Credible Persuasion

Xiao Lin

University of Pennsylvania

Ce Liu

Michigan State University

We propose a new notion of credibility for Bayesian persuasion problems. A disclosure policy is credible if the sender cannot profit from tampering with her messages while keeping the message distribution unchanged. We show that the credibility of a disclosure policy is equivalent to a cyclical monotonicity condition on the policy's induced distribution over states and actions. We also characterize how credibility restricts the sender's ability to persuade under different payoff structures. In particular, when the sender's payoff is state independent, all disclosure policies are credible. We apply our results to the market for lemons and show that no useful information can be credibly disclosed by the seller.

I. Introduction

When an informed party (sender; she) discloses information to persuade her audience (receiver; he), it is in her interest to convey only messages

We are indebted to Nageeb Ali for his continuing guidance and support. This paper has benefited from the thoughtful and constructive feedback of the editor, Emir Kamenica, and three anonymous reviewers. We are also grateful for comments and suggestions from Ian Ball, Carl Davidson, Jon Eguia, Henrique de Oliveira, Piotr Dworczak, Alex Frankel, Nima Haghpanah, Rick Harbaugh, Marc Henry, Tetsuya Hoshino, Yuhta Ishii, Navin Kartik, Vijay Krishna, SangMok Lee, George Mailath, Meg Meyer, Moritz Meyer-ter-Vehn, Arijit Mukherjee, Harry Pei, Daniel Rappoport, Andrew Rhodes, Ron Siegel, Alex Smolin, Juuso Toikka, Rakesh Vohra, Jia Xiang, Takuro Yamashita, and participants at various conferences and seminars. Siqi Li provided excellent research assistance.

Electronically published May 10, 2024

Journal of Political Economy, volume 132, number 7, July 2024.
© 2024 The University of Chicago. All rights reserved. Published by The University of Chicago Press.
https://doi.org/10.1086/728745

that steer the outcome in her own favor: schools may want to inflate their grading policies to improve their job placement records; credit rating agencies may publish higher ratings in exchange for future business. Even when the sender claims to have adopted a disclosure policy, she may still find it difficult to commit to following its prescriptions, since the adherence to such policies is often impossible to monitor. By contrast, what is often publicly observable is the final distribution of the sender's messages: students' grade distributions at many universities are publicly available, and so are the distributions of rating scores from credit rating agencies.

Motivated by this observation, we propose a notion of credible persuasion. In contrast to standard Bayesian persuasion, our sender cannot commit to a disclosure policy; however, to avoid detection, she must keep the final message distribution unchanged when deviating from her disclosure policy. For example, in the context of grade distributions, if a university had announced a disclosure policy that features certain fractions of A's, B's, and C's, it cannot switch to a distribution that assigns every student an A without being detected. Analogously, if a credit rating agency were to tamper with its rating scheme, any resulting change in the overall distribution of ratings would be detected. Our notion of credibility closely adheres to this definition of detectability: we say that a disclosure policy is credible if, given how the receiver reacts to her messages, the sender has no profitable deviation to any other disclosure policy that has the same message distribution.

Can the sender persuade the receiver by using credible disclosure policies? We find that in many settings, no informative disclosure policy is credible. An important case where this effect is exhibited is the market for lemons (Akerlof 1970). Here, we show that the seller of an asset cannot credibly disclose any useful information to the buyer; this effect arises even though the seller benefits from persuasion when she can fully commit to her disclosure policy. Conversely, we also provide conditions for when the sender is guaranteed to benefit from credible persuasion so that credibility does not entirely eliminate the scope for persuasion. In general, we show that credibility is characterized by a cyclical monotonicity condition, which is analogous to those studied in decision theory and mechanism design (Rochet 1987).

To illustrate these ideas, consider the following example. A buyer (receiver) chooses whether to buy a car from a used car seller (sender). It is common knowledge that 30% of the cars are of high quality and the remaining 70% are of low quality. For simplicity, suppose that all cars are sold at an exogenously fixed price.[1] The payoffs in this example are in

[1] In sec. III, we study a competitive market for lemons with endogenous prices and emerge with similar findings.

TABLE 1
USED CAR EXAMPLE PAYOFFS

	Buy	Not Buy
Seller:		
High	2	1
Low	2	0
Buyer:		
High	1	0
Low	−1	0

table 1. The seller always prefers selling a car, but the buyer is willing to purchase if and only if he believes its quality is high with at least 0.5 probability. Conditional on a car being sold, the seller obtains the same payoff regardless of its quality, but when a car is not sold, she receives a higher value from retaining a high-quality car.

As a benchmark, let us first see what the seller achieves if she could commit to a disclosure policy. We depict the optimal disclosure policy in figure 1. The policy uses two messages, pass and fail: all high-quality cars pass, along with 3/7 of the low-quality cars; the remaining 4/7 of the low-quality cars receive a failing grade. Conditional on the car passing, the buyer believes that the car is of high quality with probability 0.5, which is just enough to convince him to make the purchase. If a car fails, the buyer believes that the car is of low quality for sure and will refuse to buy. With this disclosure policy, the buyer expects to see the seller pass 60% of the cars and fail the remaining 40%.[2]

The policy above is optimal for the seller if she can commit to following its prescriptions. But suppose the buyer cannot observe how the seller rates her cars. Instead, the buyer observes only the fraction of cars being passed and failed. In such a setting, the seller can profitably deviate from the above disclosure policy without being detected by the buyer. Specifically, the seller can switch to failing all high-quality cars while adding an equal number of low-quality cars to the passing grade. This disclosure policy, illustrated in figure 2, induces the same distribution of messages (i.e., 60% pass, 40% fail). Holding fixed the buyer's behavior, this deviation is profitable for the seller because she still ends up selling the same number of cars but now is able to retain more high-quality cars. Accordingly, we view the optimal full-commitment policy to be not credible: after having promised to share information according to a disclosure policy, the seller would not find it rational to follow through and would instead profit from an undetectable deviation.

More generally, we introduce the following notion of credibility for disclosure policies. Consider a profile consisting of the sender's disclosure

[2] This example, by design, has the same solution as the prosecutor-judge example in Kamenica and Gentzkow (2011).

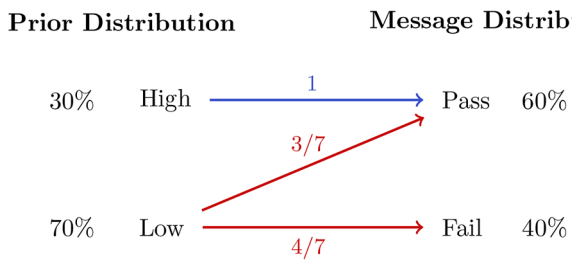

FIG. 1.—Optimal commitment policy.

policy and the receiver's strategy (mapping messages to actions). We say that a profile is receiver incentive compatible (R-IC) if the receiver's strategy best responds to the sender's disclosure policy—this requirement is standard in Bayesian persuasion problems. We say that a profile is credible if, given the receiver's strategy, the sender has no profitable deviation to any other disclosure policy that induces the same message distribution. Together, credibility and receiver incentive compatibility require that, conditional on the sender's message distribution, the sender and receiver best respond to each other.[3]

We have just argued that in the used car example, the optimal full-commitment disclosure policy was not credible, given the receiver's best response. Can any car be sold in a profile that is both credible and R-IC? The answer is no. Note that zero sales is also the outcome when no information is disclosed. In other words, credibility completely shuts down the possibility for useful information transmission.

To see why, suppose toward a contradiction that the buyer purchases a car after observing a message m_1 that is sent with positive probability. By receiver incentive compatibility, the buyer must believe that the car is of high quality with at least 0.5 probability after observing m_1. Since m_1 is sent with positive probability, the martingale property of beliefs implies that there must be another message m_2, also sent with positive probability, that makes the buyer assign less than 0.5 to the car's quality being high. Necessarily, when the buyer observes the message m_2, he does not make a purchase. This creates an incentive for the seller to tamper with her disclosure policy: by exchanging some of the good cars being mapped into m_1 with an equal number of bad cars being mapped into m_2, she can improve her payoff without changing the distribution of messages.

[3] Our solution concept is therefore analogous to an equilibrium condition in which the set of feasible deviations for the sender is to other disclosure policies that induce the same message distribution.

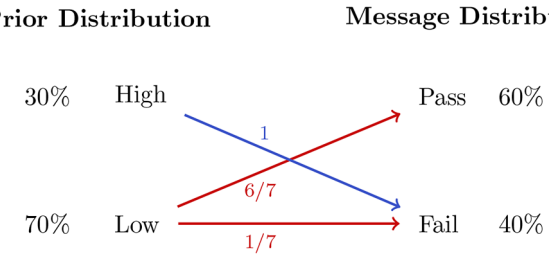

FIG. 2.—Undetectable deviation.

One may wonder whether credibility always shuts down communication entirely. The next example features a setting in which the optimal full-commitment disclosure policy is credible. Consider the disclosure problem faced by a school (sender) and an employer (receiver).[4] Just as in the used car example, a student's ability is either high with probability 0.3 or low with probability 0.7. Payoffs are as shown in table 2. The employer is willing to hire a student if he believes the student has high ability with at least 0.5 probability. The school would like all its students to be employed but derives a higher payoff from placing a good student than it does from placing a bad one.

The school's optimal full-commitment disclosure policy is identical to the one in the used car example (fig. 1), and so are the employer's best responses. But unlike the used car example, the school cannot profitably deviate without changing the message distribution.

To see why, note that without the message distribution being changed, any deviation must involve passing some low-ability students while failing an equal number of high-ability students. This would increase the employment of low-ability students at the expense of their high-ability counterparts, which makes the school worse off. Since the school cannot profit from undetectable deviations, the optimal full-commitment policy is credible. In contrast to the previous example, where credibility shuts down all useful communication, the current example shows that credibility sometimes imposes no cost on the sender relative to persuasion with full commitment.

In the two examples above, credibility has starkly different implications for information transmission. The key difference is that in the used car example, when the car's quality is higher, the sender has a weaker incentive to trade while the receiver's incentive to trade is stronger; in the school

[4] See Ostrovsky and Schwarz (2010) for an early study of how schools strategically design their grading policies in a competitive setting.

TABLE 2
SCHOOL EXAMPLE PAYOFFS

	Hire	Not Hire
School:		
High	2	0
Low	1	0
Employer:		
High	1	0
Low	−1	0

example, by contrast, both the sender and the receiver have a stronger incentive to trade as the student's ability increases. Our results formalize this intuition.

Proposition 2 shows that when the sender and receiver's preferences have opposite modularities (e.g., when the sender's payoff is strictly supermodular and the receiver's payoff is submodular), no useful information can be credibly communicated. Even when players' preferences share the same modularity, the sender does not always benefit from credible persuasion relative to the no-information benchmark. Propositions 3 and 4 provide additional conditions that guarantee that the sender does benefit from credible persuasion as well as conditions under which the optimal full-commitment disclosure policy is credible. Proposition 5 provides a comparative statics result on preference alignment.

Generalizing further, we use optimal transport theory to characterize credibility using a familiar condition from mechanism design and decision theory: cyclical monotonicity. Theorem 1 shows that for every profile of sender's disclosure policy and receiver's strategy, the credibility of the profile is equivalent to a cyclical monotonicity condition on its induced distribution over states and actions. As is illustrated in the examples above, credibility requires that the sender cannot benefit from any pairwise swapping in the matching of states and actions. The cyclical monotonicity condition generalizes this idea to cyclical swapping: for every sequence of state-action pairs in the support, the sum of the sender's utility should be lower after the matchings of states and actions in this sequence are permuted. In appendix section B.1 (app. B is available online), we discuss the connection of theorem 1 to Rochet (1987).

Our paper offers foundations for Bayesian persuasion models in settings where the sender provides information about a population of objects. In such environments, if the sender's payoff is state independent, all disclosure policies are credible, so the full-commitment assumption in the Bayesian persuasion approach is nonessential as long as the message distribution is observable. Additionally, our model also provides a rationale for considering monotone disclosure policies, which are credible when the sender's payoff is supermodular.

The rest of the paper is organized as follows. Section II introduces our credibility notion as well as the main results. Section III considers an application: in the market for lemons with endogenous prices, we show that the seller cannot credibly disclose any useful information to the buyers, even though full disclosure would maximize the seller's profit. Section IV discusses the effect of receiver mixing. Section V concludes. All omitted proofs are in appendix A. The remainder of this introduction places our contribution within the context of the broader literature.

Related literature.—Our work contributes to the study of strategic communication. The Bayesian persuasion model in Kamenica and Gentzkow (2011) studies a sender who can fully commit to a disclosure policy.[5] By contrast, the cheap talk approach pioneered by Crawford and Sobel (1982) models a sender who observes the state privately and, given the receiver's strategy, chooses an optimal (sequentially rational) message. The partial-commitment setting that we model is between these two extremes: here, the sender can commit to a distribution over messages but not the entire disclosure policy.

Our model considers a sender who can misrepresent her messages as long as the misrepresentation still produces the original message distribution. This contrasts with existing approaches to modeling limited commitment in Bayesian persuasion. One approach—pioneered by Min (2021), Fréchette, Lizzeri, and Perego (2022), and Lipnowski, Ravid, and Shishkin (2022)—is to allow the sender to alter the messages from her chosen disclosure policy with some fixed probability. A different method of modeling limited commitment is to consider settings where the sender can misreport at a cost.[6] For example, Guo and Shmaya (2021) study a sender who pays a cost when the posterior beliefs induced by her messages are miscalibrated from their literal meanings; Nguyen and Tan (2021) consider a sender who can costly revise the messages from her chosen disclosure policy; and Perez-Richet and Skreta (2021) consider a sender who can falsify the state, or input, of the disclosure policy. Another approach, taken in Libgober (2022), is to consider a sender who publicly chooses some dimension of the signal structure while privately choosing the other dimension. Finally, Perez-Richet (2014), Hedlund (2017), Koessler and Skreta (2021), and Zapechelnyuk (2023) allow the sender to have private information before choosing the disclosure policy. In these settings, the receiver infers the state through the messages from the disclosure policy as well as the signaling effect of the sender's choice of information structures.

The way that we model the sender's feasible deviations is closely related to the literature on quota mechanisms, which use message budgets to

[5] Brocas and Carrillo (2007) and Rayo and Segal (2010) also study optimal disclosure policy in more specific settings.

[6] This approach was initially introduced by Kartik (2009) to study language inflation.

induce truth-telling (see, e.g., Jackson and Sonnenschein 2007; Matsushima, Miyazaki, and Yagi 2010; Rahman 2010; Frankel 2014). Similar ideas have also been explored in communication games. For example, Chakraborty and Harbaugh (2007) consider multi-issue cheap talk problems and study equilibria where the sender assigns a ranking to each issue. In such equilibria, a message is a complete or partial ordering of all the issues, and any on-path deviation is a different ordering that maintains the same distribution of rankings. Renault, Solan, and Vieille (2013) study repeated cheap talk models where only messages and the receiver's actions are publicly observable. They characterize equilibria in the repeated communication game via a static reporting game where the sender directly reports her type. The key condition in their characterization requires truthful reporting to be optimal among all reporting strategies that replicate the true type distribution, which is akin to Rahman's (2010) characterization of implementable direct mechanisms. Margaria and Smolin (2018) use a different approach to study the case where the sender's payoff is state independent, and Meng (2021) provides a unified approach to characterizing the receiver's optimal value in these repeated cheap talk models. Kuvalekar, Lipnowski, and Ramos (2022) study a related model where the receiver is short-lived and show that the equilibrium payoffs can be characterized via a static cheap talk model with capped money burning.

A different strand of the repeated cheap talk literature studies models where the receiver can observe feedback about past state realizations. Best and Quigley (2020) consider how coarse feedback of past states can substitute for commitment; Mathevet, Pearce, and Stacchetti (2022) allow for the possibility of nonstrategic commitment types; and Pei (2020) studies a setting where the sender has persistent private information about her lying cost.

Finally, our approach to credible persuasion is reminiscent of how Akbarpour and Li (2020) model credible auctions. They study mechanism design problems where the designer's deviations are safe so long as they lead to outcomes that are possible when she is acting honestly, and they characterize mechanisms that ensure the designer has no safe and profitable deviations. By contrast, we study persuasion problems where the sender's deviations are undetectable if they do not alter the message distribution, and we characterize disclosure policies where the sender has no profitable and undetectable deviations.

II. Model

A. *Setup*

We consider an environment with a single sender (S; she) and a single receiver (R; he). Both players' payoffs depend on an unknown state $\theta \in \Theta$

and the receiver's action $a \in A$. Both Θ and A are finite sets.[7] The payoff functions are given by $u_S : \Theta \times A \to \mathbb{R}$ and $u_R : \Theta \times A \to \mathbb{R}$. Players hold full-support common prior $\mu_0 \in \Delta(\Theta)$.

Let M be a finite message space that contains A. The sender chooses an information structure to influence the receiver's action. Specifically, an information structure $\lambda \in \Delta(\Theta \times M)$ is a joint distribution of states and messages, so that the marginal distribution of states agrees with the prior; that is, $\lambda_\Theta = \mu_0$.[8] The receiver chooses an action after observing each message according to a pure strategy $\sigma : M \to A$.[9]

Our interest is in understanding the sender's incentives to deviate from her information structure, which depends on the receiver's strategy. To avoid ambiguity, we refer explicitly to pairs of (λ, σ)—or *profiles*—that consist of a sender's information structure and a receiver's strategy. For each profile (λ, σ), the players' expected payoffs are

$$U_S(\lambda, \sigma) = \sum_{\theta,m} u_S(\theta, \sigma(m))\lambda(\theta, m) \text{ and } U_R(\lambda, \sigma) = \sum_{\theta,m} u_R(\theta, \sigma(m))\lambda(\theta, m).$$

We consider a setting where the sender cannot commit to her information structure and can deviate to another information structure so long as it leaves the final message distribution unchanged. This embodies the notion that the distribution of the sender's messages is observable, even though it may be difficult to observe exactly how these messages are generated. Formally, if λ is an information structure promised by the sender, let $D(\lambda) \equiv \{\lambda' \in \Delta(\Theta \times M) : \lambda'_\Theta = \mu_0, \lambda'_M = \lambda_M\}$ denote the set of information structures that induce the same distribution of messages as λ: these information structures are indistinguishable from λ from the receiver's perspective. Our credibility notion requires that, conditioning on how the receiver responds to the sender's messages, no deviation in $D(\lambda)$ can be profitable for the sender.

DEFINITION 1. A profile (λ, σ) is *credible* if

$$\lambda \in \arg\max_{\lambda' \in D(\lambda)} \sum_{\theta,m} u_S(\theta, \sigma(m))\lambda'(\theta, m). \tag{1}$$

Moreover, the receiver's strategy is required to be a best response to the sender's information structure.

[7] In app. sec. B.6, we show that our main characterization result extends to the case where Θ and A are compact Polish spaces.

[8] For a probability measure P defined on some product space $X \times Y$, we use P_X and P_Y to denote its marginal distribution on X and Y, respectively.

[9] We focus on pure strategies to abstract from the receiver using randomization to deter the sender's deviations. This restriction is not without loss of generality, though some of our results can be extended to allow receiver mixing. See sec. IV for a more detailed discussion of this assumption.

DEFINITION 2. A profile (λ, σ) is *receiver incentive compatible* (R-IC) if

$$\sigma \in \arg\max_{\sigma' : M \to A} \sum_{\theta, m} u_R(\theta, \sigma'(m))\lambda(\theta, m). \tag{2}$$

Together, credibility and R-IC ensure that, conditioning on the message distribution of the sender's information structure, both the sender and the receiver best respond to each other. An immediate observation is that there always exists a babbling profile (λ, σ) that is both credible and R-IC: a degenerate information structure that sends only one message and a receiver strategy that best responds to the prior after observing any message.

Note that the credibility notion can be viewed as merely incorporating an additional constraint in the design of information structures. Some of our results focus on sender optimality, but the notion can be applied to different design objectives. It is also worth noting that credibility is a constraint that is independent from receiver incentive compatibility. As a result, our credibility notion can be applied more broadly to settings where the consequences of the sender's messages can be specified via an outcome function. As an application, we apply our credibility notion to a setting with multiple receivers in section III.

Finally, our credibility notion is motivated by the observability of the sender's message distribution, which we model as a restriction on the sender's feasible deviations. The observability of message distributions is best understood through a population interpretation of persuasion models,[10] where there is a continuum of objects with types distributed according to $\mu_0 \in \Delta(\Theta)$. The sender's information structure λ assigns each object a message based on its type, which generates a message distribution λ_M. Working with a continuum population affords us a cleaner exposition by abstracting from sampling variation. In appendix section B.2, we consider a finite approximation where the sender privately observes N independently and identically distributed samples from $\mu_0 \in \Delta(\Theta)$ and assigns each realization a message $m \in M$ subject to quotas on message frequencies; the receiver then chooses an action after observing the sender's message. We show that credible and R-IC profiles in our continuum model are approximated by those in the finite-sample model when the sample size N becomes large.

B. *Stable Outcome Distributions*

We characterize credible and R-IC profiles through the induced probability distribution of states and actions. Formally, an *outcome distribution* is a

[10] For a more detailed discussion of various interpretations of Bayesian persuasion models, see, e.g., sec. 2.2 of Kamenica (2019).

distribution $\pi \in \Delta(\Theta \times A)$ that satisfies $\pi_\Theta = \mu_0$: this is a consistency requirement that stipulates that the marginal distribution of states must conform to the prior. We say that an outcome distribution π is induced by a profile (λ, σ) if for every $(\theta, a) \in \Theta \times A$, $\pi(\theta, a) = \lambda(\theta, \sigma^{-1}(a))$, where σ^{-1} is the inverse mapping of σ. We are interested in characterizing outcome distributions that can be induced by profiles that are both credible and R-IC, and we refer to such outcome distributions as stable.

DEFINITION 3. An outcome distribution $\pi \in \Delta(\Theta \times A)$ is *stable* if it is induced by a profile (λ, σ) that is both credible and R-IC.

Our first result characterizes stable outcome distributions.

THEOREM 1. An outcome distribution $\pi \in \Delta(\Theta \times A)$ is stable if and only if

1. π is u_R-obedient: for each $a \in A$ such that $\pi(\Theta, a) > 0$,

$$\sum_{\theta \in \Theta} \pi(\theta, a) u_R(\theta, a) \geq \sum_{\theta \in \Theta} \pi(\theta, a) u_R(\theta, a') \text{ for all } a' \in A;$$

2. π is u_S-cyclically monotone: for any sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in \text{supp}(\pi)$ where $a_{n+1} \equiv a_1$,

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) \geq \sum_{i=1}^{n} u_S(\theta_i, a_{i+1}).$$

The first condition is the standard obedience constraint (Bergemann and Morris 2016; Taneva 2019), which specifies that the receiver finds it incentive compatible to follow the recommended action, given the belief that she forms when receiving that recommendation. The second condition—namely, u_S-cyclical monotonicity—is the new constraint that maps directly to our notion of credibility. While both the necessity and sufficiency of u_S-cyclical monotonicity can be proven by invoking the Kantorovich duality from optimal transport theory, below we outline a direct proof to better illustrate the intuition behind u_S-cyclical monotonicity. The full version of this proof can be found in lemma 2 in appendix A.

Consider an outcome distribution π and a sequence $(\theta_i, a_i)_{i=1}^n$ in the support of π. For intuition, let us regard π as a direct recommendation information structure. A cyclical deviation in this case consists of subtracting ε mass from (θ_i, a_i) while adding it to (θ_i, a_{i+1}) for each $i = 1, \ldots, n$, where $a_{n+1} \equiv a_1$. Each step of this cyclical deviation changes the sender's payoff by $\varepsilon[u_S(\theta_i, a_{i+1}) - u_S(\theta_i, a_i)]$, so the total change in the sender's payoff is

$$\varepsilon\left[\sum_{i=1}^{n} u_S(\theta_i, a_{i+1}) - \sum_{i=1}^{n} u_S(\theta_i, a_i)\right].$$

The cyclical monotonicity condition requires that the sender can find no profitable cyclical deviations.

To see why this is necessary for credibility, observe that cyclical deviations do not change the distribution of action recommendations, so any such deviation cannot be detected solely on the basis of the distribution of messages. Credibility requires that these undetectable deviations are not profitable, which implies the cyclical monotonicity condition.

For sufficiency, a key observation is that any outcome distribution $\pi \in \Delta(\Theta \times A)$ can be approximated by a distribution with rational marginals, which can then be normalized and transformed into doubly stochastic matrices. According to the Birkhoff–von Neumann theorem, permutation matrices form the extreme points of all doubly stochastic matrices. In addition, each permutation matrix corresponds to a cyclical deviation. So in a rough sense, cyclical deviations are (approximately) the extreme points of all undetectable sender deviations. It is therefore sufficient to ensure that no cyclical deviations are profitable.

When verifying the cyclical monotonicity condition, one can in fact restrict attention to deviations of length $n \leq \min\{|\Theta|, |A|\}$. This is because if there is any profitable cyclical deviation exceeding this length, we can split it into two shorter cyclical deviations, at least one of which is profitable. This observation, which is formalized in appendix section B.3, implies that it is sufficient to check a finite number of deviations. However, the total number of deviations can still be quite large, which may make it challenging to verify cyclical monotonicity. In section II.D, we impose additional structures on players' payoffs to gain further tractability.

The next result establishes the existence of a sender-optimal credible and R-IC profile and shows that it need not involve more than $\min\{|\Theta|, |A|\}$ messages.

PROPOSITION 1. There exists a sender-optimal credible and R-IC profile (λ^*, σ^*) where λ^* has no more than $\min\{|\Theta|, |A|\}$ messages.

The existence follows from the fact that the set of stable outcome distributions is compact. The bound on the number of messages in proposition 1 parallels a similar result for optimal persuasion under full commitment. For any sender-optimal stable outcome distribution π^*, we can take π^* as the direct recommendation information structure that uses no more than $|A|$ messages. For each $a \in \text{supp}(\pi^*_A)$, let μ_a denote the posterior belief induced by π^*, and v_a denote the sender's value under posterior μ_a. These (μ_a, v_a) pairs reside in $\mathbb{R}^{|\Theta|}$, but by the optimality of π^*, it can be further shown that all such (μ_a, v_a) pairs must lie on the same hyperplane in $\mathbb{R}^{|\Theta|}$, which has dimension $|\Theta| - 1$. Applying Carathéodory's theorem on this hyperplane allows us to obtain the $|\Theta|$ bound while reducing the support of the outcome distribution, which relaxes both the u_S-cyclical monotonicity and u_R-obedience constraints.

C. The Case of State-Independent Preferences

If $u_S(\theta, a)$ is state independent, then u_S-cyclical monotonicity is automatically satisfied. So we have the following observation.[11]

OBSERVATION 1. If $u_S(\theta, a) = h(a)$ for some $h: A \to \mathbb{R}$, then every outcome distribution that satisfies u_R-obedience is stable.

Therefore, in this case, there is no gap between what is achievable by a sender who can fully commit to an information structure relative to a sender who can commit to only a distribution of messages.

State-independent payoffs feature in many analyses of communication and persuasion (e.g., Chakraborty and Harbaugh 2010; Alonso and Câmara 2016; Lipnowski and Ravid 2020; Lipnowski, Ravid, and Shishkin 2022; Gitmez and Molavi 2022). In these settings, when the sender is not disclosing information about a population (thus making it difficult to observe the message distribution), generally the optimal full-commitment outcome cannot be achieved. By contrast, our analysis suggests that when one can adopt the population interpretation for Bayesian persuasion models, the sender can exercise full commitment power by making public the distribution of her messages.

D. When Is Credibility Restrictive?

When the state and action interact in the sender's payoff, credibility limits the sender's choice of information structures. The goal of this section is to understand how these limits can restrict the sender's ability to persuade the receiver.

In the examples in section I, we see that whether the sender can credibly persuade the receiver depends crucially on the alignment of their marginal incentives to trade. To understand this logic more generally, we assume that Θ and A are totally ordered sets, which without loss of generality can be assumed to be subsets of $\mathbb{R}$. Recall that a payoff function $u: \Theta \times A \to \mathbb{R}$ is *supermodular* if for all $\theta > \theta'$ and $a > a'$, we have

$$u(\theta, a) + u(\theta', a') \geq u(\theta, a') + u(\theta', a),$$

and it is *submodular* if

$$u(\theta, a) + u(\theta', a') \leq u(\theta, a') + u(\theta', a).$$

Furthermore, the function is *strictly supermodular* or *strictly submodular* if the inequalities above are strict for $\theta > \theta'$ and $a > a'$.

The modularity of players' payoff functions captures how the marginal utility from switching to a higher action varies with the state. This generalizes

[11] The same observation holds if $u(\theta, a) = r(\theta) + h(a)$ for some $r: \Theta \to \mathbb{R}$ and $h: A \to \mathbb{R}$, as adding an action-independent nuisance term does not change the sender's preferences over outcome distributions given the exogenous prior distribution on Θ.

the marginal incentive to trade in the examples in section I: intuitively, the sender and the receiver have aligned marginal incentives when both players' payoff functions share the same modularity and opposed marginal incentives when their payoff functions have opposite modularities. To fix ideas, we will assume that the sender's payoff is supermodular and vary the modularity of the receiver's payoff.

We now introduce a lemma that simplifies the u_S-cyclical monotonicity condition in theorem 1 when the sender's payoff is supermodular. Say that an outcome distribution $\pi \in \Delta(\Theta \times A)$ is *comonotone* if for all (θ, a), $(\theta', a') \in \text{supp}(\pi)$ satisfying $\theta < \theta'$, we have $a \leq a'$. Comonotonicity requires that the states and the receiver's actions are positive assortatively matched in the outcome distribution. The following lemma, whose variant appears in Rochet (1987), shows that u_S-cyclical monotonicity reduces to comonotonicity when the sender's preference is supermodular.

LEMMA 1. If u_S is supermodular, then every comonotone outcome distribution is u_S-cyclically monotone. Furthermore, if u_S is strictly supermodular, then every u_S-cyclically monotone outcome distribution is also comonotone.

Combined with theorem 1, lemma 1 implies that when the sender's preference is strictly supermodular, the credibility of a profile (λ, σ) is equivalent to the comonotonicity of its induced outcome distribution. Comonotone outcome distributions have attracted much attention in the persuasion literature in part due to their simplicity and ease of implementation (see, e.g., Goldstein and Leitner 2018; Kolotilin 2018; Dworczak and Martini 2019; Ivanov 2020; Kolotilin and Li 2021; Mensch 2021). Our credibility notion provides an additional motivation for focusing on monotone information structures.

REMARK 1. Lemma 1 is particularly relevant when $u_S(\theta, a)$ is affine in θ: that is, when there exist $\eta_0(a)$ and $\eta_1(a)$ such that $u_S(\theta, a) = \eta_0(a) + \eta_1(a)\theta$ for all θ, a. In this case, an outcome distribution π is u_S-cyclical monotone if and only if for all $(\theta, a), (\theta', a') \in \text{supp}(\pi)$ with $\theta < \theta'$, we have $\eta_1(a) \leq \eta_1(a')$. In other words, higher states are matched with actions that lead to higher slope terms in $u_S(\theta, a)$. The reason is that we can define an order on A: $a' \succcurlyeq a$ if and only if $\eta_1(a') \geq \eta_1(a)$, so that u_S is strictly supermodular with respect to such order.[12] The payoff function $u_S(\theta, a) = \eta_0(a) + \eta_1(a)\theta$ underlies much of the literature on posterior mean problems, which includes several of the papers cited above.

As benchmarks, we will often draw comparisons to what the sender can achieve when she can fully commit to her information structure as well as what is achievable when all or no information is disclosed. We say an outcome distribution π^* is an *optimal full-commitment outcome* if it maximizes the sender's payoff among outcome distributions that satisfy u_R-obedience. An

[12] Note that the order $\succcurlyeq$ defined as such may not be antisymmetric. Nevertheless, the proof of lemma 1 holds as long as $\succcurlyeq$ is complete and transitive. In app. A, we prove lemma 1 without assuming antisymmetry.

outcome distribution $\bar{\pi}$ is a *fully revealing outcome* if the receiver always chooses a best response to every state; that is,

$$a \in \arg\max_{a' \in A} u_R(\theta, a') \text{ for every } (\theta, a) \in \text{supp}(\bar{\pi}).$$

Finally, an outcome distribution $\underline{\pi}$ is a *no-information outcome* if the receiver always chooses the same action that best responds to the prior belief μ_0; in other words, there exists

$$a^* \in \arg\max_{a \in A} \sum_{\theta \in \Theta} \mu_0(\theta) u_R(\theta, a) \text{ such that } \underline{\pi}_A(a^*) = 1.$$

We say that the sender *benefits from persuasion* if an optimal full-commitment outcome gives the sender a higher payoff than every no-information outcome. Similarly, we say that the sender *benefits from credible persuasion* if there exists a stable outcome distribution that gives the sender a strictly higher payoff than every no-information outcome.

1. When Credibility Shuts Down Communication

The next result generalizes the used car example in section I. To simplify the statement of the result, we impose the following regularity assumption on the receiver's payoff function.

ASSUMPTION 1. There exist no distinct $a, a' \in A$ such that $u_R(\theta, a) = u_R(\theta, a')$ for all $\theta \in \Theta$.

In other words, from the receiver's perspective, there are no duplicate actions. This assumption is not without loss but greatly simplifies the statement of propositions 2 and 3.

PROPOSITION 2. Under assumption 1, if u_S is strictly supermodular and u_R is submodular, then every stable outcome distribution is a no-information outcome.

Proposition 2 says that when the players have opposed marginal incentives, credibility completely shuts down information transmission. The logic generalizes what we saw in the used car example: if two distinct messages resulted in different actions from the receiver, the sender and receiver would have diametrically opposed preferences regarding which action to induce in which state. Therefore, whenever R-IC is satisfied, the sender will have an incentive to deviate to another information structure that swaps states and induces the same marginal distribution of messages.

2. When the Sender Benefits from Credible Persuasion

In light of the school example in section I, one might expect credibility to not limit the sender's ability to persuade when her marginal incentives

are aligned with the receiver's. However, this is false without imposing additional assumptions. For an illustration, consider the following example, in which both the sender and receiver have supermodular payoffs. The sender benefits from persuasion when she can fully commit to her information structure, but no stable outcome distribution can give her a higher payoff than the best no-information outcome.

EXAMPLE 1. Suppose $\Theta = \{H, L\}$ with prior $\mu_0 = P(\theta = H) = 0.6$ and $A = \{a_1, a_2, a_3, a_4\}$. The sender and receiver's payoffs are as given in table 3. Note that both players' payoffs are strictly supermodular. The receiver's best response is a_1 when $\mu_0 \in [0, 0.25)$, a_2 when $\mu_0 \in [0.25, 0.5)$, a_3 when $\mu_0 \in [0.5, 0.75)$, and a_4 when $\mu_0 \in [0.75, 1]$; this leads to the sender's indirect utility function (solid line) and its concave envelope (dotted line) depicted in figure 3. From Kamenica and Gentzkow (2011), the dotted line represents the sender's optimal value under full commitment. It is clear that at $\mu_0 = 0.6$, the sender strictly benefits from persuasion if she can fully commit to her information structure. However, no stable outcome distribution can make the sender better off than the no-information outcome. To see why, first note that according to proposition 1, it is without loss to look for sender-optimal credible and R-IC profiles that induce only two posterior beliefs $\mu_1 < \mu_2$. Now consider the receiver's actions induced by these two posteriors. By lemma 1, at most one of these actions can be matched with more than one state, for otherwise the outcome distribution would not be comonotone. So at most one of the actions can be induced by interior posterior beliefs. However, it is clear from figure 3 that in order for the sender to benefit from using only two posteriors, she must induce both a_2 and a_3, both of which can happen only when the receiver holds interior beliefs. As a result, no credible and R-IC profiles can make the sender better off.

Example 1 above shows that besides the comodularity of preferences, additional conditions are needed in order to ensure the sender can benefit from credible persuasion. Proposition 3 offers several such conditions.

Let $A^\circ \equiv \{a \in A : a \in \arg\max_{a'} \Sigma_\theta \mu(\theta) u_R(\theta, a') \text{ for some } \mu \in \Delta(\Theta)\}$ denote the set of actions that are best responses to some belief of the receiver; clearly, actions that are not in A° would never be played by the receiver in

TABLE 3
PLAYERS' PAYOFFS IN EXAMPLE 1

	a_1	a_2	a_3	a_4
$u_S(\theta, a)$:				
$\theta = H$	−1	.75	1	0
$\theta = L$	0	.75	.5	−1
$u_R(\theta, a)$:				
$\theta = H$	0	.6	.8	1
$\theta = L$	1	.8	.6	0

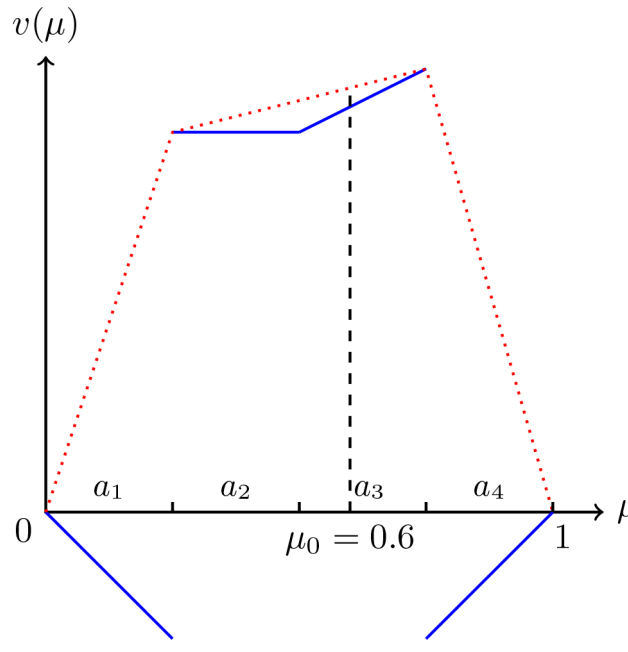

FIG. 3.—Concavification.

any R-IC profile and can without loss be discarded from the action set A. Let $\bar{a} \equiv \max A^\circ$ and $\underline{a} \equiv \min A^\circ$ denote the highest and lowest actions in A°, and let $\bar{\theta} \equiv \max \Theta$ and $\underline{\theta} \equiv \min \Theta$ denote the highest and lowest states.

PROPOSITION 3. Suppose that both u_S and u_R are supermodular and assumption 1 holds; then

1. if the highest action is dominant for the sender, that is, if $u_S(\theta, \bar{a}) > u_S(\theta, a)$ for all θ and $a \in A^\circ \backslash \{\bar{a}\}$, then for generic priors,[13] the sender benefits from credible persuasion as long as she benefits from persuasion;
2. if the sender favors extreme actions in extreme states, that is, if $u_S(\bar{\theta}, \bar{a}) > u_S(\bar{\theta}, a)$ for all $a \neq \bar{a}$ and $u_S(\underline{\theta}, \underline{a}) > u_S(\underline{\theta}, a)$ for all $a \neq \underline{a}$, then for generic priors, the sender benefits from credible persuasion; and

[13] Formally, by generic we mean that the result holds under a set of priors $T \subset \Delta(\Theta)$ that is open and dense and has full Lebesgue measure.

3. if the sender is strictly better off from a fully revealing outcome than from every no-information outcome, then the sender benefits from credible persuasion.

The first condition in proposition 3 is satisfied in settings like the school example, where the school and the employer's preferences are both supermodular, and the school would always want to place a student regardless of the student's ability. The second condition is applicable in environments where both parties have agreement on extreme states. For example, both doctors and patients favor aggressive treatment if the patient's condition is severe, and both favor no treatment if the patient is healthy, but they might disagree in intermediate cases. Last, a special case of the third condition is quadratic loss preferences, as commonly used in models of strategic communication (e.g., Crawford and Sobel 1982).[14] However, note that the conditions in proposition 3 do not guarantee the sender her optimal full-commitment payoff. In appendix section B.4, we provide an example satisfying the first condition in proposition 3. The sender in this example can benefit from credible persuasion but is unable to achieve the optimal full-commitment payoff.

The first two parts of proposition 3 are based on belief splitting. Let us briefly describe the proof for the first condition; the proof for the second part follows similar arguments. Note that if $\bar{a}$ is a dominant action for the sender, and the sender can benefit from persuasion (under full commitment), then $\bar{a}$ must not already be a best response for the receiver under the prior μ_0. The sender can then split the prior into a point mass posterior δ_{θ} and some other posterior $\tilde{\mu}$ that is close to μ_0. At δ_{θ}, the receiver is induced to choose $\bar{a}$ since his payoff is supermodular. In addition, for generic priors, the receiver's best response to $\tilde{\mu}$ remains the same as his best response to μ_0. The sender benefits from this belief splitting since the same action is still played most of the time, but in addition her favorite action is now played with positive probability. Moreover, the resulting outcome distribution matches higher states with higher actions, so it is stable because of the supermodularity of u_S and lemma 1.

The third part of proposition 3 follows because the fully revealing outcome distribution is always credible when both players' preferences are supermodular. The intuition of this result is most transparent when the sender's payoff is strictly supermodular. Consider (θ, a) and (θ', a') in the support of a fully revealing outcome distribution π, so a and a' best respond to θ and θ', respectively. From Topkis (2011), it follows that $a \geq a'$ if $\theta > \theta'$. Therefore, π is comonotone and satisfies u_S-cyclical monotonicity

[14] The model in this section has finite action spaces, so we need to additionally assume that the action space is rich enough such that the sender's indirect utility function approximates the one under a continuous action space.

by lemma 1. By construction, π also satisfies u_R-obedience, so π is stable by theorem 1. This result is closely related to theorem 1 and theorem 2 of Chakraborty and Harbaugh (2007). They show that in multi-issue cheap talk problems, truthfully revealing the rankings of the issues is an equilibrium under supermodular preferences; in addition, when the number of issues grows to infinity, revealing their rankings is asymptotically equivalent to revealing their values. The credibility of the fully revealing outcome can therefore be viewed as the limit of a rank revealing equilibrium in Chakraborty and Harbaugh (2007).

3. When Credibility Imposes No Cost to the Sender

In observation 1, we see that when the sender's payoff is additively separable, credibility does not restrict the set of stable outcomes. Proposition 4 provides a condition that guarantees that credibility imposes no loss on the sender's optimal value, even when credibility does restrict the set of stable outcomes.

PROPOSITION 4. Suppose $|A| = 2$. If both u_S and u_R are supermodular, then at least one optimal full-commitment outcome is stable; if in addition u_S is strictly supermodular, then every optimal full-commitment outcome is stable.

Proposition 4 says that in settings where both players have supermodular payoffs and the receiver faces a binary decision, such as accept or reject, then credibility imposes no cost to the sender. This result follows from combining our theorem 1 and lemma 1 with theorem 1 in Mensch (2021). He shows that under the assumptions in our proposition 4, there exists an optimal full-commitment outcome that is comonotone. The intuition is that for any outcome distribution π that is u_R-obedient but not comonotone, the sender can weakly improve her payoff by swapping the noncomonotone pairs in the support of π, so that they become matched assortatively. Such swapping also benefits the receiver because of the supermodularity of u_R, so u_R-obedience remains satisfied. As a result, the sender can always transform a noncomonotone outcome distribution into one that is comonotone without violating u_R-obedience while weakly improving her own payoff. Therefore, there must be an optimal full-commitment outcome that is comonotone, which is also stable by theorem 1 and lemma 1.

4. Comparative Statics

Our analysis thus far demonstrates that the mode of preference alignment plays a crucial role in determining the scope of credible persuasion. In this section, we provide a comparative statics result relating the sender's optimal credible persuasion payoff to the degree of preference alignment.

In order to measure the sender's utility on a constant scale, we will keep the sender's payoff function unchanged and adjust only the receiver's payoff.[15] Following section IV of Kamenica and Gentzkow (2011), we say preferences (u_S, u'_R) are more aligned than (u_S, u_R) if for any $a \in A$ and any $\mu \in \Delta(\Theta)$,

$$E_\mu[u_S(\theta, \hat{a}(\mu))] \geq E_\mu[u_S(\theta, a)] \Rightarrow E_\mu[u_S(\theta, \hat{a}'(\mu))] \geq E_\mu[u_S(\theta, a)],$$

where $\hat{a}(\mu) \in \arg\max_{a \in A} \Sigma_\theta \mu(\theta) u_R(\theta, a)$ and $\hat{a}'(\mu) \in \arg\max_{a \in A} \Sigma_\theta \mu(\theta) u'_R(\theta, a)$ denote the receiver's best response function, with ties broken in the sender's favor.

The following result shows that when payoffs are supermodular and preferences become more aligned, the sender is guaranteed a higher payoff from credible persuasion.[16]

PROPOSITION 5. Suppose that u_S, u_R, and u'_R are strictly supermodular payoff functions. If in addition the preferences (u_S, u'_R) are more aligned than (u_S, u_R), then under (u_S, u'_R), the sender obtains a higher payoff from a sender-optimal stable outcome distribution compared with under (u_S, u_R).

To prove proposition 5, we take an optimal stable outcome distribution π under the less aligned preferences (u_S, u_R) and show that when this same π is used as an information structure under the more aligned preferences (u_S, u'_R), it induces a stable outcome distribution that offers the sender a superior payoff. Specifically, consider the outcome distribution π' induced by the receiver choosing the sender-favored best responses to π under (u_S, u'_R). Since π is a stable outcome distribution under (u_S, u_R) and u_S is supermodular, it follows that π must be comonotone; this combined with the fact that u'_R is supermodular implies that π' is also comonotone and therefore stable under (u_S, u'_R). Moreover, as (u_S, u'_R) is more aligned than (u_S, u_R), following each message from π, the receiver's chosen action in π' is more favorable to the sender than the recommended action from π. The sender obtains a higher payoff from π' compared with π and therefore must be better off under (u_S, u'_R) than (u_S, u_R).

It is worth noting that under the assumptions of proposition 5, more aligned preferences do not guarantee a larger set of stable outcome distributions. We illustrate this point with an example in appendix

[15] In fact, the change of u_S would affect only the sender's optimal credible persuasion payoff through a scaling effect: according to theorem 1 and lemma 1, credibility is equivalent to the outcome distribution being comonotone as long as u_S is strictly supermodular. So under our maintained assumptions on payoff functions, the set of stable outcome distributions would be unaffected by modifications in the sender's payoff function.

[16] As an example, app. sec. B.5.1 provides a class of preferences that meets the requirements of proposition 5.

section B.5.2, where the set of stable outcome distributions first expands and then shrinks as the players' preferences become more aligned.

III. Application: The Market for Lemons

A classic insight from Akerlof (1970) is that in markets with asymmetric information, adverse selection can lead to substantial efficiency loss. In practice, buyers and sellers often rely on warranty or third-party certification to overcome this inefficiency. A seemingly more direct solution to their predicament is for the seller to fully reveal her private information, so that there is no information asymmetry between players. In this section, however, we show that this apparently easy fix to the adverse selection problem relies on unrealistic assumptions on the seller's ability to commit. Indeed, we show that any information disclosure that improves efficiency cannot be credible.

To fix ideas, we adapt the formulation in Mas-Colell, Whinston, and Green (1995) and consider a seller who values an asset she owns (say, a car) at $\theta \in \Theta \subseteq [0, 1]$; two buyers (1 and 2) both value the car at $v(\theta)$, which is weakly increasing in θ. Buyers share a common prior belief $\mu_0 \in \Delta(\Theta)$. We assume $v(\theta) > \theta$ for all $\theta \in \Theta$ so there is common knowledge of gain from trade. Moreover, we assume $E_{\mu_0}[v(\theta)] < 1$ so that without information disclosure, some cars will not be traded because of adverse selection. Below we first describe the base game without information disclosure and then augment the base game to allow the seller to choose an information structure to influence the buyers' beliefs.

The base game G.—The seller and the buyers move simultaneously. The seller learns her value and chooses an ask price $a_s \in A_S = [0, v(1)]$; each buyer $i = 1, 2$ chooses a bid $b_i \in A_i = [0, v(1)]$. If the ask price is lower than or equal to the highest bid, the car is sold at the highest bid to the winning buyer, and ties are broken evenly. If the ask price is higher than the highest bid, the seller keeps the car and receives the reserve value θ, while both buyers get 0. More formally, the seller's payoff function is

$$u_S(\theta, a_S, b_1, b_2) = \begin{cases} \max\{b_1, b_2\} & \text{if } a_S \leq \max\{b_1, b_2\}, \\ \theta & \text{if } a_S > \max\{b_1, b_2\}, \end{cases}$$

and buyer i's payoff is

$$u_i(\theta, a_S, b_1, b_2) = \begin{cases} v(\theta) - b_i & \text{if } b_i > b_{-i} \text{ and } b_i \geq a_S, \\ \frac{1}{2}[v(\theta) - b_i] & \text{if } b_i = b_{-i} \text{ and } b_i \geq a_S, \\ 0 & \text{otherwise.} \end{cases}$$

The game with disclosure.—Let M be the set of messages, which we assume is a Polish space. Before the base game is played, the seller chooses an information structure λ to publicly disclose information to the buyers.[17] Together the information structure λ and the base game G define a Bayesian game $\langle G, \lambda \rangle$. Every message m from the information structure λ induces a posterior belief $\mu_m \equiv \lambda(\cdot|m) \in \Delta(\Theta)$ for the buyers. The buyers $i = 1, 2$ choose their respective bids $\beta_i(m)$, while the seller chooses an ask price $\alpha_S(\theta, m)$. We restrict attention to Bayesian Nash equilibria where the seller plays her weakly dominant strategy $\alpha_S(\theta, m) = \theta$, and buyers play pure strategies. As we show in lemma 6, such equilibria exist in $\langle G, \lambda \rangle$ for every λ. These equilibria also give rise to the familiar fixed-point characterization of equilibrium price: buyers' bids satisfy

$$\max\{\beta_1(m), \beta_2(m)\} = E_{\mu_m}[v(\theta)|\theta \leq \max\{\beta_1(m), \beta_2(m)\}].$$

The trading game above differs from the sender-receiver setting in section II in two ways: first, the sender in the current setting publicly discloses information to multiple receivers; second, in addition to the receivers, the sender also chooses an action (ask price) after observing the realization of the information structure. Nevertheless, the notion of stable outcome distribution extends to the current setting. In particular, the credibility notion is based on the same idea that the sender cannot profitably deviate to a different information structure without changing the message distribution. The R-IC condition, meanwhile, is replaced by a new incentive compatible condition that asks both the sender and receivers to play according to a Bayesian Nash equilibrium in $\langle G, \lambda \rangle$. As mentioned above, in the market for lemons, we will focus on a special class of Bayesian Nash equilibria in the game $\langle G, \lambda \rangle$ where the seller plays her weakly dominant strategy $\alpha_S(\theta, m) = \theta$, and the buyers do not mix. We will call such profiles (λ, σ) WD-IC to distinguish from the weaker incentive compatible requirement. The formal discussion of our credibility notion in this multiple receiver setting is notationally cumbersome and is deferred to appendix section B.7.

Next, we state our result, discuss its implications, and provide intuition for its proof. As a benchmark, fix an arbitrary message $m_0 \in M$, and let $\lambda_0 \equiv \mu_0 \times \delta_{m_0}$ be a null information structure. Let R_0 denote the supremum of the seller's payoffs among profiles (λ_0, σ) that are WD-IC, so R_0 represents the highest equilibrium payoff the seller can achieve when providing no information.

PROPOSITION 6. Under every credible and WD-IC profile, the seller's payoff is no more than R_0.

[17] In our setting, λ determines only the buyers' information structure, and the seller is perfectly informed about θ. That is, the seller cannot prevent herself from learning the true quality of the car. This differs from Kartik and Zhong (2019), who fully characterize payoffs in the market for lemons under all possible information structures.

Proposition 6 implies that any information that can be credibly disclosed is not going to improve the seller's payoff compared with the no-information benchmark. This is in sharp contrast to the full-commitment case, where the seller would like to fully reveal the car's quality, and all car types θ are sold at $v(\theta)$, which would allow the seller to capture all surplus from trade.

Let us describe the intuition behind the proof for proposition 6.[18] For each message m from the seller's information structure λ, let $\Theta(m)$ denote the support of the buyer's posterior belief after observing m. A key step in proving proposition 6 is to show that there exists a common trading threshold τ such that for each message m, a car of quality $\theta \in \Theta(m)$ is traded if and only if $\theta \leq \tau$. To see why, suppose toward a contradiction that the trading threshold in message m is higher than the threshold in another message m'. We show in the proof that the seller would then have a profitable deviation by swapping some of the cars slightly below the higher threshold in message m with an equal amount of cars from m' that are of worse quality.[19] Because this deviation does not change the seller's message distribution, it is also undetectable. Therefore, credibility demands a common threshold τ that applies across messages. Given this common threshold τ, we then apply Tarski's fixed point theorem to show that when no information is disclosed, there is an equilibrium that features a higher trading threshold $\tau' \geq \tau$. Since a higher threshold means more cars are being traded, which in turn increases the seller's payoff, the seller's payoff under every stable outcome is therefore weakly worse than her payoff from a no-information outcome, and this proves our result.

IV. Discussion: Receiver Mixing

While our paper focuses on the receiver playing pure strategies, the notion of credible and R-IC profiles can be extended to allow for receiver mixing. Suppose the message space M is a Polish space that contains $\Delta(A)$ as a subset. A profile (λ, σ) consisting of the sender's information structure $\lambda \in \Delta(\Theta \times M)$ and the receiver strategy $\sigma : M \to \Delta(A)$ is (mixed strategy) credible if

[18] While the message of proposition 6 is reminiscent of proposition 2, it requires a different proof since the seller has a private action, so theorem 1 does not apply. Instead of working with the outcome distribution $\pi \in \Delta(\Theta \times A)$, here we apply the cyclical monotonicity characterization directly to the seller's information structure $\lambda \in \Delta(\Theta \times M)$ by invoking an optimal transport result from Beiglböck et al. (2009).

[19] This deviation is profitable because it allows the seller to replace the higher-quality cars traded in m with the lower-quality, untraded cars in m'. After this swapping, the lower-quality cars are now sold at the price for the higher-quality cars in m, while the higher-quality cars are now retained by the seller in m'.

$$\lambda \in \arg\max_{\lambda' \in D(\lambda)} \int_{\Theta \times M} \tilde{u}_S(\theta, \sigma(m)) d\lambda'(\theta, m),$$

and (mixed strategy) R-IC if

$$\sigma \in \arg\max_{\sigma' : M \to \Delta(A)} \int_{\Theta \times M} \tilde{u}_R(\theta, \sigma'(m)) d\lambda(\theta, m),$$

where $\tilde{u}_S : \Theta \times \Delta(A) \to \mathbb{R}$ and $\tilde{u}_R : \Theta \times \Delta(A) \to \mathbb{R}$ are extensions of u_S and u_R to mixed strategies, respectively.

As is illustrated in the following example, allowing mixed strategies can sometimes enlarge the set of payoffs achievable through credible persuasion.[20] This is based on similar ideas that appeared in Chakraborty and Harbaugh (2010) and Lipnowski and Ravid (2020): by mixing actions that the sender finds unappealing with those that she finds desirable, the receiver can reduce the scope of the sender's profitable deviations.

EXAMPLE 2. Suppose $\Theta = \{\theta_1, \theta_2\}$ with equal priors, and $A = \{a_1, a_2, a_3\}$. Consider the payoff matrices in table 4. In this example, a_3 is the most desirable action for the sender. We will show that without receiver mixing, the sender can never induce the receiver to play a_3 through credible persuasion; however, with receiver mixing, the sender can achieve a higher payoff by persuading the receiver to take a_3 with positive probability.

First, we show that without mixing, the receiver will never play a_3. In particular, we argue that any stable outcome distribution π^* must satisfy $\pi^*_A(a_3) = 0$. Suppose by contradiction that $\pi^*_A(a_3) > 0$. Since a_3 is weakly dominated by a_2, the receiver will play a_3 only under the point mass belief on θ_2. It follows that $\pi^*(\theta_2 | a_3) = 1$, so $\pi^*(\theta_1, a_3) = 0$. Therefore, either $\pi(\theta_1, a_1) > 0$ or $\pi(\theta_1, a_2) > 0$. However, recall that (θ_2, a_3) is in the support of π^* and

$$u_S(\theta_1, a_1) + u_S(\theta_2, a_3) < u_S(\theta_1, a_3) + u_S(\theta_2, a_1),$$
$$u_S(\theta_1, a_2) + u_S(\theta_2, a_3) < u_S(\theta_1, a_3) + u_S(\theta_2, a_2).$$

So both cases violate u_S-cyclical monotonicity. This proves that only a_1 and a_2 can be induced in any stable outcome distribution. In fact, the best the sender can do with credible persuasion is to fully reveal the states, which gives the sender a payoff of 1.

Next, we show that the sender can achieve a strictly higher payoff with receiver mixing. Consider the profile where the sender fully reveals the state ($\lambda(\theta_1, m_1) = \lambda(\theta_2, m_2) = 1/2$), and the receiver plays $\sigma(m_1) = \delta_{a_1}$ and $\sigma(m_2) = (1/2)\delta_{a_2} + (1/2)\delta_{a_3}$, with δ denoting the Dirac measure. This profile is clearly R-IC. Moreover, without changing the distribution of

[20] We thank a referee for suggesting this example.

TABLE 4
SENDER AND RECEIVER'S PAYOFFS

	a_1	a_2	a_3
u_S:			
θ_1	1	0	4
θ_2	0	1	2
u_R:			
θ_1	1	0	−1
θ_2	0	1	1

messages, the only deviation the sender has is pairwise-swapping probability mass from (θ_1, m_1) and (θ_2, m_2) to be placed on (θ_1, m_2) and (θ_2, m_1). This is not profitable because

$$\tilde{u}_S(\theta_1, \delta_{a_1}) + \tilde{u}_S\left(\theta_2, \frac{1}{2}\delta_{a_2} + \frac{1}{2}\delta_{a_3}\right) = 1 + 1.5 > 2 = \tilde{u}_S\left(\theta_1, \frac{1}{2}\delta_{a_2} + \frac{1}{2}\delta_{a_3}\right) + \tilde{u}_S(\theta_2, \delta_{a_1}).$$

Therefore, this strategy profile is (mixed strategy) credible and R-IC. Moreover, the sender achieves a strictly higher payoff of 1.25 from this mixed strategy profile than any pure strategy credible and R-IC profile.

Despite the gap between pure and mixed strategies illustrated by the example above, some of our results can be extended to cover receiver mixing. In appendix section B.6, we provide a variant of theorem 1 (theorem 1^*) for the case when Θ and A are both compact Polish spaces. If we view an outcome distribution $\pi \in \Delta(\Theta \times \Delta(A))$ as direct recommendations for mixed strategies, theorem 1^* then characterizes credibility as $\tilde{u}_S$-cyclical monotonicity on the space $\Theta \times \Delta(A)$.

As a more specific example, when the receiver's action is binary, proposition 2 holds even when allowing for receiver mixing. In particular, proposition 2^* in appendix section B.6 extends proposition 2 to the case when Θ and A are both compact subsets of $\mathbb{R}$. When the receiver's action is binary, the set of mixed strategies can be identified with the interval $[0, 1]$, and the extended payoff functions $\tilde{u}_S$ and $\tilde{u}_R$ preserve the super(sub-)modularity of u_S and u_R. So as a corollary of proposition 2^*, no information can be credibly transmitted in this case, and focusing on pure strategies in proposition 2 is without loss of generality when the receiver's action is binary.

V. Conclusion

This paper offers a new notion of credibility for persuasion problems. We model a sender who can commit to an information structure only up to the details that are observable to the receiver. The receiver does not observe the chosen information structure but observes the distribution of

messages. This leads to a model of partial commitment, where the sender can undetectably deviate to information structures that induce the same distribution of messages. Our framework characterizes when, given the receiver's best response, the sender has no profitable deviation.

We show that this consideration eliminates the prospects for credible persuasion in settings with adverse selection. In some other settings, we show that the requirement is compatible with the sender still benefiting from persuasion. More generally, we show that our requirement translates to a cyclical monotonicity condition on the induced distribution of states and actions.

Our work also speaks to why certain industries (such as education) can effectively disclose information by utilizing their own rating systems, while some other industries (such as car dealerships) must resort to other means to address asymmetric information, such as third-party certification or warranties. Our results provide a rationale: in industries that exhibit adverse selection, the informed party cannot credibly disclose information through its own ratings even if it wishes to do so.

The notion of credibility we consider in this paper is motivated by the observability of the sender's message distribution. In some settings, the receiver may observe more than the distribution of messages; for example, she may observe some further details of the information structure, such as how some states of the world map into messages. In other settings, the receiver may observe less; for example, she may see the average grade but not its distribution. To capture these various cases, one would then formulate the problem of detectable deviations differently. We view it to be an interesting direction for future research to understand how different notions of detectability map into different conditions on the outcome distribution.

Appendix A

Proofs

A1. Proof of Theorem 1

A1.1. Lemma 2 and Its Proof

The following lemma, which will play a key role in the proof of theorem 1, is a finite version of theorem 5.10 of Villani (2008). Below we present a direct proof of the lemma to better illustrate the intuition behind theorem 1.

LEMMA 2. Suppose that both X and Y are finite sets and $u : X \times Y \to \mathbb{R}$ is a real function. Let $\mu \in \Delta(X)$ and $\nu \in \Delta(Y)$ be two probability measure on X and Y, respectively, and $\Pi(\mu, \nu)$ be the set of joint probability measure on $X \times Y$ such that the marginals on X and Y are μ and ν. The following two statements are equivalent:

1. $\pi^* \in \arg\max_{\pi \in \Pi(\mu,\nu)} \Sigma_{x,y} \pi(x, y) u(x, y)$; and
2. π^* is u-cyclically monotone; that is, for any n and $(x_1, y_1), \ldots, (x_n, y_n) \in \text{supp}(\pi^*)$,

$$\sum_{i=1}^{n} u(x_i, y_i) \geq \sum_{i=1}^{n} u(x_i, y_{i+1}),$$

where $y_{n+1} \equiv y_1$.

Proof. $(1 \Rightarrow 2)$ To see the necessity of u-cyclical monotonicity, suppose by contraposition that π^* is not u_S-cyclically monotone, which implies that there exists a sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \text{supp}(\pi^*)$ such that

$$\sum_{i=1}^{n} u(x_i, y_i) < \sum_{i=1}^{n} u(x_i, y_{i+1}),$$

where $y_{n+1} = y_1$. Take $0 < \varepsilon \leq (1/n)\min_{i=1,\ldots n} \pi^*(x_i, y_i)$,[21] and define

$$\pi^\varepsilon \equiv \pi^* + \varepsilon \sum_{i=1}^{n} \left[\delta_{(x_i, y_{i+1})} - \delta_{(x_i, y_i)}\right],$$

where $\delta_{(x,y)}$ denotes the Dirac measure on (x, y). Note that $\pi^\varepsilon \in \Pi(\mu, \nu)$ and satisfies

$$\begin{aligned} &\sum_{x,y} u(x, y)\pi^\varepsilon(x, y) \\ &= \sum_{x,y} u(x, y)\pi^*(x, y) + \varepsilon\left[\sum_{i=1}^{n} u(x_i, y_{i+1}) - \sum_{i=1}^{n} u(x_i, y_i)\right] \\ &> \sum_{x,y} u(x, y)\pi^*(x, y), \end{aligned}$$

which implies $\pi^* \notin \arg\max_{\pi \in \Pi(\mu,\nu)} \Sigma_{x,y} \pi(x, y) u(x, y)$.

$(1 \Leftarrow 2)$ First note that π^* being u-cyclically monotone is equivalent to the following: for any n and $(x_1, y_1), \ldots, (x_n, y_n) \in \text{supp}(\pi^*)$ and for any permutation $s : \{1, \ldots, n\} \to \{1, \ldots, n\}$,

$$\sum_{i=1}^{n} u(x_i, y_i) \geq \sum_{i=1}^{n} u(x_i, y_{s(i)}).$$

This is because any permutation can be written as the composition of disjoint cycles.

We now prove the sufficiency of u-cyclical monotonicity by contraposition. Suppose that there exists $\pi' \in \Pi(\mu, \nu)$ such that

$$\sum_{x,y} \pi'(x, y) u(x, y) > \sum_{x,y} \pi^*(x, y) u(x, y). \tag{3}$$

We will show that there exists a sequence $(x_1, y_1), \ldots, (x_n, y_n) \in \text{supp}(\pi^*)$ and a permutation $s : \{1, \ldots n\} \to \{1, \ldots, n\}$, such that

[21] The scaling factor $1/n$ is added to ensure that π^ε is a positive measure, in case the same pair (x, y) appears multiple times in the sequence.

$$\sum_{i=1}^{n} u(x_i, y_{t(i)}) > \sum_{i=1}^{n} u(x_i, y_i).$$

The argument proceeds in three steps.

Step 1.—Approximate π^* and π' with $\tilde{\pi}^*$ and $\tilde{\pi}'$, respectively, while preserving inequality (3): both $\tilde{\pi}^*$ and $\tilde{\pi}'$ are joint distributions that share the same rational marginals; in addition, $\tilde{\pi}^*$ shares the same support as π^*.

Choose $\varepsilon_0 > 0$ so that

$$\sum_{x,y} \pi^*(x, y) u(x, y) < \sum_{x,y} \pi'(x, y) u(x, y) - \varepsilon_0.$$

By continuity, there exists $\delta_1 > 0$ such that for all $|\pi - \pi^*| < \delta_1$, we have

$$\sum_{x,y} \pi(x, y) u(x, y) < \sum_{x,y} \pi'(x, y) u(x, y) - \frac{\varepsilon_0}{2}. \tag{4}$$

By lemma 10 in appendix section B.2.1, there exists $\delta_2 > 0$ such that for all $|\tilde{\mu} - \mu| < \delta_2$ and $|\tilde{\nu} - \nu| < \delta_2$, there exists $\pi \in \Pi(\tilde{\mu}, \tilde{\mu})$ with

$$\sum_{x,y} \pi(x, y) u(x, y) > \sum_{x,y} \pi'(x, y) u(x, y) - \frac{\varepsilon_0}{2}. \tag{5}$$

Let $\delta_3 \equiv \min_{x,y}\{\pi^*(x, y) : \pi^*(x, y) > 0\}$ denote the smallest probability weight among the support of π^*.

Now let $\delta = \min\{\delta_1, \delta_2/(|X| \times |Y|), \delta_3\}$ and consider a rational joint distribution $\tilde{\pi}^* \in Q^{X \times Y} \cap \Delta(X \times Y)$ such that $|\tilde{\pi}^* - \pi^*| < \delta$. Note that $\text{supp}(\tilde{\pi}^*) = \text{supp}(\pi^*)$. By inequality (4),

$$\sum_{x,y} \tilde{\pi}^*(x, y) u(x, y) < \sum_{x,y} \pi'(x, y) u(x, y) - \frac{\varepsilon_0}{2}.$$

Furthermore, the marginals of $\tilde{\pi}^*$, $p \equiv \tilde{\pi}^*_X$ and $q \equiv \tilde{\pi}^*_Y$, are also rational and satisfy $|p - \mu| < \delta_2$ and $|q - \nu| < \delta_2$. By inequality (5), there exists $\tilde{\pi}' \in \Pi(p, q)$ such that

$$\sum_{x,y} \tilde{\pi}'(x, y) u(x, y) > \sum_{x,y} \pi'(x, y) u(x, y) - \frac{\varepsilon_0}{2},$$

so

$$\sum_{x,y} \tilde{\pi}'(x, y) u(x, y) > \sum_{x,y} \tilde{\pi}^*(x, y) u(x, y). \tag{6}$$

Step 2.—Normalize and transform the above two joint distributions with the same rational marginals, $\tilde{\pi}^*$ and $\tilde{\pi}'$, into doubly stochastic matrices. Through the Birkhoff–von Neumann theorem, express inequality (6) in terms of permutation matrices.

Let N be an integer such that $Np(x)$ and $Nq(y)$ are integers for all $x \in X$ and $y \in Y$. Let $S : X \to 2^{\{1, \ldots, N\}}$ be a partition of $\{1, \ldots, N\}$ such that $|S(x)| = Np(x)$ for each $x \in X$; similarly, let $T : Y \to 2^{\{1, \ldots, N\}}$ be a partition of $\{1, \ldots, N\}$ such that

$|T(y)| = Nq(y)$ for each $y \in Y$. For each $i = 1, \ldots, N$, let $\tilde{x}_i \equiv \{x : i \in S(x)\}$ denote the $x \in X$ indexing the partition that contains i; similarly, for each column j, let $\tilde{y}_j \equiv \{y : j \in T(y)\}$ denote the $y \in Y$ indexing the partition that contains j.

Consider the matrix $[B^*_{ij}]_{1 \le i,j \le N}$, defined by

$$B^*_{ij} = \frac{\tilde{\pi}^*(\tilde{x}_i, \tilde{y}_j)}{Np(\tilde{x}_i)q(\tilde{y}_j)} \text{ for all } 1 \le i, j \le N,$$

and the matrix $[B'_{ij}]_{1 \le i,j \le N}$, defined by

$$B'_{ij} = \frac{\tilde{\pi}'(\tilde{x}_i, \tilde{y}_j)}{Np(\tilde{x}_i)q(\tilde{y}_j)} \text{ for all } 1 \le i, j \le N.$$

Notice that the matrix B^* is doubly stochastic: for any i, we have

$$\sum_j B^*_{ij} = \sum_{y \in Y} \left(\frac{\tilde{\pi}^*(\tilde{x}_i, y)}{Np(\tilde{x}_i)q(y)} \cdot Nq(y) \right) = \frac{p(\tilde{x}_i)}{p(\tilde{x}_i)} = 1.$$

Similarly, we can show that $\Sigma_i B^*_{ij} = 1$ for every j. And following similar arguments, the matrix B' is also doubly stochastic.

Note that

$$\sum_{i,j} B^*_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) = \sum_{i,j} \frac{\tilde{\pi}^*(\tilde{x}_i, \tilde{y}_j)}{Np(\tilde{x}_i)q(\tilde{y}_j)} u(\tilde{x}_i, \tilde{y}_j)$$

$$= \sum_{x,y} \frac{\tilde{\pi}^*(x, y)}{Np(x)q(y)} \cdot Np(x) \cdot Nq(y) \cdot u(x, y) = N \sum_{x,y} \tilde{\pi}^*(x, y) u(x, y),$$

and similarly,

$$\sum_{i,j} B'_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) = N \sum_{x,y} \tilde{\pi}'(x, y) u(x, y).$$

Now since

$$\sum_{x,y} \tilde{\pi}'(x, y) u(x, y) > \sum_{x,y} \tilde{\pi}^*(x, y) u(x, y),$$

we have

$$\sum_{i,j} B'_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) > \sum_{i,j} B^*_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j).$$

Let $\mathcal{P}$ denote the set of $N \times N$ permutation matrices. By the Birkhoff–von Neumann theorem, both B^* and B' are in the convex hull of $\mathcal{P}$. It follows that there exist permutation matrices P^* and P' such that

$$\sum_{i,j} P'_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) > \sum_{i,j} P^*_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j), \tag{7}$$

and in addition, $P^*_{ij} = 1$ implies that the corresponding entry in B^* satisfies $B^*_{ij} > 0$.

Step 3.—Convert inequality (7) into a cyclical deviation.

Note that the permutation matrix P^* is equivalent to a mapping $t : \{1, \ldots, N\} \to \{1, \ldots, N\}$ such that $P^*_{ij} = 1$ if and only if $j = t(i)$. So

$$\sum_{i,j} P^*_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) = \sum_i u(\tilde{x}_i, \tilde{y}_{t(i)}).$$

In particular, every element of $\{(\tilde{x}_i, \tilde{y}_{t(i)})\}_{i=1}^N$ is in the support of $\tilde{\pi}^*$, since $P^*_{ij} = 1$ implies $B^*_{ij} > 0$, which further implies $\tilde{\pi}^*(\tilde{x}, \tilde{y}_{t(i)}) > 0$. Since π^* and $\tilde{\pi}^*$ share the same support, every element of $\{(\tilde{x}_i, \tilde{y}_{t(i)})\}_{i=1}^N$ is in the support of π^* as well.

Let $t' : \{1, \dots, N\} \to \{1, \dots, N\}$ denote the permutation mapping induced by the matrix P'_{ij}, so

$$\sum_{i,j} P'_{ij} \cdot u(\tilde{x}_i, \tilde{y}_j) = \sum_i u(\tilde{x}_i, \tilde{y}_{t'(i)}).$$

It follows that inequality (7) can be rewritten as

$$\sum_i u(\tilde{x}_i, \tilde{y}_{t'(i)}) > \sum_i u(\tilde{x}_i, \tilde{y}_{t(i)}). \tag{8}$$

Since t and t' are both permutations, there exists a permutation $s : \{1, \dots, N\} \to \{1, \dots, N\}$ such that $s(t(i)) = t'(i)$. Consider the sequence $\{(x_i, y_i)\}_{i=1}^N$, defined by

$$x_i = \tilde{x}_i \text{ and } y_i = \tilde{y}_{t(i)} \text{ for } i = 1, \dots, N.$$

Then (8) becomes

$$\sum_{i=1}^N u(x_i, y_i) < \sum_{i=1}^N u(x_i, y_{s(i)}),$$

where $(x_i, y_i) \in \text{supp}(\pi^*)$ for each $1 \le i \le N$, which violates u-cyclical monotonicity. QED

A1.2. Proof of Theorem 1

For the "if" direction, suppose that π is u_R-obedient, u_S-cyclically monotone, and satisfies $\pi_\Theta = \mu_0$. The proof is by construction.

Since $\pi_\Theta = \mu_0$, we can construct an information structure (M, λ^*) by setting $M = A$ and $\lambda^* = \pi$; furthermore, let σ^* be the identity map from M to A. It is straightforward to see that the profile (λ^*, σ^*) induces the outcome distribution π. We first show that (λ^*, σ^*) is R-IC. Since π is u_R-obedient, we have that for each $a \in A$,

$$a \in \arg\max_{a'} \sum_\Theta u_R(\theta, a')\pi(\theta, a).$$

Since σ^* is an identity map, it follows that for each $m \in M$,

$$\sigma^*(m) \in \arg\max_{a'} \sum_\Theta u_R(\theta, a')\pi(\theta, \sigma^*(m)).$$

Furthermore, since $\lambda^* = \pi$ and σ^* is injective, we have $\lambda^*(\theta, m) = \pi(\theta, \sigma^*(m))$ for all $\theta \in \Theta$ and $m \in M$. So

$$\sigma^* \in \arg\max_{\sigma : M \to A} \sum_{\Theta \times M} u_R(\theta, \sigma(m))\lambda^*(\theta, m),$$

which means σ^* is a best response to λ^*.

It remains to show that the sender does not benefit from choosing any other information structure in $D(\lambda^*)$. Observe that since π is u_S-cyclically monotone, every sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n)$ in $\text{supp}(\pi)$ where $a_{n+1} \equiv a_1$ satisfies

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) \geq \sum_{i=1}^{n} u_S(\theta_i, a_{i+1}).$$

Since $\lambda^* = \pi$ and σ^* is the identity mapping, this further implies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) \geq \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1}));$$

for every sequence $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in \text{supp}(\lambda^*)$ with $m_{n+1} = m_1$. In addition, $\lambda^*_\theta = \mu_0$ and $\lambda^*_M = \lambda^*_M$ by construction. By lemma 2, λ^* satisfies

$$\lambda^* \in \arg\max_{\lambda \in D(\lambda^*)} \sum_{\Theta \times M} u_S(\theta, \sigma(m))\lambda(\theta, m),$$

which means λ^* is sender optimal conditional on its message distribution.

For the "only if" direction, suppose that π is stable and thus induced by a credible and R-IC profile (λ^*, σ^*). Since σ^* best responds to the messages from λ^*, the u_R-obedience of π follows from Bergemann and Morris (2016).

It remains to show that π is u_S-cyclical monotone. Suppose by contradiction that π is not u_S-cyclically monotone, which implies that there exists a sequence $(\theta_1, a_1), \ldots, (\theta_n, a_n) \in \text{supp}(\pi)$ such that

$$\sum_{i=1}^{n} u_S(\theta_i, a_i) < \sum_{i=1}^{n} u_S(\theta_i, a_{i+1}),$$

where $a_{n+1} = a_1$. Since π is induced by (λ^*, σ^*), for each $i = 1, \ldots, n$ there exists m_i such that $m_i \in \sigma^{*-1}(a_i)$ and $(\theta_i, m_i) \in \text{supp}(\lambda^*)$, so we have a sequence $(\theta_1, m_1), \ldots, (\theta_n, m_n) \in \text{supp}(\lambda^*)$ that satisfies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) < \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1})), \tag{9}$$

where $m_{n+1} = m_1$. Define $v(\theta, m) \equiv u_S(\theta, \sigma^*(m))$. Since (λ^*, σ^*) is credible, we have

$$\lambda^* \in \arg\max_{\lambda \in D(\lambda^*)} \sum_{\Theta \times M} v(\theta, m)\lambda(\theta, m).$$

Lemma 2 implies that λ^* is v-cyclically monotone. Since $(\theta_1, m_1), \ldots, (\theta_n, m_n)$ is in $\text{supp}(\lambda^*)$, the v-cyclical monotonicity of λ^* implies

$$\sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_i)) \geq \sum_{i=1}^{n} u_S(\theta_i, \sigma^*(m_{i+1})),$$

where $m_{n+1} = m_1$, which is a contradiction to (9). So π must be u_S-cyclically monotone. QED

A2. Proof of Proposition 1

The sender-optimal stable outcome distribution is the solution to the following problem:

$$\max_{\pi \in \Delta(\Theta \times A)} \sum_{\theta, a} \pi(\theta, a) u_S(\theta, a)$$

$$\text{subject to } \sum_{\theta} \pi(\theta|a) u_R(\theta, a) \geq \sum_{\theta} \pi(\theta|a) u_R(\theta, a') \text{ for all } a \in \text{supp}(\pi_A) \text{ and } a' \in A,$$

π is u_S-cyclically monotone,

$\pi_\Theta = \mu_0$.

We first argue that the feasible region in the optimization program above is compact, so there exists a sender-optimal stable outcome distribution. The obedience constraints are weak inequalities, so they define a compact set of outcome distributions. It suffices to establish the compactness of the set of u_S-cyclically monotone outcome distributions, denoted by $\Pi^{cyc} \equiv \{\pi \in \Delta(\Theta \times A) : \pi \text{ is } u_S\text{-cyclically monotone}\}$.

Let $\mathcal{O} \equiv \{\text{supp}(\pi) : \pi \in \Pi^{cyc}\}$ denote the set of the supports of the distributions in Π^{cyc}. For each such support $O \in \mathcal{O}$, let $\Pi_O \equiv \{\pi \in \Delta(\Theta \times A) : \text{supp}(\pi) \subseteq O\}$ denote the set of outcome distributions whose support is contained within O. Note that since reducing the support of the outcome distribution relaxes the u_S-cyclical monotonicity constraint, every distribution in the set Π_O is u_S-cyclically monotone, so we have $\Pi^{cyc} = \cup_{O \in \mathcal{O}} \Pi_O$. In addition, for each $O \in \mathcal{O}$, the set Π_O is closed since it is defined by equality constraints: $\Pi_O = \{\pi \in \Delta(\Theta \times A) : \pi(\theta, a) = 0 \ \forall \ (\theta, a) \notin O\}$. The set $\Pi^{cyc} = \cup_{O \in \mathcal{O}} \Pi_O$ is a finite union of closed and bounded sets and is therefore compact, so there exists a sender-optimal stable outcome distribution.

Next, we show that there exists a sender-optimal credible and R-IC profile that does not involve more than $\min\{|\Theta|, |A|\}$ messages. Let π^* denote a sender-optimal stable outcome distribution.

To establish the $|\Theta|$ bound, let $v^* \equiv \Sigma_a \pi^*_A(a) \Sigma_\theta \pi^*(\theta|a) u_S(\theta, a)$ denote the sender's value from π^* and $A^* \equiv \text{supp}(\pi^*_A)$ denote the support of π^*'s action distribution. Consider the following set

$$\mathcal{E} \equiv \left\{ \left(\pi^*(\cdot|a), \sum_{\theta} u_S(\theta, a) \pi^*(\theta|a) \right) \in \mathbb{R}^{|\Theta|} | a \in A^* \right\},$$

where $\pi^*(\cdot|a) \in \Delta(\Theta)$ and $\Sigma_\theta u_S(\theta, a) \pi^*(\theta|a) \in \mathbb{R}$ denote the posterior belief and sender's value conditioning on a, respectively. We will use (μ_a, v_a) to denote an element of $\mathcal{E}$.

Recall that $\mu_0 \in \Delta(\Theta)$ is the prior distribution over states. Clearly $(\mu_0, v^*) \in \text{conv}(\mathcal{E}) \subset \mathbb{R}^{|\Theta|}$. We will show that (μ_0, v^*) can be represented as a convex combination of at most $|\Theta|$ number of points in $\mathcal{E}$. To this end, we argue that (μ_0, v^*) must be a boundary point of $\text{conv}(\mathcal{E})$. Suppose not; then there exists $\hat{p} \in \Delta(A^*)$ and $\{(\mu_a, v_a)\}_{a \in A^*} \in \mathcal{E}$ such that $\Sigma_{a \in A^*} (\mu_a, v_a) \hat{p}(a) = (\mu_0, \hat{v})$, where $\hat{v} > v^*$. Let $\hat{\pi} \in \Delta(\Theta \times A)$ be the outcome distribution induced by $\hat{p}$: that is,

$$\hat{\pi}(\theta, a) = \begin{cases} \hat{p}(a) \pi^*(\theta|a) & \text{for all } a \in A^* \text{ and } \theta \in \Theta, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\hat{\pi}$ satisfies obedience. It also satisfies cyclical monotonicity because the support of $\hat{\pi}$ is a subset of the support of π^*, and smaller support means that there are fewer inequalities to check in the cyclical monotonicity condition. Moreover,

$\hat{\pi}$ yields a strictly higher value to the sender, which contradicts π^* being the sender-optimal stable outcome distribution. Therefore, (μ_0, v^*) is on the boundary of $\text{conv}(\mathcal{E})$.

By the supporting hyperplane theorem, there exists a supporting hyperplane of $\text{conv}(\mathcal{E})$ containing (μ_0, v^*); that is, there exists a vector $r \in \mathbb{R}^{|\Theta|}$ and scalar $w \in \mathbb{R}$ such that $r \cdot (\mu_0, v^*) = w$ and $r \cdot (\mu, v) \geq w$ for all $(\mu, v) \in \text{conv}(\mathcal{E})$. In particular, $r \cdot (\mu_a, v_a) \geq w$ for all $a \in A^*$. Recall that $(\mu_0, v^*) = \Sigma_{a \in A^*} \pi^*_A(a)(\mu_a, v_a)$. Since π^*_A has full support on A^*, this implies $r \cdot (\mu_a, v_a) = w$ for all $a \in A^*$, so all points in $\mathcal{E}$ lie on the same hyperplane, which has dimension $|\Theta| - 1$. By Carathéodory's theorem, (μ_0, v^*) can be represented as a convex combination of at most $|\Theta|$ number of points in $\mathcal{E}$ according to some mixture probabilities $\tilde{p} \in \Delta(A^*)$.

Let $\tilde{\pi}$ be the outcome distribution induced by $\tilde{p}$ (in particular, it is obtained from the same construction as that for $\hat{\pi}$ above but with $\tilde{p}$ replacing $\hat{p}$). By construction, $|\text{supp}(\tilde{\pi}_A)| \leq \min\{|\Theta|, |A|\}$, and the sender's value from $\tilde{p}$ is also v^*. In addition, $\tilde{\pi}$ clearly satisfies obedience; it also satisfies cyclical monotonicity because the support of $\tilde{\pi}$ is a subset of π^*, which relaxes the cyclical monotonicity constraint. Therefore, $\tilde{\pi}$ is a sender-optimal stable outcome distribution. Now by using $\tilde{\pi}$ as a direct recommendation information structure, we can derive a sender-optimal credible and R-IC profile that uses no more than $\min\{|\Theta|, |A|\}$ messages, following the same construction outlined in the "if" direction of the proof of theorem 1.

A3. Proof of Lemma 1

In light of remark 1, we shall prove lemma 1 without assuming that the order on either Θ or A is antisymmetric. Suppose $(\Theta, \succcurlyeq_1)$ and $(A, \succcurlyeq_2)$ are finite ordered sets and $\succcurlyeq_1$, $\succcurlyeq_2$ are weak orders (complete and transitive). The notions of supermodularity and comonotonicity are extended naturally with weak orders $\succcurlyeq_1$ and $\succcurlyeq_2$ replacing the total orders on Θ and A.

In particular, we say a function $u : \Theta \times A \to \mathbb{R}$ is supermodular if for any $\theta \succcurlyeq \theta'$ and $a \succcurlyeq a'$, we have

$$u(\theta, a) + u(\theta', a') \geq u(\theta, a') + u(\theta', a);$$

the function is strictly supermodular if in additional for any $\theta \succ \theta'$ and $a \succ a'$,

$$u(\theta, a) + u(\theta', a') > u(\theta, a') + u(\theta', a).$$

An outcome distribution π is comonotone if for any $(\theta, a), (\theta', a') \in \text{supp}(\pi)$, $\theta \succ \theta'$ implies $a \succcurlyeq a'$. We shall prove the following result, which is a restatement of lemma 1 but based on the weak orders $\succcurlyeq_1$ and $\succcurlyeq_2$.

LEMMA 1*. If u_S is supermodular, then every comonotone outcome distribution is u_S-cyclically monotone. Furthermore, if u_S is strictly supermodular, then every u_S-cyclically monotone outcome distribution is also comonotone.

We begin the proof by establishing the following lemma.

LEMMA 3. Let $t : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a bijection. Suppose that t is not the identity mapping; then there exists k^* such that $t(k^*) > k^*$ and $t(t(k^*)) < t(k^*)$.

Proof. Define $K := \{k \in \{1, \dots, n\} : t(k) \neq k\}$. Since t is not the identity mapping, K is nonempty. Since t is a bijection, $t(k) \neq k$ if and only if $t(t(k)) \neq t(k)$, so K is t-invariant. Let $k^* = t^{-1}(\max K) \in K$, then $k^* < \max K = t(k^*)$ and $t(k^*) = \max K > t(\max K) = t(t(k^*))$. QED

Proof of lemma 1.* First, we show that comonotonicity implies u_S-cyclical monotonicity when u_S is supermodular. Suppose that an outcome distribution $\pi \in \Delta(\Theta \times A)$ is comonotone, then the product order of $\succcurlyeq_1$ and $\succcurlyeq_2$ is also a weak order on $\text{supp}(\pi)$. Take any sequence $(\theta_1, a_1), \dots, (\theta_n, a_n) \in \text{supp}(\pi)$ and assume without loss of generality that (θ_i, a_i) is nondecreasing in $i \in \{1, \dots, n\}$ with respect to the product order. We will show that for any permutation $t : \{1, \dots, n\} \to \{1, \dots, n\}$,

$$u_S(\theta_1, a_1) + \cdots + u_S(\theta_n, a_n) \geq u_S(\theta_1, a_{t(1)}) + \cdots + u_S(\theta_n, a_{t(n)}),$$

which then proves the statement. In particular, for each permutation t, let $v(t) \equiv u_S(\theta_1, a_{t(1)}) + \cdots + u_S(\theta_n, a_{t(n)})$ denote the value obtained from summing u_S according to the state-action pairings in t and let I denote the identity map. We show that $v(I) \geq v(t)$ for every permutation t.

To this end, take any permutation t that is not an identity mapping, and let $l(t)$ denote the number of fixed points of t (which may be zero). By lemma 3, there exists k^* such that $t(k^*) > k^*$ and $t(t(k^*)) < t(k^*)$. The supermodularity of u_S implies

$$u_S(\theta_{t(k^*)}, a_{t(k^*)}) + u_S(\theta_{k^*}, a_{t(t(k^*))}) \geq u_S(\theta_{k^*}, a_{t(k^*)}) + u_S(\theta_{t(k^*)}, a_{t(t(k^*))}). \tag{10}$$

Define a new permutation $\hat{t}$ so that k is mapped to $t(t(k))$ while $t(k)$ is mapped to $t(k)$, while all other pairings remain unchanged. Formally,

$$\hat{t}(k) = \begin{cases} t(k) & \text{for all } k \neq k^*, t(k^*), \\ t(t(k^*)) & \text{if } k = k^*, \\ t(k^*) & \text{if } k = t(k^*). \end{cases}$$

By (10), we have

$$u_S(\theta_1, a_{\hat{t}(1)}) + \cdots + u_S(\theta_n, a_{\hat{t}(n)}) \geq u_S(\theta_1, a_{t(1)}) + \cdots + u_S(\theta_n, a_{t(n)}),$$

so we have constructed another permutation $\hat{t}$ with $v(\hat{t}) \geq v(t)$ and $l(\hat{t}) = l(t) + 1$. Each time we iterate the process above, $v(\cdot)$ weakly increases while the number of fixed points increases by 1. Since $n < \infty$, the iteration terminates at the identity map I, so $v(I) \geq v(t)$ for every permutation t.

Next, suppose that u_S is strictly supermodular. We want to show that u_S-cyclical monotonicity implies comonotonicity. We prove this statement by contraposition: suppose that an outcome distribution π is not comonotone; we will show that π is not u_S-cyclically monotone. Since π is not comonotone, there exists $(\theta, a), (\theta', a') \in \text{supp}(\pi)$ such that $\theta \prec \theta'$, $a \succ a'$. Since u_S is strictly supermodular, we have

$$u_S(\theta, a) + u_S(\theta', a') < u_S(\theta, a') + u_S(\theta', a). \tag{11}$$

Consider a cycle of length 2 where $(\theta_1, a_1) = (\theta, a)$ and $(\theta_2, a_2) = (\theta', a')$; then inequality (11) above implies that π is not u_S-cyclically monotone. QED

A4. Proof of Proposition 2

Let π be a stable outcome distribution, and suppose by contradiction that there exist two distinct actions $a_1, a_2 \in \text{supp}(\pi_A)$, say, $a_1 < a_2$. Let $I_1 \equiv \{\theta \in \Theta | \pi(\theta, a_1) > 0\}$ and $I_2 \equiv \{\theta \in \Theta | \pi(\theta, a_2) > 0\}$ be the states associated with a_1 and a_2, respectively, in the support of π. By theorem 1, since π is stable, it must be u_R-obedient, which implies

$$\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_A(a_1)} \geq 0 \geq \sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_A(a_2)}. \quad (12)$$

Furthermore, since u_S is strictly supermodular, π is also comonotone by theorem 1 and lemma 1, so any $\theta \in I_1$ and $\theta' \in I_2$ satisfies $\theta \leq \theta'$. Since u_R is submodular, we have $u_R(\theta, a_1) - u_R(\theta, a_2) \leq u_R(\theta', a_1) - u_R(\theta', a_2)$ for all $\theta \in I_1$ and $\theta' \in I_2$, which implies

$$\max_{\theta \in I_1} \{u_R(\theta, a_1) - u_R(\theta, a_2)\} \leq \min_{\theta' \in I_2} \{u_R(\theta', a_1) - u_R(\theta', a_2)\}.$$

So

$$\begin{aligned} \sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_A(a_1)} &\leq \max_{\theta \in I_1} \{u_R(\theta, a_1) - u_R(\theta, a_2)\} \\ &\leq \min_{\theta' \in I_2} \{u_R(\theta', a_1) - u_R(\theta', a_2)\} \\ &\leq \sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_A(a_2)}. \end{aligned} \quad (13)$$

Combining (12) and (13), we have

$$\sum_{\theta \in I_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\pi(\theta, a_1)}{\pi_A(a_1)} = \max_{\theta \in I_1} \{u_R(\theta, a_1) - u_R(\theta, a_2)\} = 0$$

and

$$\sum_{\theta' \in I_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\pi(\theta', a_2)}{\pi_A(a_2)} = \min_{\theta' \in I_2} \{u_R(\theta', a_1) - u_R(\theta', a_2)\} = 0.$$

So $u_R(\theta, a_1) = u_R(\theta, a_2)$ for all $\theta \in I_1 \cup I_2$.

Since the argument above applies to any $a_1, a_2 \in \text{supp}(\pi_A)$, we have that for all $a_i, a, a' \in \text{supp}(\pi_A)$,

$$u_R(\theta, a_i) - u_R(\theta, a) = u_R(\theta, a_i) - u_R(\theta, a') = 0 \ \forall \ \theta \in I_i,$$

so for all i and $\theta \in I_i$, we have

$$u_R(\theta, a) - u_R(\theta, a') = 0 \ \forall \ a, a' \in \text{supp}(\pi_A),$$

and therefore

$$u_R(\theta, a) - u_R(\theta, a') = 0 \ \forall \ a, a' \in \text{supp}(\pi_A) \text{ and } \theta \in \Theta.$$

However, this is a contradiction since by assumption, there exists no $a, a' \in A$ such that $a \neq a'$ and $u_R(\theta, a) = u_R(\theta, a')$ for all θ.

Therefore, $\text{supp}(\pi_A)$ must be a singleton, denoted by a^*. Then u_R-obedience implies $a^* \in \arg\max_{a \in A} \Sigma_\theta \mu_0(\theta) u(\theta, a)$. So π is a no-information outcome.

A5. *Proof of Proposition 3*

Proof of statement 1. For each $a \in A$, let

$$P_a \equiv \{\mu \in \Delta(\Theta) | \sum_\theta \mu(\theta) u_R(\theta, a) > \sum_\theta \mu(\theta) u_R(\theta, a'), \forall\ a' \neq a\},$$

which denotes the set of beliefs such that a is the receiver's strict best response. We prove our claim under the assumption that there exists $a^\circ \in A$ such that $\mu_0 \in P_{a^\circ}$ (i.e., a° is the unique best response to μ_0). Later we will show that this assumption holds for generic priors.

When the sender's information structure is uninformative, the receiver best responds to the sender's messages by choosing a°. The sender's payoff is

$$v_0 \equiv \sum_{\theta \in \Theta} \mu_0(\theta) u_S(\theta, a^\circ).$$

We will show that there exists a stable outcome distribution that gives the sender a higher payoff than v_0.

We consider the case where the sender benefits from persuasion, so $a^\circ \neq \bar{a}$; otherwise, the receiver is already choosing the sender's favorite action under the prior. For ε sufficiently small, consider the outcome distribution $\pi^\varepsilon \in \Delta(\Theta \times A)$, defined by

$$\pi^\varepsilon(\theta, a) = \begin{cases} \mu_0(\theta) & \text{if } \theta \neq \bar{\theta}, a = a^\circ, \\ \mu_0(\bar{\theta}) - \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, a^\circ), \\ \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, \bar{a}), \\ 0 & \text{otherwise.} \end{cases}$$

We will show that for ε sufficiently small, π^ε is stable and gives the sender a higher payoff than v_0.

It can be easily seen that the support of π^ε is comonotone. Since u_S is supermodular, π^ε is u_S-cyclically monotone by lemma 1. Next, we verify that for ε sufficiently small, π^ε satisfies u_R-obedience at the two actions $\{\bar{a}, a^\circ\}$. For a°, note that since $\mu_0 \in P_{a^\circ}$, we have

$$\sum_{\theta \in \Theta} \mu_0(\theta) u(\theta, a^\circ) > \sum_{\theta \in \Theta} \mu_0(\theta) \pi(\theta, a') \text{ for all } a' \in A,$$

so for ε sufficiently small,

$$\sum_{\theta \in \Theta} \mu_0(\theta) u(\theta, a^\circ) - \varepsilon u(\bar{\theta}, a^\circ) \geq \sum_{\theta \in \Theta} \mu_0(\theta) \pi(\theta, a') - \varepsilon u(\bar{\theta}, a') \text{ for all } a' \in A,$$

which means π^ε satisfies u_R-obedience at a°.

As $\bar{a} \in A^\circ$, there exists $\bar{\mu} \in \Delta(\Theta)$ such that $\bar{a} \in \arg\max_a \Sigma_\theta \bar{\mu}(\theta) u_R(\theta, a)$. So for every $a' \neq \bar{a}$,

$$\sum_\theta \bar{\mu}(\theta)[u_R(\theta, \bar{a}) - u_R(\theta, a')] \geq 0.$$

Since u_R is supermodular, $u_R(\theta, \bar{a}) - u_R(\theta, a')$ is weakly increasing in θ, so if a belief μ' first-order stochastically dominates $\bar{\mu}$, then

$$\sum_{\theta} \mu'(\theta)[u_R(\theta, \bar{a}) - u(\theta, a')] \geq \sum_{\theta} \bar{\mu}(\theta)[u_R(\theta, \bar{a}) - u(\theta, a')] \geq 0 \text{ for all } a' \neq \bar{a}.$$

In particular, the Dirac measure $\delta_{\bar{\theta}}$ first-order stochastically dominates $\bar{\mu}$, so the inequality above implies

$$u_R(\bar{\theta}, \bar{a}) - u_R(\bar{\theta}, a') \geq 0 \text{ for all } a' \neq \bar{a}.$$

So $\bar{a} \in \arg\max_a u_R(\bar{\theta}, a)$, and π^ε is u_R-obedient at action $\bar{a}$.

Finally, we show that the sender obtains higher payoff from π^ε than v_0. Note that since by our assumption, $u_S(\bar{\theta}, a') < u_S(\bar{\theta}, \bar{a})$ for all $a' \neq \bar{a}$, we have

$$\begin{aligned}
\sum_{\theta, a} \pi^\varepsilon(\theta, a) u_S(\theta, a) &= \sum_{\theta \neq \bar{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\bar{\theta}) - \varepsilon) u_S(\bar{\theta}, a^\circ) + \varepsilon u_S(\bar{\theta}, \bar{a}) \\
&> \sum_{\theta \neq \bar{\theta}} \mu_0(\theta) u_S(\theta, a^\circ) + (\mu_0(\bar{\theta}) - \varepsilon) u_S(\bar{\theta}, a^\circ) + \varepsilon u_S(\bar{\theta}, a^\circ) \\
&= \sum_{\theta} \mu_0(\theta) u_S(\theta, a^\circ) = v_0.
\end{aligned}$$

Therefore, the sender receives a strictly higher payoff from π^ε than v_0. This completes the proof.

The rest of the proof shows that $\cup_{a \in A} P_a$ contains an open, dense, and full measure set. In particular, we show that $\Delta(\Theta)/\{\cup_{a \in A} P_a\}$ is included in a negligible, closed, and nowhere dense subset in $\Delta(\Theta)$.

Define $H_{a,a'} \equiv \{\mu \in \Delta(\Theta) | \Sigma_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0\}$ for any $a \neq a'$. Since by assumption 1, $u_R(\cdot, a) - u_R(\cdot, a') \neq \mathbf{0}$, which implies $J_{a,a'} \equiv \{\mu \in \mathbb{R}^{|\Theta|} | \Sigma_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0\}$ is a hyperplane in $\mathbb{R}^{|\Theta|}$. Notice that $H_{a,a'} = J_{a,a'} \cap \Delta(\Theta)$, which is the intersection of a hyperplane with a simplex. Since the hyperplane includes $\mathbf{0}$ and $\Delta(\Theta)$ does not, either they have to be parallel with no intersection or their intersection is in a lower-dimensional subspace, which is negligible, closed, and nowhere dense.

For any $\mu \in \Delta(\Theta)/\{\cup_{a \in A} P_a\}$, since the maximizer of $\Sigma_\theta \mu(\theta) u_R(\theta, a)$ is not unique, there exists a, a' such that $\Sigma_\theta \mu(\theta)(u_R(\theta, a) - u_R(\theta, a')) = 0$. So $\Delta(\Theta)/\{\cup_{a \in A} P_a\} \subseteq \cup_{a \neq a'} H_{a,a'}$, where the latter is a negligible, closed, and nowhere dense subset of $\Delta(\Theta)$. QED

Proof of statement 2. For any generic prior $\mu^\circ \in \cup_{a \in A} P_a$, either $\mu^\circ \notin P_{\underline{a}}$ or $\mu^\circ \notin P_{\underline{a}}$. We consider the case $\mu^\circ \notin P_{\underline{a}}$, and the other case can be shown symmetrically. Similar to the previous argument, for ε sufficiently small, consider the outcome distribution $\pi^\varepsilon \in \Delta(\Theta \times A)$:

$$\pi^\varepsilon(\theta, a) = \begin{cases} \mu_0(\theta) & \text{if } \theta \neq \bar{\theta}, a = a^\circ, \\ \mu_0(\bar{\theta}) - \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, a^\circ), \\ \varepsilon & \text{if } (\theta, a) = (\bar{\theta}, \bar{a}), \\ 0 & \text{otherwise.} \end{cases}$$

As we have shown in the proof of statement 1, for ε sufficiently small, π^ε is stable and gives the sender a higher payoff than v_0. Therefore, the sender benefits from credible persuasion. QED

Proof of statement 3. Let Π_F denote the set of fully revealing outcome distributions, which is compact because it is a closed subset of $\Delta(\Theta \times A)$. Let

$$\Pi_F^* \equiv \arg\max_{\pi \in \Pi_F} \sum_{\theta,a} \pi(\theta, a) u_S(\theta, a)$$

be the subset of Π_F that maximizes sender's payoff, which is also compact by Berge's theorem of maximum. Note that by definition, every fully revealing outcome distribution is obedient. We will show that there exists an outcome distribution $\pi^* \in \Pi_F^*$ that is comonotone. This implies that as long as the sender benefits from one fully revealing outcome distribution, she must also benefit from π^*, which is a comonotone (and obedient) fully revealing outcome distribution. This will then complete our proof following theorem 1 and lemma 1.

To this end, let us choose

$$\pi^* \in \arg\max_{\pi \in \Pi_F^*} \sum_{\theta,a} \pi(\theta, a) \theta a.$$

Suppose by contradiction that π^* is not comonotone; we construct another outcome distribution $\pi' \in \Pi_F^*$ that satisfies $\Sigma_{\theta,a} \pi'(\theta, a) \theta a > \Sigma_{\theta,a} \pi^*(\theta, a) \theta a$, which contradicts $\pi^* \in \arg\max_{\pi \in \Pi_F^*} \Sigma_{\theta,a} \pi(\theta, a) \theta a$.

Since π^* is not comonotone, there exists a pair (θ_1, a_1), (θ_2, a_2) in the support of π^* such that $\theta_1 < \theta_2$ and $a_1 > a_2$. Take $\varepsilon = \min\{\pi^*(\theta_1, a_1), \pi^*(\theta_2, a_2)\}$, and construct the outcome distribution π' where

- $\pi'(\theta_1, a_1) = \pi^*(\theta_1, a_1) - \varepsilon$, $\pi'(\theta_2, a_2) = \pi^*(\theta_2, a_2) - \varepsilon$;
- $\pi'(\theta_1, a_2) = \pi^*(\theta_1, a_2) + \varepsilon$, $\pi'(\theta_2, a_1) = \pi^*(\theta_2, a_1) + \varepsilon$; and
- $\pi'(\theta, a) = \pi^*(\theta, a)$ for all other (θ, a).

We first argue that $\pi' \in \Pi_F^*$. Let $A^*(\theta) \equiv \arg\max_{a \in A} u_R(\theta, a)$ denote the receiver's best response correspondence. Since $u_R(\theta, a)$ is supermodular, by lemma 2.8.1 of Topkis (2011), $A^*(\theta)$ is weakly increasing in θ in the induced set order. That is, for any $\theta > \theta'$, $a \in A^*(\theta)$, and $a' \in A^*(\theta')$, we have $\max\{a, a'\} \in A^*(\theta)$ and $\min\{a, a'\} \in A^*(\theta')$. Since $a_1 \in A^*(\theta_1)$ and $a_2 \in A^*(\theta_2)$, we have $a_1 \in A^*(\theta_2)$ and $a_2 \in A^*(\theta_1)$. Therefore, π' is also a fully revealing outcome distribution. Moreover, since u_S is supermodular,

$$\begin{aligned}\sum_{\theta,a} [\pi'(\theta, a) - \pi^*(\theta, a)] u_S(\theta, a) &= \varepsilon[u_S(\theta_1, a_2) + u_S(\theta_2, a_1) - u_S(\theta_1, a_1) - u_S(\theta_2, a_2)] \\ &\geq 0,\end{aligned}$$

so the sender's payoff from π' is weakly greater than from π^*, and therefore $\pi' \in \Pi_F^*$.

Next, we argue that $\Sigma_{\theta,a} \pi'(\theta, a) \theta a > \Sigma_{\theta,a} \pi^*(\theta, a) \theta a$. To this end, note that

$$\begin{aligned}\sum_{\theta,a} [\pi'(\theta, a) - \pi^*(\theta, a)] \theta a &= \varepsilon[\theta_1 a_2 + \theta_2 a_1 - \theta_1 a_1 - \theta_2 a_2] \\ &= (\theta_2 - \theta_1)(a_1 - a_2) > 0.\end{aligned}$$

This contradicts $\pi^* \in \arg\max_{\pi \in \Pi_F^*} \Sigma_{\theta,a} \pi(\theta, a) \theta a$. QED

A6. Proof of Proposition 4

From theorem 1 of Mensch (2021), if both u_S and u_R are supermodular and $|A| = 2$, there exists a full-commitment optimal outcome distribution that is comonotone. Then by theorem 1 and lemma 1, such an outcome distribution is stable. Moreover, if in addition u_S is strictly supermodular, any full-commitment optimal outcome distribution is comonotone. So any full-commitment optimal outcome distribution is stable.

A7. Proof of Proposition 5

We begin by establishing a lemma that will be useful for proving proposition 5.

LEMMA 4. Suppose that the message space M is a finite subset of $\mathbb{R}$, the information structure $\lambda \in \Delta(\Theta \times M)$ is comonotone, and the receiver's payoff function u_R is strictly supermodular. Consider a receiver strategy $\sigma : M \to A$ defined by

$$\sigma(m) \in \arg\max_{a \in A} \sum_{\theta} \lambda(\theta, m) u_R(\theta, a).$$

The outcome distribution $\pi \in \Delta(\Theta \times A)$ induced by (λ, σ) is comonotone and u_R-obedient.

Proof. The fact that π is u_R-obedient follows from Bergemann and Morris (2016). We will prove that π is comonotone. Suppose by contradiction that π is not comonotone, so there exists $(\theta_1, a_1), (\theta_2, a_2) \in \text{supp}(\pi)$, such that $a_1 > a_2$ and $\theta_1 < \theta_2$. We will show that this leads to a contradiction.

Let $M_1 = \{m \in M : \lambda(\theta_1, m) > 0\}$ and $M_2 = \{m \in M : \lambda(\theta_2, m) > 0\}$. Since $(\theta_1, a_1), (\theta_2, a_2) \in \text{supp}(\pi)$, there exists $m_1 \in M_1$ and $m_2 \in M_2$ such that $\sigma(m_1) = a_1$ and $\sigma(m_2) = a_2$. In additional, $m_1 \le m_2$ because $\theta_1 < \theta_2$ and λ is comonotone; furthermore, $m_1 \ne m_2$ because $\sigma(m_1) \ne \sigma(m_2)$, so $m_1 < m_2$.

Let $\Theta_1 = \{\theta \in \Theta : \lambda(\theta, m_1) > 0\}$ and $\Theta_2 = \{\theta \in \Theta : \lambda(\theta, m_2) > 0\}$. Since σ best responds to each message, we have

$$\sum_{\theta \in \Theta_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\lambda(\theta, m_1)}{\lambda_M(m_1)} \ge 0 \ge \sum_{\theta' \in \Theta_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\lambda(\theta', m_2)}{\lambda_M(m_2)}. \tag{14}$$

Furthermore, since λ is comonotone and $m_1 < m_2$, for any $\theta \in \Theta_1$ and $\theta' \in \Theta_2$, $\theta \le \theta'$, which implies $\max \Theta_1 \le \min \Theta_2$. Together with the supermodularity of u_R, we have

$$\max_{\theta \in \Theta_1} \{u_R(\theta, a_1) - u_R(\theta, a_2)\} \le \min_{\theta' \in \Theta_2} \{u_R(\theta', a_1) - u_R(\theta', a_2)\}.$$

So

$$\begin{aligned} \sum_{\theta \in \Theta_1} [u_R(\theta, a_1) - u_R(\theta, a_2)] \frac{\lambda(\theta, m_1)}{\lambda_M(m_1)} &\le \max_{\theta \in \Theta_1} \{u_R(\theta, a_1) - u_R(\theta, a_2)\} \\ &\le \min_{\theta' \in \Theta_2} \{u_R(\theta', a_1) - u_R(\theta', a_2)\} \\ &\le \sum_{\theta' \in \Theta_2} [u_R(\theta', a_1) - u_R(\theta', a_2)] \frac{\lambda(\theta', m_2)}{\lambda_M(m_2)}. \end{aligned} \tag{15}$$

Combining (14) and (15), we have

$$\sum_{\theta\in\Theta_1}[u_R(\theta, a_1) - u_R(\theta, a_2)]\frac{\lambda(\theta, m_1)}{\lambda_M(m_1)} = \max_{\theta\in\Theta_1}\{u_R(\theta, a_1) - u_R(\theta, a_2)\} = 0$$

and

$$\sum_{\theta'\in\Theta_2}[u_R(\theta', a_1) - u_R(\theta', a_2)]\frac{\lambda(\theta', m_2)}{\lambda_M(m_2)} = \min_{\theta'\in\Theta_2}\{u_R(\theta', a_1) - u_R(\theta', a_2)\} = 0,$$

so $u_R(\theta, a_1) = u_R(\theta, a_2)$ for all $\theta \in \Theta_1 \cup \Theta_2$.

But recall that $\theta_1, \theta_2 \in \Theta_1 \cup \Theta_2$ and $\theta_1 < \theta_2$, and from the strict supermodularity of u_R,

$$u_R(\theta_1, a_1) - u_R(\theta_1, a_2) < u_R(\theta_2, a_1) - u_R(\theta_2, a_2),$$

which leads to a contradiction. QED

Proof of proposition 5. Let π be a sender-optimal stable outcome distribution under preferences (u_S, u_R). By theorem 1 and lemma 1, π is comonotone.

Now under the more aligned preferences (u_S, u'_R), suppose that the sender uses the information structure $\lambda = \pi$ with message space $M = \text{supp}(\pi_A)$, and let σ' be the receiver strategy that best responds to each message from π, with ties broken in favor of the sender. By lemma 4, the outcome distribution π' induced by the profile (π, σ') is comonotone and u'_R-obedient. By theorem 1 and lemma 1, π' is a stable outcome distribution under preferences (u_S, u'_R).

It remains to show that the sender obtains a higher payoff from π'. For each belief $\mu \in \Delta(\Theta)$,

$$\hat{a}(\mu) \in \arg\max_{a\in A}\sum_{\theta}\mu(\theta)u_R(\theta, a) \text{ and } \hat{a}'(\mu) \in \arg\max_{a\in A}\sum_{\theta}\mu(\theta)u'_R(\theta, a)$$

denote the receiver's best response to belief μ, with ties broken in favor of the sender. Note that since σ' breaks ties in favor of the sender,

$$E_{\pi(\cdot|a)}[u_S(\theta, \sigma(a))] = E_{\pi(\cdot|a)}[u_S(\theta, \hat{a}'(\pi(\cdot|a)))] \text{ for all } a \in M. \tag{16}$$

By contrast,

$$E_{\pi(\cdot|a)}[u_S(\theta, a)] \le E_{\pi(\cdot|a)}[u_S(\theta, \hat{a}(\pi(\cdot|a)))] \text{ for all } a \in M \tag{17}$$

since π may not be the result of a sender-favoring tie-breaking strategy.

So

$$\begin{aligned}
E_{\pi'}[u_S(\theta, a)] &= E_{\pi}[u_S(\theta, \sigma(a))] \\
&= E_{\pi_A}[E_{\pi(\cdot|a)}[u_S(\theta, \sigma(a))]] \\
&= E_{\pi_A}[E_{\pi(\cdot|a)}[u_S(\theta, \hat{a}'(\pi(\cdot|a)))]] \\
&\ge E_{\pi_A}[E_{\pi(\cdot|a)}[u_S(\theta, \hat{a}(\pi(\cdot|a)))]] \\
&\ge E_{\pi_A}[E_{\pi(\cdot|a)}[u_S(\theta, a)]] \\
&= E_{\pi}[u_S(\theta, a)],
\end{aligned}$$

where the first line above follows from the definition of π', the second line is the law of iterated expectation, the third follows from (16), the fourth line follows from the preferences (u_S, u'_R) being more aligned than (u_S, u_R), the fifth follows from (17), and the last equality is again the law of iterated expectation.

So the sender obtains a higher payoff from π' than π. Since π' is also stable under preferences (u_S, u'_R), this completes our proof. QED

A8. Proof of Proposition 6

For each buyer's belief over quality, $\mu \in \Delta(\Theta)$, let $\underline{\theta}_\mu$ denote the smallest θ in the support of μ; in addition, let $\phi_\mu(x) \equiv E_\mu[v(\theta)|\theta \leq x]$ denote the corresponding expected value to buyers when the quality threshold is $\theta \leq x$.[22] Clearly, $\phi_\mu(\cdot)$ is weakly increasing and $\phi_\mu(1) = E_\mu[v(\theta)]$.

LEMMA 5. For every $\mu \in \Delta(\Theta)$, there exists a largest fixed point $\theta^*_\mu \in (\underline{\theta}_\mu, 1)$ such that $\phi_\mu(\theta^*_\mu) = \theta^*_\mu$. Moreover, for any $\theta \in (\theta^*_\mu, 1]$, $\phi_\mu(\theta) < \theta$.

Proof. Since $\phi_\mu(\underline{\theta}_\mu) = v(\underline{\theta}_\mu) > \underline{\theta}_\mu$, $\phi_\mu(1) = E_\mu[v(\theta)] < 1$, and $\phi_\mu(\cdot)$ is weakly increasing, from Tarski's fixed point theorem, there exists a largest fixed point $\theta^*_\mu \in (\underline{\theta}_\mu, 1)$ such that $\phi_\mu(\theta^*_\mu) = \theta^*_\mu$. To see the second statement, suppose that there exists $\theta \in (\theta^*_\mu, 1)$ such that $\phi_\mu(\theta) \geq \theta$; again from Tarski's fixed point theorem, there exists a fixed point $\theta' \in (\theta^*_\mu, 1)$, which contradicts θ^*_μ being the largest fixed point. QED

LEMMA 6. Let $\lambda \in \Delta(\Theta \times M)$ be an information structure, and for every $m \in M$, let $\mu_m \in \Delta(\Theta)$ denote the buyers' posterior belief after observing message m. The following strategy profile is a Bayesian Nash equilibrium in the game $\langle G, \lambda \rangle$: $\alpha_S(\theta, m) = \theta$, $\beta_1(m) = \beta_2(m) = \theta^*_{\mu_m}$.

Proof. For every message m, since $\phi_{\mu_m}(\theta^*_{\mu_m}) = \theta^*_{\mu_m}$, each buyer's expected payoff is 0. Any deviation to a lower bid also gives a payoff of zero. From lemma 5, for any $\theta \in (\theta^*_{\mu_m}, 1]$, $\phi_{\mu_m}(\theta) < \theta$, so any deviation to a bid higher than $\theta^*_{\mu_m}$ would lead to a negative payoff. Therefore, no buyer has an incentive to deviate. QED

LEMMA 7. Let (λ^*, σ^*) be a WD-IC profile. For each message m, let $p(m) \equiv \max\{\beta^*_1(m), \beta^*_2(m)\}$ denote the equilibrium market price in the game $\langle G, \lambda^* \rangle$. Then $\phi_{\mu_m}(p(m)) = p(m)$ for each $m \in M$.

Proof. Suppose $\phi_{\mu_m}(p(m)) < p(m)$; then the winning buyer's payoff is negative and can profitably deviate to bid 0. Now suppose $\phi_{\mu_m}(p(m)) > p(m)$; we show that at least one buyer has an incentive to bid a higher price.

If $\beta^*_1(m) \neq \beta^*_2(m)$, then the losing bidder can profitably deviate. Since $\phi_{\mu_m}(\cdot)$ is weakly increasing, there exists small enough ε such that $\phi_{\mu_m}(p(m) + \varepsilon) > p(m) + \varepsilon$. So the losing bidder can deviate to bidding $p(m) + \varepsilon$ and receives a strictly positive payoff.

If $\beta^*_1(m) = \beta^*_2(m) = b$ for some b, we show that both buyers have an incentive to deviate. Let $K \equiv \phi_{\mu_m}(b) - b > 0$. Since ties are broken evenly, each buyer's payoff is $(1/2)P_{\mu_m}(\theta \leq b)K$. By letting $\varepsilon < (K/2)$, we have

$$\phi_{\mu_m}(b + \varepsilon) - b - \varepsilon \geq \phi_{\mu_m}(b) - b - \varepsilon = K - \varepsilon > \frac{K}{2}.$$

[22] For x less than $\underline{\theta}_\mu$, we set $\phi_\mu(x) = v(\underline{\theta}_\mu)$.

So if either of the bidders deviates to bidding $b + \varepsilon$, he receives a payoff of $P_{\mu_u}(\theta \leq b + \varepsilon)[\phi_{\mu_u}(b + \varepsilon) - b - \varepsilon] > (1/2)P_{\mu_u}(\theta \leq b)K$, which is profitable. QED

LEMMA 8. If a profile (λ^*, σ^*) is credible and WD-IC, then there exists a set $E \subset \Theta \times M$ such that $\lambda^*(E) = 1$, and for any $(\theta, m), (\theta', m') \in E$,

$$\max\{\theta, p(m)\} + \max\{\theta', p(m')\} \geq \max\{\theta, p(m')\} + \max\{\theta', p(m)\}.$$

Proof. Since (λ^*, σ^*) is WD-IC, trade happens only when the seller's ask price $\alpha^*(\theta, m) = \theta$ is higher than the prevailing market price $p(m) = \max\{\beta_1^*(m), \beta_2^*(m)\}$. The seller's payoff function can therefore be simplified as

$$u_S(\theta, \sigma^*(\theta, m)) = u_S(\theta, \alpha^*(\theta, m), \beta_1^*(m), \beta_2^*(m)) = \max\{\theta, p(m)\}.$$

Recall that credibility requires

$$\lambda \in \arg\max_{\lambda' \in D(\lambda)} \int u_S(\theta, \sigma^*(\theta, m)) d\lambda'(\theta, m).$$

Let $v_S(\theta, m) \equiv u_S(\theta, \sigma^*(\theta, m)) = \max\{\theta, p(m)\}$. From theorem 1 of Beiglböck et al. (2009), λ is v_S-cyclically monotone. That is, there exists a set $E \subset \Theta \times M$ such that $\lambda^*(E) = 1$, and for any sequence $(\theta_k, m_k)_{k=1}^n \in E$,

$$\sum_{k=1}^{n} v_S(\theta_k, m_k) \geq \sum_{k=1}^{n} v_S(\theta_k, m_{k+1}).$$

Suppose $(\theta, m), (\theta', m') \in E$; then v_s-cyclical monotonicity implies that

$$v_S(\theta, m) + v_S(\theta', m') \geq v_S(\theta, m') + v_S(\theta', m),$$

which is

$$\max\{\theta, p(m)\} + \max\{\theta', p(m')\} \geq \max\{\theta, p(m')\} + \max\{\theta', p(m)\}.$$

QED

In light of lemma 8, for every credible profile (λ^*, σ^*), we will focus only on pairs $(\theta, m) \in E$. We will use $\text{proj}_M(E) \equiv \{m \in M : (\theta, m) \in E\}$ to denote the projection of E onto the message space.

Let $\underline{p} = \inf\{p(m) | m \in \text{proj}_M(E)\}$ be the infimum of trading prices across all messages. For each message m, let $\Theta(m) = \{\theta : (\theta, m) \in E\}$ be the set of θ that is matched with m.

LEMMA 9. Let (λ^*, σ^*) be a credible and WD-IC profile. For every message $\hat{m} \in proj_M(E)$ such that $\hat{p} \equiv p(\hat{m}) = \max\{\beta_1^*(\hat{m}), \beta_2^*(\hat{m})\} > \underline{p}$, we have $\Theta(\hat{m}) \cap (\underline{p}, \infty) = \varnothing$.

Proof. To prove the lemma, suppose by contradiction that there exists $\hat{\theta} \in \Theta(\hat{m}) \cap (\underline{p}, \infty)$. By the definition of $\underline{p}$, there exists p' with $\underline{p} < p' < \hat{\theta}$ such that $p' = p(m')$ for some $m' \in \text{proj}_M(E)$. Since in equilibrium $p' = E_{\mu_{m'}}[v(\theta) | \theta \in \Theta(m') \cap [0, p']]$, there also exists $\theta' \in \Theta(m')$ such that $\theta' < p'$. Since $(\theta', m'), (\hat{\theta}, \hat{m}) \in E$, by lemma 8, we have

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} \leq \max\{\theta', p'\} + \max\{\hat{\theta}, \hat{p}\}.$$

Since $\theta' < p'$ by construction, we have

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} \leq p' + \max\{\hat{\theta}, \hat{p}\}. \tag{18}$$

Note also that

$$p' + \max\{\hat{\theta}, \hat{p}\} < \hat{p} + \hat{\theta}. \tag{19}$$

The inequality above follows by considering two possibilities for $\max\{\hat{\theta}, \hat{p}\}$: either $\hat{\theta} \geq \hat{p}$, in which case $p' + \max\{\hat{\theta}, \hat{p}\} = p' + \hat{\theta} < \hat{p} + \hat{\theta}$; or $\hat{\theta} < \hat{p}$, in which case $p' + \max\{\hat{\theta}, \hat{p}\} = p' + \hat{p} < \hat{p} + \hat{\theta}$ as well.

Combining (18) and (19) and noticing $\theta' < p' < \hat{p}$ and $p' < \hat{\theta}$ yield

$$\max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\} < \hat{p} + \hat{\theta} = \max\{\theta', \hat{p}\} + \max\{\hat{\theta}, p'\},$$

which is a contradiction. QED

Proof of proposition 6. To prove our result, we first calculate the seller's profit from an arbitrary credible and WD-IC profile (λ^*, σ^*). We then show that there exists another credible and WD-IC profile (λ^0, σ^0), where λ^0 is a null information structure, that leads to weakly higher profit for the seller.

Recall that the seller's payoff function can be written as $u_S(\theta, \sigma^*(\theta, m)) = \max\{\theta, p(m)\}$, so her ex ante profit is

$$\begin{aligned}
&\int_{\Theta \times M} \max\{\theta, p(m)\} d\lambda^*(\theta, m) = \int_M \int_0^1 \max\{\theta, p(m)\} d\lambda^*(\theta|m) d\lambda_M^*(m) \\
&= \int_M \left[\int_0^{p(m)} p(m) d\lambda^*(\theta|m) + \int_{p(m)}^1 \theta d\lambda^*(\theta|m) \right] d\lambda_M^*(m) \\
&= \int_M \left[p(m) P_{\lambda^*(\theta|m)}(\theta \leq p(m)) + \int_{p(m)}^1 \theta d\lambda^*(\theta|m) \right] d\lambda_M^*(m).
\end{aligned}$$

By lemma 7, $p(m) = E_{\lambda^*(\theta|m)}[v(\theta)|\theta \leq p(m)]$, so we can write the integral above as

$$\begin{aligned}
&\int_M \left[E_{\lambda^*(\theta|m)}[v(\theta)|\theta \leq p(m)] P_{\lambda^*(\theta|m)}(\theta \leq p(m)) + \int_{p(m)}^1 \theta d\lambda^*(\theta|m) \right] d\lambda_M^*(m) \\
&= \int_M \left[\int_0^{p(m)} v(\theta) d\lambda^*(\theta|m) + \int_{p(m)}^1 \theta d\lambda^*(\theta|m) \right] d\lambda_M^*(m).
\end{aligned}$$

By lemma 9, for every $m \in \text{proj}_M(E)$, if $p(m) > \underline{p}$, then $\Theta(m) \cap (\underline{p}, \infty) = \varnothing$, so the seller's profit from (λ^*, σ^*) can be further simplified to

$$\begin{aligned}
&\int_M \left[\int_0^{\underline{p}} v(\theta) d\lambda^*(\theta|m) + \int_{\underline{p}}^1 \theta d\lambda^*(\theta|m) \right] d\lambda_M^*(m) \\
&= \int_0^{\underline{p}} v(\theta) d\mu_0(\theta) + \int_{\underline{p}}^1 \theta d\mu_0(\theta).
\end{aligned} \tag{20}$$

Having calculated the seller's profit from (λ^*, σ^*), next we will construct another credible and WD-IC profile (λ^0, σ^0) with a weakly higher profit, where λ_0 is the null information structure $\mu_0 \times \delta_{m_0}$.

From lemma 9, for every $m \in \text{proj}_M(E)$,

$$\phi_{\mu_m}(p(m)) = E_{\mu_m}[v(\theta)|\theta \leq p(m)] = E_{\mu_m}[v(\theta)|\theta \leq \underline{p}] = \phi_{\mu_m}(\underline{p});$$

in addition, from lemma 7, $\phi_{\mu_m}(p(m)) = p(m)$ for every message $m \in \text{proj}_M(E)$. Combining these yields

$$\phi_{\mu_\infty}(\underline{p}) = \phi_{\mu_\infty}(p(m)) = p(m) \geq \underline{p}.$$

Taking expectation over all messages, we have $\phi_{\mu_0}(\underline{p}) \geq \underline{p}$. By Tarski's fixed point theorem, there exists a largest $p^0 \in [\underline{p}, 1)$ such that the $\phi_{\mu_0}(p^0) = p^0$.

When we use a similar argument as that in lemma 6, the strategy profile σ^0 where the seller plays her weakly dominant strategy $\alpha^0(\theta, m_0) = \theta$ and buyers play $\beta_1^0(m_0) = \beta_2^0(m_0) = p^0$ is a Bayesian Nash equilibrium in the game $\langle G, \lambda \rangle$.

It remains to show that the seller's profit from (λ^0, σ^0) is weakly higher than that from (λ^*, σ^*). Under (λ^0, σ^0), the seller's profit is

$$\begin{aligned}
\int_0^1 \max\{\theta, p^0\} d\mu_0(\theta) &= \int_0^{p^0} p^0 d\mu_0(\theta) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= p^0 P_{\mu_0}(\theta \leq p^0) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= E_{\mu_0}[v(\theta) | \theta \leq p^0] P_{\mu_0}(\theta \leq p^0) + \int_{p^0}^1 \theta d\mu_0(\theta) \\
&= \int_0^{p^0} v(\theta) d\mu_0(\theta) + \int_{p^0}^1 \theta d\mu_0(\theta).
\end{aligned} \tag{21}$$

When we compare (20) and (21), since $p^0 \geq \underline{p}$ and $v(\theta) > \theta$ for all θ, it follows that

$$\int_0^{p^0} v(\theta) d\mu_0 + \int_{p^0}^1 \theta d\mu_0 \geq \int_0^{\underline{p}} v(\theta) d\mu_0 + \int_{\underline{p}}^1 \theta d\mu_0.$$

The seller's profit under (λ^0, σ^0) is therefore weakly higher than that from (λ^*, σ^*). QED

References

Akbarpour, Mohammad, and Shengwu Li. 2020. "Credible Auctions: A Trilemma." *Econometrica* 88 (2): 425–67.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Q.J.E.* 84 (3): 488–500.

Alonso, Ricardo, and Odilon Câmara. 2016. "Persuading Voters." *A.E.R.* 106 (11): 3590–605.

Beiglböck, Mathias, Martin Goldstern, Gabriel Maresch, and Walter Schachermayer. 2009. "Optimal and Better Transport Plans." *J. Functional Analysis* 256 (6): 1907–27.

Bergemann, Dirk, and Stephen Morris. 2016. "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games." *Theoretical Econ.* 11 (2): 487–522.

Best, James, and Daniel Quigley. 2020. "Persuasion for the Long Run." Working paper.

Brocas, Isabelle, and Juan D Carrillo. 2007. "Influence through Ignorance." *RAND J. Econ.* 38 (4): 931–47.

Chakraborty, Archishman, and Rick Harbaugh. 2007. "Comparative Cheap Talk." *J. Econ. Theory* 132 (1): 70–94.

———. 2010. "Persuasion by Cheap Talk." *A.E.R.* 100 (5): 2361–82.
Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431–51.
Dworczak, Piotr, and Giorgio Martini. 2019. "The Simple Economics of Optimal Persuasion." *J.P.E.* 127 (5): 1993–2048.
Frankel, Alexander. 2014. "Aligned Delegation." *A.E.R.* 104 (1): 66–83.
Fréchette, Guillaume R., Alessandro Lizzeri, and Jacopo Perego. 2022. "Rules and Commitment in Communication: An Experimental Analysis." *Econometrica* 90 (5): 2283–318.
Gitmez, A. Arda, and Pooya Molavi. 2022. "Polarization and Media Bias." Working paper.
Goldstein, Itay, and Yaron Leitner. 2018. "Stress Tests and Information Disclosure." *J. Econ. Theory* 177:34–69.
Guo, Yingni, and Eran Shmaya. 2021. "Costly Miscalibration." *Theoretical Econ.* 16 (2): 477–506.
Hedlund, Jonas. 2017. "Bayesian Persuasion by a Privately Informed Sender." *J. Econ. Theory* 167:229–68.
Ivanov, Maxim. 2020. "Optimal Monotone Signals in Bayesian Persuasion Mechanisms." *Econ. Theory* 72:955–1000.
Jackson, Matthew O., and Hugo F. Sonnenschein. 2007. "Overcoming Incentive Constraints by Linking Decisions." *Econometrica* 75 (1): 241–57.
Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Ann. Rev. Econ.* 11:249–72.
Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *A.E.R.* 101 (6): 2590–615.
Kartik, Navin. 2009. "Strategic Communication with Lying Costs." *Rev. Econ. Studies* 76 (4): 1359–95.
Kartik, Navin, and Weijie Zhong. 2019. "Lemonade from Lemons: Information Design and Adverse Selection." Working paper.
Koessler, Frédéric, and Vasiliki Skreta. 2021. "Information Design by an Informed Designer." Working paper.
Kolotilin, Anton. 2018. "Optimal Information Disclosure: A Linear Programming Approach." *Theoretical Econ.* 13 (2): 607–35.
Kolotilin, Anton, and Hongyi Li. 2021. "Relational Communication." *Theoretical Econ.* 16 (4): 1391–430.
Kuvalekar, Aditya, Elliot Lipnowski, and Joao Ramos. 2022. "Goodwill in Communication." *J. Econ. Theory* 203:105467.
Libgober, Jonathan. 2022. "False Positives and Transparency." *American Econ. J. Microeconomics* 14 (2): 478–505.
Lipnowski, Elliot, and Doron Ravid. 2020. "Cheap Talk with Transparent Motives." *Econometrica* 88 (4): 1631–60.
Lipnowski, Elliot, Doron Ravid, and Denis Shishkin. 2022. "Persuasion via Weak Institutions." *J.P.E.* 130 (10): 2705–30.
Margaria, Chiara, and Alex Smolin. 2018. "Dynamic Communication with Biased Senders." *Games and Econ. Behavior* 110:330–39.
Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*, vol. 1. New York: Oxford Univ. Press.
Mathevet, Laurent, David Pearce, and Ennio Stacchetti. 2022. "Reputation for a Degree of Honesty." Working paper.
Matsushima, Hitoshi, Koichi Miyazaki, and Nobuyuki Yagi. 2010. "Role of Linking Mechanisms in Multitask Agency with Hidden Information." *J. Econ. Theory* 145 (6): 2241–59.

Meng, Delong. 2021. "On the Value of Repetition for Communication Games." *Games and Econ. Behavior* 127:227–46.
Mensch, Jeffrey. 2021. "Monotone Persuasion." *Games and Econ. Behavior* 130:521–42.
Min, Daehong. 2021. "Bayesian Persuasion under Partial Commitment." *Econ. Theory* 72:743–64.
Nguyen, Anh, and Teck Yong Tan. 2021. "Bayesian Persuasion with Costly Messages." *J. Econ. Theory* 193:105212.
Ostrovsky, Michael, and Michael Schwarz. 2010. "Information Disclosure and Unraveling in Matching Markets." *American Econ. J. Microeconomics* 2 (2): 34–63.
Pei, Harry. 2020. "Repeated Communication with Private Lying Cost." Working paper.
Perez-Richet, Eduardo. 2014. "Interim Bayesian Persuasion: First Steps." *A.E.R.* 104 (5): 469–74.
Perez-Richet, Eduardo, and Vasiliki Skreta. 2021. "Test Design under Falsification." Working paper.
Rahman, David. 2010. "Detecting Profitable Deviations." Working paper.
Rayo, Luis, and Ilya Segal. 2010. "Optimal Information Disclosure." *J.P.E.* 118 (5): 949–87.
Renault, Jérôme, Eilon Solan, and Nicolas Vieille. 2013. "Dynamic Sender–Receiver games." *J. Econ. Theory* 148 (2): 502–34.
Rochet, Jean-Charles. 1987. "A Necessary and Sufficient Condition for Rationalizability in a Quasi-Linear Context." *J. Math. Econ.* 16 (2): 191–200.
Taneva, Ina. 2019. "Information Design." *American Econ. J. Microeconomics* 11 (4): 151–85.
Topkis, Donald M. 2011. *Supermodularity and Complementarity*. Princeton, NJ: Princeton Univ. Press.
Villani, Cédric. 2008. *Optimal Transport: Old and New*, vol. 338. Heidelberg: Springer.
Zapechelnyuk, Andriy. 2023. "On the Equivalence of Information Design by Uninformed and Informed Principals." *Econ. Theory* 76:1051–67.