$ implies that $\overline{m}_k^{new} - \mu_k^{new} = \frac{1}{2}\left(|m_{k-1}| + |m_k|\right)$. By Equations (11), (12), and (16) we have that

$$\mu_{k+1} - \mu_k \ge |m_k| + 2\left(b - \frac{\alpha}{2}\right). \tag{19}$$

By plugging inequality (19) into inequality (18) and simplifying we only need to show that:

$$|m_k|^2 |m_{k-1}|^2 + |m_k|^3 |m_{k-1}| - 8\Phi_k |m_k|^2 - 4\Phi_k^2 - 16\Phi_k |m_k| \left(b - \frac{\alpha}{2}\right) \le 0. \tag{20}$$

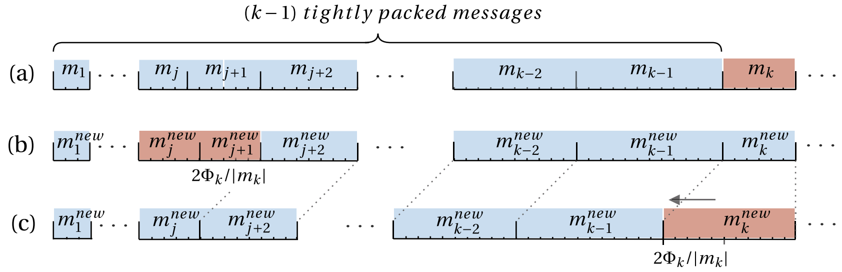


Fig. 7. Step 1, Case II.

Finally, to see why this last inequality is satisfied, notice that the left-hand side of the inequality is negative when $|m_{k-1}| = 0$. A simple computation shows that it is negative also for $|m_{k-1}| = \frac{2\Phi_k}{|m_k|}$. Because the left-hand side of the inequality is convex in $|m_{k-1}|$, it is negative for any $|m_{k-1}| \in \left[0, \frac{2\Phi_k}{|m_k|}\right]$.

Case II. Suppose that $|m_{k-1}| > \frac{2\Phi_k}{|m_k|}$. Find the index $0 \le j < k-1$ for which $\left|m_j\right| < \frac{2\Phi_k}{|m_k|} \le \left|m_{j+1}\right|$. For ease of notation assume that $m_0 = \emptyset$ and $|m_0| = 0$. Repartition the union of the two messages $m_j \cup m_{j+1}$ into two new messages: an interval $m_j^{new} = \left[\overline{m}_{j-1}, \overline{m}_{j+1} - \frac{2\Phi_k}{|m_k|}\right]$ with measure $|m_j^{new}| = \left|m_j\right| + \left|m_{j+1}\right| - \frac{2\Phi_k}{|m_k|}$ and an interval $m_j^{new} = \left[\overline{m}_{j+1} - \frac{2\Phi_k}{|m_k|}, \overline{m}_{j+1}\right]$ with measure $|\overline{m}_{j+1}^{new}| = \frac{2\Phi_k}{|m_k|}$. Relabel all the other messages from m_j as m_j^{new}, as illustrated in Fig. 7(b). This modified partition weakly *improves* the value of the objective function, compared to the original partition, because: (i) the convexification of m_k weakly decreases the variance of all messages, and (ii) the repartitioning of $m_j \cup m_{j+1}$ into m_j^{new} and m_{j+1}^{new} makes the two messages "more equal" in their measures compared to m_j and m_{j+1} in the original partition, and so decreases the weighted variance further.

After repartitioning, the constraints ICup(j^{new}) and ICup($(k-1)^{new}$) are perhaps violated. To fix this, we eliminate message m_{j+1}^{new} whose measure is exactly equal to $\frac{2\Phi_k}{|m_k|}$, as follows: we "shift" to the left messages $m_{j+2}^{new}, \ldots, m_{k-1}^{new}$ by length $\frac{2\Phi_k}{|m_k|}$, and add $\frac{2\Phi_k}{|m_k|}$ to message m_k^{new} from the left, as illustrated in Fig. 7(c).[36]

After this modification, all the ICup constraints are satisfied:

1. The constraint ICup($(j-1)^{new}$) is satisfied because $\mu_j^{new} \ge \mu_j$, whereas $\mu_{j-1}^{new} = \mu_{j-1}$ and $\overline{m}_{j-1}^{new} = \overline{m}_{j-1}$. Thus, the fact that ICup($j-1$) is satisfied in the original partition implies that ICup($(j-1)^{new}$) is satisfied in the modified partition.
2. The constraint ICup(j^{new}) is satisfied. To see this note first that, by construction, $|m_j^{new}| < \left|m_{j+1}\right|$ and $|m_{j+2}| = |m_{j+2}^{new}|$. Next, notice that credibility of the original partition, and the fact that m_{j+1} and m_{j+2} are two convex and adjacent messages, imply, by Equation (9), that $\left|m_{j+1}\right| \le \left|m_{j+2}\right| - 4\left(b - \frac{\alpha}{2}\right)$. Therefore, $\left|m_j^{new}\right| \le \left|m_{j+2}^{new}\right| - 4\left(b - \frac{\alpha}{2}\right)$, which guarantees by Equation (9) that ICup(j^{new}) is satisfied.

[36] We say that a convex message (interval) m is shifted to the left by x if $\underline{m}^{new} := \underline{m} - x$ and $\overline{m}^{new} := \overline{m} - x$ where m^{new} denotes message m after the shift.

3. The constraint ICup$\big((k-1)^{new}\big)$ is satisfied. To see this, note that $\mu_k^{new} = \mu_k - \frac{\Phi_k}{|m_k|} - \frac{1}{2} \cdot \frac{2\Phi_k}{|m_k|}$ (the convexification of m_k to the left decreases μ_k by $\frac{\Phi_k}{|m_k|}$, and the expansion from the left further decreases the mean by $\frac{1}{2} \cdot \frac{2\Phi_k}{|m_k|}$). Furthermore, $\mu_{k-1}^{new} = \mu_{k-1} - \frac{2\Phi_k}{|m_k|}$ and $\overline{m}_{k-1}^{new} = \overline{m}_{k-1} - \frac{2\Phi_k}{|m_k|}$ due to the shift of messages to the left. Taken together, the last three observations imply that since ICup$(k-1)$ is satisfied in the original partition, then ICup$((k-1)^{new})$ is satisfied in the merged partition.
4. The constraint ICup(k^{new}) is satisfied. This is because the convexification of m_k to the left implies that $\mu_{k+1}^{new} \geq \mu_{k+1}$. Shifting the messages to the left implies that $\mu_k^{new} = \mu_k - \frac{\Phi_k}{|m_k|} - \frac{1}{2} \cdot \frac{2\Phi_k}{|m_k|}$ (as explained above) and $\overline{m}_k^{new} = \overline{m}_k - L$. Taken together, these observations and equation (11) imply that since ICup(k) was satisfied in the original partition, then ICup(k^{new}) is satisfied in the new partition.
5. All the other ICup constraints are unaffected by the shift.

The modification improves the value of the objective function compared to the original partition. To see this, recall first that the partition illustrated in Fig. 7(a), which is the outcome of convexifying message m_k to the left (performed in Part I of the algorithm), improves the value of the objective function relative to the original partition. Next, as explained above, the partition depicted in Fig. 7(b) improves on the partition depicted in Fig. 7(a). Finally, inspection of Fig. 7(c) reveals that it consists of messages with the same lengths as the partition depicted in Fig. 7(b), except for message m_k^{new} in Fig. 7(c), which can be viewed as a union between messages m_k^{new} and m_{j+1}^{new} in Fig. 7(b). It is useful to perform this merge in two steps: first, shift message m_{j+1}^{new} to the right so that it lies between messages m_{k-1}^{new} and m_k^{new} in Fig. 7(b); second, merge messages m_{j+1}^{new} and m_k^{new} as illustrated in Fig. 7(c). Because the measure of message m_{j+1}^{new} is exactly $\frac{2\Phi_k}{|m_k|}$, the argument used in Case I above can be applied here, where m_{j+1}^{new} takes the place of message m_{k-1} in the argument presented in Case I.

Step 2: Show that Part II of the algorithm improves the objective function's value further
Part I of the algorithm, followed by the modifications described above (according to Case I or Case II), produces a partition with convex messages on the interval $[0, l_k]$ that satisfies all the ICup constraints and improves upon the value of the objective function compared to the original partition. The next lemma asserts that executing Part II of the algorithm on this partition preserves all the ICup constraints and further improves the value of the objective function.

Lemma A.1. *Let P be a partition that satisfies all the $ICup$ constraints with $\hat{J}$ convex messages on the interval $[0, l_{\hat{J}}]$. Then, tightly packing the messages on the interval $[0, l_{\hat{J}}]$ in a maximal manner preserves all the $ICup$ constraints and improves the value of the objective function.*

Finally, to complete the proof of the proposition, notice that when $I(1)$ messages are maximally tightly packed on Ω then all the ICup constraints are binding (by definition). In this case, all the ICdown constraints are satisfied as well. To see this, fix j and notice that

$$\frac{\mu_{j-1} + \mu_j}{2} - \underline{m}_j = \frac{\mu_{j-1} + \mu_j}{2} - \overline{m}_{j-1} = b - \frac{\alpha}{2},$$

where the first equality is by definition and the second equality follows from the fact that the ICup$(j-1)$ constraint is binding. It follows that $\frac{\mu_{j-1}+\mu_j}{2} - \underline{m}_j \leq b + \frac{\alpha}{2}$. This completes the proof of the proposition. ■

A.4. Proof of Lemma A.1

Suppose that messages 1 through $\hat{J}$ are not tightly packed. It follows that the ICup(j) constraint is not binding for some message m_j, $j < \hat{J}$. Redefine messages m_j and m_{j+1} as $m_j^{new} = [\underline{m}_j, \overline{m}_j + \varepsilon]$ and $m_{j+1}^{new} = [\underline{m}_{j+1} + \varepsilon, \overline{m}_{j+1}]$ for $\varepsilon > 0$ sufficiently small so that the ICup(j) constraint is satisfied for the new messages. Observe that this change weakly relaxes all the other ICup constraints. This change improves the value of the objective function (3) because it makes the probability masses of messages m_j^{new} and m_{j+1}^{new} closer together relative to messages m_j and m_{j+1}, which decreases the weighted variance. This implies that tightly packing the $\hat{J}$ messages on the interval $[0, l_{\hat{J}}]$ satisfies all the ICup constraints (by definition) and improves the value of the objective function.

If messages 1 through $\hat{J}$ are tightly packed, but not maximally tightly packed, then maximally tightly packing messages on the interval $[0, l_{\hat{J}}]$ satisfy all the ICup constraints and improve the value of the objective function.

To see this, suppose that P is a partition that satisfies all the ICup constraints with k messages that are tightly packed on the interval $[0, l]$, where $l \equiv l_k$. Suppose that it is possible to tightly pack $k + 1$ messages on the interval $[0, l]$. Let Q be a partition that tightly packs $k + 1$ messages on the interval $[0, l]$ and coincides with P on $\Omega \setminus [0, l]$. Denote the messages of P and Q restricted to the interval $[0, l]$ by $m_1^P, \dots, m_k^P$ and $m_1^Q, \dots, m_{k+1}^Q$, respectively. Denote the value of the objective function (3) restricted to the interval $[0, l]$ that is induced by these two partitions by $V(P) = \sum_{i=1}^{k} |m_i^P| \mathrm{Var}(m_i^P)$ and $V(Q) = \sum_{i=1}^{k+1} |m_i^Q| \mathrm{Var}(m_i^Q)$, respectively.

All the ICup(i), $1 \le i \le k$, constraints are binding in both P and Q. By equation (9), which in this case holds as an equality, it follows that $|m_i^P| > |m_{i+1}^Q| > |m_1^Q|$ for all $1 \le i \le k$. It follows that

$$
\begin{aligned}
V(P) &= \sum_{i=1}^{k} |m_{i+1}^Q| \mathrm{Var}(m_i^P) + \sum_{i=1}^{k} \left(|m_i^P| - |m_{i+1}^Q| \right) \mathrm{Var}(m_i^P) \\
&\ge \sum_{i=1}^{k} |m_{i+1}^Q| \mathrm{Var}(m_{i+1}^Q) + \sum_{i=1}^{k} \left(|m_i^P| - |m_{i+1}^Q| \right) \mathrm{Var}(m_1^Q) \\
&= \sum_{i=2}^{k+1} |m_i^Q| \mathrm{Var}(m_i^Q) + \left(l - \sum_{i=2}^{k+1} |m_i^Q| \right) \mathrm{Var}(m_1^Q) \\
&= V(Q)
\end{aligned}
$$

where the inequality follows from the fact that the variance of an interval increases in its length.

Finally, the fact that the ICup(k) constraint is satisfied in partition P and the fact that $\mu_{k+1}^Q > \mu_k^P$ imply that the ICup$(k + 1)$ constraint is satisfied in partition Q. Hence, partition Q satisfies all the ICup constraints.

A.5. Proof of Proposition 3

Fix a belief p^* and a cost parameter $\alpha^* \ge 0$. We show that for any α close to α^*, $V(p^*, \alpha)$ is close to $V(p^*, \alpha^*)$.

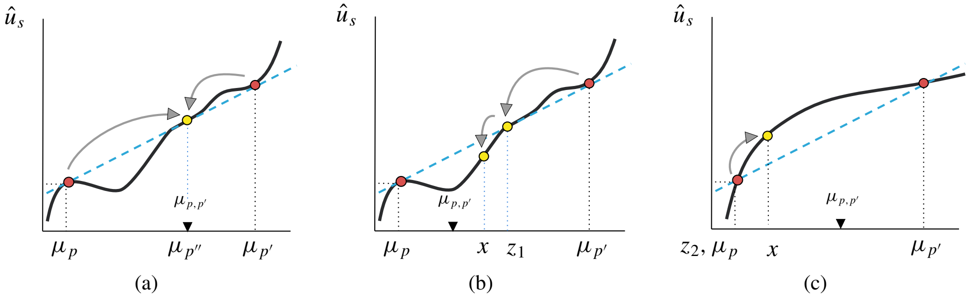


Fig. 8. Modifications of the induced posterior beliefs.

Denote the set of posterior beliefs induced by a Sender-optimal equilibrium (under the prior belief p^* and the cost parameter α^*) by P and denote the induced distribution over P by τ. For any two posterior beliefs $p, p' \in P$, denote the weighted mean of p and p' by

$$\mu_{p,p'} \equiv \frac{\tau(p)}{\tau(p)+\tau(p')} \cdot \mu_p + \frac{\tau(p')}{\tau(p)+\tau(p')} \cdot \mu_{p'}.$$

Define

$$g(\mu_p, \mu_{p'}) \equiv \frac{\hat{u}_S(\mu_{p'}) - \hat{u}_S(\mu_p)}{\mu_{p'} - \mu_p}.$$

In the definition of $g(\mu_p, \mu_{p'})$, as well as below, whenever a specific equilibrium is considered, $\hat{u}_S(\mu_p)$ is Sender's payoff under posterior belief p, under this equilibrium. The value of $g(\mu_p, \mu_{p'})$ can be interpreted as the slope of the line that connects the point $(\mu_p, \hat{u}_S(\mu_p))$ with the point $(\mu_{p'}, \hat{u}_S(\mu_{p'}))$ on the mean/payoff plane. In Fig. 8(a) this is the slope of the dashed line. Notice that, given three posterior beliefs p, p', and p'' that are such that $\mu_p < \mu_{p'} < \mu_{p''}$, if $g(\mu_p, \mu_{p'}) = g(\mu_{p'}, \mu_{p''}) = \alpha^*$ then $g(\mu_p, \mu_{p''}) = \alpha^*$. In this case, the three points $(\mu_p, \hat{u}_S(\mu_p))$, $(\mu_{p'}, \hat{u}_S(\mu_{p'}))$, and $(\mu_{p''}, \hat{u}_S(\mu_{p''}))$ are all on the same line in the mean/payoff plane.

Credibility of the optimal message function implies that $|g(\mu_p, \mu_{p'})| \leq \alpha^*$ for any pair of posterior beliefs $p, p' \in P$. If the inequality is strict for all such pairs (i.e., the credibility constraint is not binding), then clearly the same value of V can be achieved by employing the same distribution of posteriors τ over the set of posterior beliefs P for any α that is sufficiently close to α^*.

We therefore assume that there is at least one pair of posterior beliefs $p, p' \in P$ for which $g(\mu_p, \mu_{p'}) = \alpha^*$ (the case of $-g(\mu_p, \mu_{p'}) = \alpha^*$ is analogous and is therefore omitted). The next two lemmas are useful for the analysis that follows.

Lemma A.2. *Let $p, p' \in P$ be such that $\mu_p < \mu_{p'}$. For any two posterior means μ_x, μ_y such that $\mu_p \leq \mu_x < \mu_{p,p'} < \mu_y \leq \mu_{p'}$ there exists a set of posterior beliefs $\hat{P} = P \setminus \{p, p'\} \cup \{x, y\}$, where x and y are posterior beliefs that induce the means μ_x and μ_y, respectively, and a Bayes' plausible distribution $\hat{\tau}$ over $\hat{P}$ such that*

$$\hat{\tau}(x) = (\tau(p) + \tau(p')) \cdot \frac{\mu_y - \mu_{p,p'}}{\mu_y - \mu_x}$$

$$\hat{\tau}(y) = (\tau(p) + \tau(p')) \cdot \frac{\mu_{p,p'} - \mu_x}{\mu_y - \mu_x}$$

and $\hat{\tau} = \tau$ otherwise. We refer to the substitution of p, p' by x, y as the replacement of p and p' by x and y. *Furthermore, if $g(\mu_p, \mu_x) = g(\mu_x, \mu_y) = g(\mu_y, \mu_{p'})$, then the value of V induced by τ is the same as the value of V induced by $\hat{\tau}$.*[37]

Lemma A.3. *Suppose that $p, p', p'' \in P$ are three posterior beliefs with means $\mu_p < \mu_{p'} < \mu_{p''}$ such that $g(\mu_p, \mu_{p'}) = g(\mu_{p'}, \mu_{p''}) = \alpha^*$. Then, it is possible to eliminate either p, or p'', or both, from P, and adjust the distribution over posteriors τ, in a way that preserves the value of V and preserves credibility.*

Fix a pair of posterior beliefs $p, p' \in P$ for which $g(\mu_p, \mu_{p'}) = \alpha^*$. By Lemma A.3, no loss of generality is implied by assuming that $g(\mu_p, \mu_y) < \alpha^*$ for all $y \in P \setminus \{p, p'\}$ (as otherwise at least one posterior belief can be eliminated from P). Credibility then implies that $\mu_y \notin (\mu_p, \mu_{p'})$ for all $y \in P \setminus \{p, p'\}$.[38]

For ease of exposition, we start with the case in which $\hat{u}_S$ is everywhere single-valued. After presenting the analysis for this case, we explain how it can be extended to the general case.

Suppose that $\hat{u}_S$ is single-valued. We distinguish between three cases:

(i) Suppose that $g(\mu_p, \mu_{p,p'}) = \alpha^*$. In this case, the point $(\mu_{p,p'}, \hat{u}_S(\mu_{p,p'}))$ lies on the line that connects the points $(\mu_p, \hat{u}_S(\mu_p))$ and $(\mu_{p'}, \hat{u}_S(\mu_{p'}))$ in the mean/payoff plane, as illustrated in Fig. 8(a).

Modify the message function such that in any state in which the messages that induce p and p' were sent, the message function now sends only one message. Denote the posterior belief induced by this new message by p'' and notice that the mean of p'' is $\mu_{p''} = \mu_{p,p'}$. The fact that $g(\mu_p, \mu_{p,p'}) = g(\mu_{p,p'}, \mu_{p'}) = \alpha^*$ implies that the value of V is unaffected by the modification (see also the proof of Lemma A.3).

After the modification, we have that $|g(\mu_{p''}, \mu_y)| < \alpha^*$ for all $y \in P \setminus \{p, p'\}$. Intuitively, this is because for any $\mu_y \notin (\mu_p, \mu_{p'})$, the slope of the line that connects the point $(\mu_y, \hat{u}_S(\mu_y))$ with the point $(\mu_{p''}, \hat{u}_S(\mu_{p''}))$ in the mean/payoff plane is between the slopes of: (A) the line that connects $(\mu_y, \hat{u}_S(\mu_y))$ with $(\mu_p, \hat{u}_S(\mu_p))$ and (B) the line that connects $(\mu_y, \hat{u}_S(\mu_y))$ with $(\mu_{p'}, \hat{u}_S(\mu_{p'}))$. Since, by credibility, both (A) and (B) are smaller than α^* in absolute value, the result follows.

Formally, fix any posterior belief $y \in P \setminus \{p, p'\}$. Since $g(\mu_p, \mu_{p''}) = \alpha^*$, we have that

$$g\left(\mu_y, \mu_{p''}\right) = \frac{u\left(\mu_{p''}\right) - u\left(\mu_y\right)}{\mu_{p''} - \mu_y} = \frac{u\left(\mu_p\right) + \alpha^*\left(\mu_{p''} - \mu_p\right) - u\left(\mu_y\right)}{\mu_{p''} - \mu_y}.$$

Differentiating g with respect to $\mu_{p''}$ yields:

$$\frac{\partial g\left(\mu_y, \mu_{p''}\right)}{\partial \mu_{p''}} = \left(\mu_p - \mu_y\right)\frac{\alpha^* - g\left(\mu_y, \mu_p\right)}{\left(\mu_{p''} - \mu_y\right)^2}. \tag{21}$$

[37] If either $\hat{u}_S(\mu_x)$ or $\hat{u}_S(\mu_y)$ is not single-valued, then fix values $\hat{u}_S(\mu_x)$ and $\hat{u}_S(\mu_y)$, respectively. If the condition $g(\mu_p, \mu_x) = g(\mu_x, \mu_y) = g(\mu_y, \mu_{p'})$ is satisfied for these values, then the conclusion follows.

[38] To see this, suppose by way of contradiction that $\mu_y \in (\mu_p, \mu_{p'})$ and that $\hat{u}_S(\mu_{p'}) > \hat{u}_S\left(\mu_p\right)$ (the other case is handled similarly). Since $g\left(\mu_p, \mu_y\right) < \alpha^*$ then $\hat{u}_S(\mu_{p'}) - \hat{u}_S(\mu_y) > \hat{u}_S(\mu_{p'}) - \hat{u}_S\left(\mu_p\right) - (\mu_y - \mu_p) \cdot \alpha^*$. Using the fact that $g(\mu_p, \mu_{p'}) = \alpha^*$ we have that $\hat{u}_S(\mu_{p'}) - \hat{u}_S(\mu_y) > (\mu_{p'} - \mu_y)\alpha^*$, and since $\mu_y \in (\mu_p, \mu_{p'})$ we have that $g\left(\mu_y, \mu_{p'}\right) > \alpha^*$, contradicting the credibility of the message function.

Recall that $g\left(\mu_y, \mu_p\right) < \alpha^*$ for all $y \in P \setminus \{p, p'\}$. Thus, if $\mu_y < \mu_p$, the right-hand side of Equation (21) is positive and so $g\left(\mu_y, \mu_p\right) < g\left(\mu_y, \mu_{p''}\right) < g(\mu_y, \mu_{p'})$. And, if $\mu_y > \mu_{p'}$, the right-hand side of Equation (21) is negative and so $g\left(\mu_y, \mu_p\right) > g\left(\mu_y, \mu_{p''}\right) > g(\mu_y, \mu_{p'})$. It follows that $|g\left(\mu_y, \mu_{p''}\right)| < \max[|g\left(\mu_y, \mu_p\right)|, |g(\mu_y, \mu_{p'})|] \leq \alpha^*$ for all $\mu_y \notin [\mu_p, \mu_{p'}]$.

Thus, the modified message function eliminates a pair of posterior beliefs for which the credibility constraint was binding, and replaces it with one posterior belief for which credibility is not binding for any other element in P.

(ii) Suppose that $g(\mu_p, \mu_{p,p'}) < \alpha^*$. In this case, the point $(\mu_{p,p'}, \hat{u}_S(\mu_{p,p'}))$ is *below* the line that connects the points $(\mu_p, \hat{u}(\mu_p))$ and $(\mu_{p'}, \hat{u}(\mu_{p'}))$ in the mean/payoff plane, as illustrated in Fig. 8b).

Let $z_1 \in [\mu_p, \mu_{p'}]$ be the lowest mean that is greater than $\mu_{p,p'}$ for which $g(\mu_p, z_1) = \alpha^*$, i.e., $z_1 = \min[x | x > \mu_{p,p'}$ and $g(\mu_p, x) = \alpha^*]$. Note that z_1 necessarily exists, by the continuity of g and the intermediate value theorem (it could be the case that $z_1 = \mu_{p'}$).

Replace the posteriors p and p' by the posteriors p and p'', where p'' is a posterior with mean z_1, in the manner described in Lemma A.2 and illustrated in Fig. 8(b). This modification does not change the value of the function V because $g(\mu_p, \mu_{p'}) = g(\mu_p, z_1)$. Credibility of the original message function, and the fact that $z_1 \in (\mu_p, \mu_{p'})$, imply that $|g(\mu_y, \mu_{p''})| < \alpha^*$ for all $y \in P \setminus \{p\}$ (the analysis is identical to the one presented in case (i) above). Therefore the modified message function satisfies credibility.

Continuity of g implies that, for any $\varepsilon > 0$, there exists a $\hat{\delta} > 0$, such that if $0 < \delta < \hat{\delta}$ then there exists $x \in [z_1 - \varepsilon, z_1]$ such that $|g(\mu_y, x)| \leq \alpha^* - \delta$ for all $y \in P \setminus \{p'\}$. Thus, when α is close to α^*, we can modify the message function (by replacing the posterior beliefs p and p'' by p and p''', where p''' is a posterior belief with mean x, in the manner described in Lemma A.2 and illustrated in Fig. 8b), such that credibility is satisfied and the value of V is only slightly affected.[39]

(iii) Suppose that $g(\mu_p, \mu_{p,p'}) > \alpha^*$. In this case, the point $(\mu_{p,p'}, \hat{u}_S(\mu_{p,p'}))$ is *above* the line that connects the points $(\mu_p, \hat{u}(\mu_p))$ and $(\mu_{p'}, \hat{u}(\mu_{p'}))$ in the mean/payoff plane, as illustrated in Fig. 8c).

Let $z_2 \in [\mu_p, \mu_{p'}]$ be the highest value that is smaller than $\mu_{p,p'}$ for which $g(\mu_p, z_2) = \alpha^*$, i.e. $z_2 = \max[x | x < \mu_{p,p'}$ and $g(\mu_p, x) = \alpha^*]$. For ease of notation we define $g(\mu_p, \mu_p) = \alpha^*$ and allow z_2 to be equal to μ_p, which is the case that is illustrated in Fig. 8(c). As in case (ii), z_2 necessarily exists by the continuity of g.

If $z_2 \neq \mu_p$, replace the posteriors p and p' by the posteriors p'' and p', where p'' is a posterior with mean z_2, in the manner described in Lemma A.2. As in case (ii) above, this modification preserves the value of V and the credibility of the message function.

Continuity of g implies that, for any $\varepsilon > 0$, there exists a $\hat{\delta} > 0$, such that if $0 < \delta < \hat{\delta}$ then there exists $x \in [z_2, z_2 + \varepsilon]$ such that $|g(\mu_y, x)| \leq \alpha^* - \delta$ for all $y \in P \setminus \{p'\}$. As in case (ii) above, when α is close to α^*, we can modify the message function (by replacing the posterior beliefs p' and p'' by p' and p''', where p''' is a posterior belief with mean x, in the manner described in Lemma A.2 and illustrated in Fig. 8c), such that credibility is satisfied and the value of V is only slightly affected.

[39] This is because $\hat{u}_S$ is a continuous function and because the distribution of posterior beliefs, $\hat{\tau}$, that is described in the statement of Lemma A.2, is only slightly affected by the modification.

Thus, for any pair of posterior beliefs $p, p' \in P$ for which credibility is binding in the original message function, and for any small change in α^*, it is either the case that this pair can be eliminated without affecting the value of V (case i), or there exists a modification of the message function that restores credibility while only slightly affecting the value of V (cases (ii) and (iii)). Therefore, if α is close to α^* then the value $V(p^*, \alpha)$ is close to $V(p^*, \alpha^*)$.

To complete the proof, suppose now that $\hat{u}_S$ is not everywhere single-valued. The analysis is similar to the analysis above up to a few technical adaptations described below.

Denote $\underline{g}(\mu_p, x) = g(\mu_p, \min\{\hat{u}_S(x)\})$ and $\overline{g}(\mu_p, x) = g(\mu_p, \max\{\hat{u}_S(x)\})$. Distinguish between the following three cases:

(i) Suppose that $\underline{g}(\mu_p, \mu_{p,p'}) < \alpha^* < \overline{g}(\mu_p, \mu_{p,p'})$. In this case, perform the change described in case (i) above, with the additional assumption that if Receiver's beliefs correspond to $\mu_{p,p'}$ then he randomizes in such a way that $g(\mu_p, \mu_{p,p'}) = \alpha^*$. The rest of the argument carries over as in case (i).

(ii) Suppose that $\overline{g}(\mu_p, \mu_{p,p'}) < \alpha^*$. There exists a point $z_1 \in [\mu_{p,p'}, \mu_{p'}]$ such that $\underline{g}(\mu_p, z_1) \leq \alpha^* \leq \overline{g}(\mu_p, z_1)$. Pick a payoff for Sender from $\hat{u}_S(z_1)$ such that $g(\mu_p, z_1) = \alpha^*$ where g is computed according to this payoff at z_1. The argument proceeds in the same way as in case (ii) except that we allow ε to be zero so that instead of a new posterior belief p''' we may adjust the value of $\hat{u}_S(z_1)$ downwards.

(iii) Suppose that $\underline{g}(\mu_p, \mu_{p,p'}) > \alpha^*$. There exists a point $z_2 \in [\mu_p, \mu_{p,p'}]$ such that $\underline{g}(\mu_p, z_2) \leq \alpha^* \leq \overline{g}(\mu_p, z_2)$. Pick a payoff for Sender from $\hat{u}_S(z_2)$ such that $g(\mu_p, z_2) = \alpha^*$ where g is computed according to this payoff at z_2. The argument proceeds in the same way as in case (iii) except that we allow ε to be zero so that instead of a new posterior belief p''' we adjust the value of $\hat{u}_S(z_2)$ upwards.

A.6. Proof of Lemma A.2

Observe that this replacement of posteriors is performed in a way that contracts the distribution of posterior means and preserves both the conditional mean $\mu_{p,p'}$ and the mean μ_{p^*}. This implies that $\hat{\tau}$ second-order-stochastically-dominates τ. This ensures that the distribution $\hat{\tau}$ is Bayes plausible.

Suppose now that $g(\mu_p, \mu_x) = g(\mu_x, \mu_y) = g(\mu_y, \mu_{p'})$. Sender's value (V) from employing the modified message function is given by:

$$\sum_{q \in P \backslash \{p,p'\} \cup \{x,y\}} \hat{\tau}(q) \cdot \hat{u}_S(q) = \sum_{q \in P \backslash \{p,p'\}} \hat{\tau}(q) \hat{u}_S(q) + \hat{\tau}(x) \cdot \hat{u}_S(\mu_x) + \hat{\tau}(y) \cdot \hat{u}_S(\mu_y). \tag{22}$$

By construction, we have that $\hat{\tau}(x) = \frac{\mu_y - \mu_p}{\mu_y - \mu_x} \cdot \tau(p) - \frac{\mu_{p'} - \mu_y}{\mu_y - \mu_x} \cdot \tau(p')$. Since $g(\mu_p, \mu_x) = g(\mu_x, \mu_y) = g(\mu_y, \mu_{p'})$, it follows that

$$\hat{\tau}(x) = \frac{\hat{u}_S(\mu_y) - \hat{u}_S(\mu_p)}{\hat{u}_S(\mu_y) - \hat{u}_S(\mu_x)} \cdot \tau(p) - \frac{\hat{u}_S(\mu_{p'}) - \hat{u}_S(\mu_y)}{\hat{u}_S(\mu_y) - \hat{u}_S(\mu_x)} \cdot \tau(p').$$

By plugging this expression of $\hat{\tau}(x)$, and $\hat{\tau}(y) = \tau(p) + \tau(p') - \hat{\tau}(x)$, into the right-hand side of Equation (22) we obtain

$$\sum_{q\in P\setminus\{p,p'\}} \hat{\tau}(q)\,\hat{u}_S(q) + \tau(p)\cdot \hat{u}_S\left(\mu_p\right) + \tau\left(p'\right)\cdot \hat{u}_S(\mu_{p'}) = \sum_{q\in P} \hat{\tau}(q)\,\hat{u}_S(q),$$

which is Sender's value under the original message function.

A.7. Proof of Lemma A.3

Suppose that a credible message function induces the three posterior beliefs p, p', p'' as described in the statement of the lemma.

If $\mu_{p,p''} = \mu_{p'}$, modify the message function so that in any state in which the messages that induced p and p'' are sent, the modified message function sends the message that induced p' instead. Thus, the mean of the posterior belief induced by this message remains $\mu_{p,p''}$. The fact that $g(\mu_p, \mu_{p'}) = g(\mu_{p'}, \mu_{p''})$ implies that the value of V remains unchanged.[40]

If $\mu_{p,p''} < \mu_{p'}$, replace the posteriors p and p'' in P by p and p', in the manner described in Lemma A.2. If $\mu_{p,p''} > \mu_{p'}$, replace the posteriors p and p'' in P by p' and p'' in the manner described in Lemma A.2. These modifications do not change the value of the function V.

Finally, note that in all the cases described above, the modified message function does not induce a posterior belief that was not induced by the original message function. Thus, the credibility constraints in Sender's problem (SP1) are only relaxed, and the fact that the original message function was credible implies that the modified one is also credible.

A.8. Proof of Proposition 4

We prove the proposition for the case in which $\hat{u}_S$ is increasing and convex. The proof for the case in which $\hat{u}_S$ is decreasing, or decreasing and then increasing, is analogous.

Suppose that the state space is binary, i.e., $\Omega = \{l, h\}$ for some two numbers $l, h \in \mathbb{R}$ with $l < h$. A belief over Ω can be described by the probability $p \in [0, 1]$ that the state is h. The prior belief is thus given by $\pi \in (0, 1)$. The mean of belief p is $\mu_p = l + (h - l)\,p$. In what follows we normalize the parameters h and l to be 1 and 0, respectively, and therefore $\mu_p = p$.

According to Corollary 2 the optimal message function induces either one posterior belief that is equal to the prior π, or two credible posterior beliefs $p_L < \pi < p_H$, whichever generates a higher expected payoff to Sender. In the former case, the ex-ante expected payoff to Sender is $\hat{u}_S(\pi)$. In the latter case, the ex-ante expected payoff to Sender is $p_H \cdot \hat{u}_S(p_H) + p_L \cdot \hat{u}_S(p_L)$. Credibility requires that $\frac{\hat{u}_S(p_H)-\hat{u}_S(p_L)}{p_H-p_L} \le \alpha$.

We distinguish between the following three cases:

(i) If $\hat{u}_S(1) - \hat{u}_S(0) \le \alpha$, then the message function that induces a distribution τ over the posterior beliefs $p_L^* = 0$ (realized with probability $1 - \pi$) and $p_H^* = 1$ (realized with probability π) is credible under α. Such a message function is optimal for Sender because it concavifies $\hat{u}_S$ on the interval [0, 1].

[40] To see this, note first that $\tau(p)\cdot\hat{u}_S(\mu_p) + \tau\left(p'\right)\cdot\hat{u}_S(\mu_{p'}) = \left(\tau(p)+\tau\left(p'\right)\right)\left(\frac{\tau(p)}{\tau(p)+\tau\left(p'\right)}\hat{u}_S(\mu_p) + \frac{\tau\left(p'\right)}{\tau(p)+\tau\left(p'\right)}\hat{u}_S(\mu_{p'})\right)$. Next, since $\mu_{p'} = \mu_{p,p''}$ and $g(\mu_p, \mu_{p'}) = \alpha$ we have that $\hat{u}_S(\mu_p) = \hat{u}_S(\mu_{p,p''}) - (\mu_{p,p''} - \mu_p)\alpha$, and since $g(\mu_{p'}, \mu_{p''}) = \alpha$ we have $\hat{u}_S(\mu_{p''}) = \hat{u}_S(\mu_{p,p''}) + \left(\mu_{p''} - \mu_{p,p''}\right)\alpha$. By definition of $\mu_{p,p''}$ we then obtain $\tau(p)\cdot\hat{u}_S(\mu_p) + \tau\left(p'\right)\cdot\hat{u}_S(\mu_{p'}) = (\tau(p) + \tau\left(p'\right))\cdot\hat{u}_S(\mu_{p,p''})$.

(ii) If $\frac{\hat{u}_S(\pi)-\hat{u}_S(0)}{\pi} < \alpha < \hat{u}_S(1) - \hat{u}_S(0)$, then the two optimally induced beliefs under α are $p_L^* = 0$ and p_H^* that is such that $\frac{\hat{u}_S(p_H^*)-\hat{u}_S(0)}{p_H^*} = \alpha$. To see this, note first that for any different pair of posterior beliefs $p_L < \pi < p_H$, decreasing p_L relaxes the credibility constraint and improves the ex-ante expected payoff to Sender. Then, it is possible to increase p_H up to p_H^*, where the credibility constraint is binding, i.e., $\frac{\hat{u}_S(p_H^*)-\hat{u}_S(0)}{p_H^*} = \alpha$, which further increases the ex-ante expected payoff to Sender.

(iii) If $\alpha \leq \frac{\hat{u}_S(\pi)-\hat{u}_S(0)}{\pi}$, then the unique feasible policy induces just one posterior belief, which is equal to the prior π. This is because the convexity of $\hat{u}_S$ implies that $\frac{\hat{u}_S(p_H)-\hat{u}_S(p_L)}{p_H-p_L}$ is increasing in p_L and in p_H and therefore $\frac{\hat{u}_S(p_H)-\hat{u}_S(p_L)}{p_H-p_L} \geq \alpha$ for any $p_L \leq \pi$ and $p_H \geq \pi$. Thus, no message function can induce two (Bayes plausible) posterior beliefs in a credible way.

Notice that decreasing the value of α does not affect Sender's optimal distribution over posteriors so long as α remains as in case (i) or (iii). As the value of α changes from case (i) to (ii), or as α decreases in case (ii), Sender's optimal distribution τ becomes more garbled. This is because the convexity of $\hat{u}$ implies that $\frac{\hat{u}_S(p_H)-\hat{u}_S(0)}{p_H}$ increases in p_H. Thus, a lower value of α implies a lower value of p_H^* (i.e., messages are less informative with respect to the state).

A.9. Proof of Proposition 5

We first show that for any equilibrium there exists another equilibrium that generates an ex-ante identical payoff to Sender and that employs no more than $|\Omega + 1|$ messages. Suppose that an equilibrium message function σ sends more than $|\Omega| + 1$ messages. Every message $m \in M$ that is sent by σ induces a posterior belief (distribution) p_m^σ over the states. This belief can be represented by a vector in $\mathbb{R}^{|\Omega|-1}$. Sender's equilibrium material payoff from inducing the posterior belief p_m^σ is $v_{p_m^\sigma}$. Thus, each message m that is sent by σ induces a vector $(p_m^\sigma, v_{p_m^\sigma}) \in \mathbb{R}^{|\Omega|}$.

Denote Sender's ex-ante expected payoff under σ by $U_S(\sigma)$. Then,

$$\left(\mathbb{E}_m\left[p_m^\sigma\right], \mathbb{E}_m\left[v_{p_m^\sigma}\right]\right) = (\pi, U_S(\sigma)) \in \mathbb{R}^{|\Omega|}$$

where $\mathbb{E}_m\left[p_m^\sigma\right] = \pi \in \mathbb{R}^{|\Omega|-1}$ follows from Bayes plausibility: the mean of the induced posterior beliefs is equal to the prior belief, and $\mathbb{E}_m\left[v_{p_m^\sigma}\right] = U_S(\sigma) \in \mathbb{R}$ by definition of $U_S(\sigma)$. Therefore, the vector $(\pi, U_S(\sigma)) \in \mathbb{R}^{|\Omega|}$ belongs to the convex hull that is generated by the set $\{(p_m^\sigma, v_{p_m^\sigma})\}_{m \in M}$.

By Carathéodory's theorem (Rockafellar 1997, Theorem 17.1) it is possible to write the vector $\{(\pi, U_S(\sigma)\}$ as a convex combination of no more than $|\Omega| + 1$ elements in the set $\{(p_m^\sigma, v_{p_m^\sigma})\}_{m \in M}$.

Suppose that the messages that induce these $|\Omega| + 1$ beliefs in the original message function σ are given by $m_1, \ldots, m_{|\Omega|+1}$. Consider a message function σ' that sends messages $m'_1, \ldots, m'_{|\Omega|+1}$ that induce the same posterior beliefs as those induced by $m_1, \ldots, m_{|\Omega|+1}$, with the probabilities determined by Carathéodory's theorem. By construction, $p_{m'_j}^{\sigma'} = p_{m_j}^\sigma$ for $j \in \{1, \cdots, |\Omega| + 1\}$. Note that the message function σ' generates the same ex-ante expected payoff to Sender as σ.

We now show that the message function σ' can also be part of an equilibrium. Denote by $q^{\sigma}(m,\omega)$ the probability that message m is sent in state ω by message function σ. Observe that:

$$q^{\sigma}(m_j,\omega)=0 \Rightarrow q^{\sigma'}(m'_j,\omega)=0 \quad \forall j\in\{1,\cdots,|\Omega|+1\}, \forall\omega\in\Omega,$$

because, otherwise, $q^{\sigma'}(m'_j,\omega)>0=q^{\sigma}(m_j,\omega)$ for some $j\in\{1,\cdots,|\Omega|+1\}$ and $\omega\in\Omega$. Then, $p^{\sigma'}_{m'_j}[\omega]>0$ while $p^{\sigma}_{m_j}[\omega]=0$. This is in contradiction to the fact that $p^{\sigma'}_{m'_j}=p^{\sigma}_{m_j}$ for $j\in\{1,\cdots,|\Omega|+1\}$. Thus, every belief that is induced by σ' in some state ω was also induced by σ in ω. Therefore, the credibility of σ implies the credibility of σ'.

Thus, we have shown that we may restrict our attention to equilibria that employ no more than $|\Omega|+1$ messages.

Consider an equilibrium message function σ that employs $|\Omega|+1$ messages, that induce posterior beliefs $p_1,\ldots,p_{|\Omega|+1}$ with probabilities $\lambda_1,\ldots,\lambda_{|\Omega|+1}$, respectively, such that $\sum_{i=1}^{|\Omega|+1}\lambda_i p_i=\pi$. Denote the set of these posterior beliefs by $P=\{p_1,\ldots,p_{|\Omega|+1}\}$ and denote the ex-ante expected payoff to Sender that is generated by σ by $\sum_{i=1}^{|\Omega|+1}\lambda_i\cdot v_{p_i}\equiv U$. We may assume that each λ_i is positive and that each p_i is different from π because otherwise it is possible to induce an ex-ante expected payoff that is at least U with no more than $|\Omega|$ messages.

We proceed with the following lemma.

Lemma A.4. *Suppose that $S=\{x_1,\ldots,x_{d+2}\}$ is a set of $d+2$ vectors in $\mathbb{R}^d$. For any vector $p\in\mathbb{R}^d$ in the convex hull generated by S, denoted by* co(S)*, there exist at least two distinct subsets $S',S''\subset S$ with no more than $d+1$ elements each, such that $p\in$ co$(S')\cap$ co(S'').*

Proof. For any vector $x\in\mathbb{R}^d$, denote the vector's i^{th} coordinate by $x_{[i]}$, and set $\bar{x}\equiv\binom{1}{x}\in\mathbb{R}^{d+1}$. Define the matrices $X=\begin{bmatrix}x_1 & x_2 & \cdots & x_{d+2}\end{bmatrix}\in\mathbb{R}^{d\times(d+2)}$ and $\overline{X}=\begin{bmatrix}\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_{d+2}\end{bmatrix}\in\mathbb{R}^{(d+1)\times(d+2)}$. Since $p\in$ co(S), there exists a vector $\lambda=(\lambda_{[1]},\ldots,\lambda_{[d+2]})^T\in\mathbb{R}^{d+2}$ such that $\sum_{i=1}^{d+2}\lambda_{[i]}=1$ and $X\lambda=p$.

The vectors $\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_{d+2}$ are linearly dependent. Hence, there is a vector $\alpha=(\alpha_{[1]},\ldots,\alpha_{[d+2]})^T\in\mathbb{R}^{d+2}$, with coordinates not all equal to zero, such that $\alpha\in\ker\left(\overline{X}\right)$. Since $\sum_{i=1}^{d+2}\alpha_{[i]}=0$ it follows that α has at least one positive coordinate and at least one negative coordinate.

Suppose without loss of generality that the coordinates in α are ordered such that $\frac{\lambda_{[1]}}{\alpha_{[1]}}\le\cdots\le\frac{\lambda_{[k]}}{\alpha_{[k]}}<0<\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\le\cdots\le\frac{\lambda_{[d+2]}}{\alpha_{[d+2]}}$. We can therefore decompose the vector p as follows:

$$p=\sum_{i=1}^{d+2}\lambda_{[i]}\bar{x}_i=\sum_{i=1}^{k}\lambda_{[i]}\bar{x}_i+\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\sum_{i=k+1}^{d+2}\alpha_{[i]}\bar{x}_i+\sum_{i=k+2}^{d+2}\left(\frac{\lambda_{[i]}}{\alpha_{[i]}}-\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\right)\alpha_{[i]}\bar{x}_i.$$

Substituting $\sum_{i=k+1}^{d+2}\alpha_{[i]}\bar{x}_i=-\sum_{i=1}^{k}\alpha_{[i]}\bar{x}_i$ and rearranging yields

$$p=\sum_{i=1}^{k}\left(\frac{\lambda_{[i]}}{\alpha_{[i]}}-\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\right)\alpha_{[i]}\bar{x}_i+\sum_{i=k+2}^{d+2}\left(\frac{\lambda_{[i]}}{\alpha_{[i]}}-\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\right)\alpha_{[i]}\bar{x}_i.$$

Therefore, the vector $\beta=(\beta_{[1]},\ldots,\beta_{[d+2]})^T$ that is defined such that $\beta_{[i]}=\left(\frac{\lambda_{[i]}}{\alpha_{[i]}}-\frac{\lambda_{[k+1]}}{\alpha_{[k+1]}}\right)\alpha_{[i]}$ satisfies $\sum_{i=1}^{d+2}\beta_{[i]}=1$ and $X\beta=p$ and all its coordinates are nonnegative. A similar argument shows that the vector $\gamma = (\gamma_{[1]}, \dots, \gamma_{[d+2]})^T$ that is defined such that $\gamma_{[i]} = \left(\frac{\lambda_{[i]}}{\alpha_{[i]}} - \frac{\lambda_{[k]}}{\alpha_{[k]}}\right)\alpha_{[i]}$ satisfies $\sum_{i=1}^{d+2} \gamma_{[i]} = 1$ and $X\gamma = p$ and all its coordinates are nonnegative.

Let $S' = \{x_1, \dots, x_k, x_{k+2}, \dots, x_{d+2}\}$ and $S'' = \{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_{d+2}\}$. We have therefore shown that $p \in \text{co}(S')$ and $p \in \text{co}(S'')$. Moreover, notice that $\lambda_{[i]} = \nu\beta_{[i]} + (1-\nu)\,\gamma_{[i]}$ where $\nu = \frac{1}{1-\frac{\lambda_{k+1}}{\alpha_{k+1}}\frac{\alpha_k}{\lambda_k}}$. ■

By Lemma A.4, given any set of beliefs $P = \{p_1, \dots, p_{|\Omega|+1}\}$ that are each different from π, and are induced with positive probabilities $\lambda_1, \dots, \lambda_{|\Omega|+1}$, respectively, such that $\sum_{i=1}^{|\Omega|+1} \lambda_i p_i = \pi$, there exist at least two subsets of beliefs P', $P'' \subset P$ with no more than $|\Omega|$ elements each, with associated probabilities λ' and λ'' that also average the prior belief π. With slight abuse of notation we also use λ' and λ'' to denote the $|\Omega| + 1$-dimensional vectors of probabilities $p_1, \dots, p_{|\Omega|+1}$ where instead of the probability associated with the belief that is missing from the subsets P' and P'', respectively, we write zero.

Inspection of the proof of Lemma A.4 reveals that the vector λ can be written as a convex combination of the vectors λ' and λ''. Therefore, the expected payoff $U = \sum_{i=1}^{|\Omega|+1} \lambda_i \cdot v_{p_i}$ can be written as a convex combination of the expected payoffs $U' = \sum_{i=1}^{|\Omega|+1} \lambda'_i \cdot v_{p_i}$ and $U'' = \sum_{i=1}^{|\Omega|+1} \lambda''_i \cdot v_{p_i}$ associated with the two vectors of probabilities λ' and λ''. It follows that either U' or U'' is greater than or equal to U.

Finally, the message functions σ' and σ'' that induce the posterior beliefs in P' and P'', respectively, satisfy credibility because of the same argument used in the first part of the proof. Namely,

$$q^{\sigma}(m_j, \omega) = 0 \Rightarrow q^{\sigma'}(m'_j, \omega) = 0,\, p^{\sigma''}(m'_j, \omega) = 0 \quad \forall j \in \{1, \cdots, |\Omega| + 1\},$$

because, otherwise, $q^{\sigma'}(m'_j, \omega), q^{\sigma''}(m'_j, \omega) = 0 > 0 = q^{\sigma}(m_j, \omega)$. Thus, it is never the case that a message is sent under σ' at a state where it was not sent under σ. Therefore, the credibility of σ implies the credibility of σ' and σ''.

References

Abeler, Johannes, Nosenzo, Daniele, Raymond, Collin, 2019. Preferences for truth-telling. Econometrica 87 (4), 1115–1153.

Aumann, Robert J., Maschler, Michael, 1995. Repeated Games with Incomplete Information. MIT Press, Cambridge, MA.

Austin, John L., 1975. How to do Things with Words. Harvard University Press.

Battigalli, Pierpaolo, Dufwenberg, Martin, 2007. Guilt in games. Am. Econ. Rev. 97 (2), 170–176.

Battigalli, Pierpaolo, Dufwenberg, Martin, 2009. Dynamic psychological games. J. Econ. Theory 144 (1), 1–35.

Battigalli, Pierpaolo, Dufwenberg, Martin, 2022. Belief-dependent motivations and psychological game theory. J. Econ. Lit. 60 (3), 833–882.

Bester, Helmut, Strausz, Roland, 2001. Contracting with imperfect commitment and the revelation principle: the single agent case. Econometrica 69 (4), 1077–1098.

Caplin, Andrew, Leahy, John, 2004. The supply of information by a concerned expert. Econ. J. 114 (497), 487–505.

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

Crawford, Vincent P., Sobel, Joel, 1982. Strategic information transmission. Econometrica 50 (6), 1431–1451.

Doval, Laura, Skreta, Vasiliki, 2018. Constrained Information Design: Toolkit. Working Paper.

Eilat, Ran, Eliaz, Kfir, Mu, Xiaosheng, 2021. Bayesian privacy. Theor. Econ. 16, 1557–1603.

Ely, Jeffrey, Frankel, Alexander, Kamenica, Emir, 2015. Suspense and surprise. J. Polit. Econ. 123 (1), 215–260.

Fischbacher, Urs, Föllmi-Heusi, Franziska, 2008. Lies in disguise: an experimental study on cheating. J. Eur. Econ. Assoc. 11, 525–547.

Fréchette, Guillaume R., Lizzeri, Alessandro, Perego, Jacopo, 2022. Rules and commitment in communication: an experimental analysis. Econometrica 90 (5), 2283–2318.

Geanakoplos, John, Pearce, David, Stacchetti, Ennio, 1989. Psychological games and sequential rationality. Games Econ. Behav. 1 (1), 60–79.

Gneezy, Uri, 2005. Deception: the role of consequences. Am. Econ. Rev. 95 (1), 384–394.

Gradwohl, Ronen, 2018. Privacy in implementation. Soc. Choice Welf. 50, 547–580.

Guo, Yingni, Shmaya, Eran, 2021. Costly miscalibration. Theor. Econ. 16 (2), 477–506.

Hagenbach, Jeanne, Koessler, Frédéric, 2022. Selective memory of a psychological agent. Eur. Econ. Rev. 142.

Heller, Yuval, Sturrock, David, 2020. Promises and endogenous reneging costs. J. Econ. Theory 187, 105024.

Heumann, Tibor, 2020. On the cardinality of the message space in sender–receiver games. J. Math. Econ. 90, 109–118.

Kamenica, Emir, Gentzkow, Matthew, 2011. Bayesian persuasion. Am. Econ. Rev. 101 (6), 2590–2615.

Kartik, Navin, 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76 (4), 1359–1395.

Krähmer, Daniel, Strausz, Roland, forthcoming. Optimal non-linear pricing with data-sensitive consumers, Am. Econ. J. Microecon.

Le Treust, Maëlle, Tomala, Tristan, 2019. Persuasion with limited communication capacity. J. Econ. Theory 184, 104940.

Lipnowski, Elliot, 2020. Equivalence of Cheap Talk and Bayesian Persuasion in a Finite Continuous Model. Working Paper.

Lipnowski, Elliot, Ravid, Doron, 2020. Cheap talk with transparent motives. Econometrica 88 (4), 1631–1660.

Lipnowski, Elliot, Ravid, Doron, Shishkin, Denis, 2022. Persuasion via weak institutions. J. Polit. Econ. 130 (10), 2705–2730.

Loginova, Uliana, 2012. Strategic communication with guilt aversion. Working Paper.

Nguyen, Anh, Tan, Yong Teck, 2019. Bayesian Persuasion with Costly Messages. Working Paper.

Ottaviani, Marco, Squintani, Francesco, 2006. Naive audience and communication bias. Int. J. Game Theory 35, 129–150.

Ottaviani, Marco, Sørensen, Peter Norman, 2006. Reputational cheap talk. Rand J. Econ. 37 (1), 155–175.

Perez-Richet, Eduardo, Skreta, Vasiliki, 2022. Test design under falsification. Econometrica 90 (3), 1109–1142.

Rockafellar, Ralph Tyrell, 1997. Convex Analysis. Princeton University Press.

Salamanca, Andres, 2021. The value of mediated communication. J. Econ. Theory 192, 105191.

Shiryaev, A.N., 1996. Probability. Springer-Verlag New York Inc.

Sobel, Joel, 2020. Lying and deception in games. J. Polit. Econ. 128 (3), 907–947.