Strategic Communication with Negative Reciprocity and Endogenous Preferences[*]

Ran Eilat[†] Kfir Eliaz[‡]

January 28, 2026

Abstract

We incorporate negative reciprocity into strategic information transmission and study how a receiver's response to perceived manipulation shapes equilibrium outcomes. We examine both an information-design environment, in which the sender can commit to an information structure, and a cheap-talk environment, in which commitment is not possible. In each setting, the receiver's behavior or preferences shift adversarially against the sender as the sender's strategy becomes less accurate. We analyze and solve these communication games in the presence of negative reciprocity, even when it generates endogenous preferences, and we characterize the novel ways in which classical results are altered. Our findings link information design to behavioral economics highlighting the implications of design-dependent preferences.

1 Introduction

The traditional approach in economics to policy and mechanism design assumes that individuals' preferences are fixed and unaffected by the very policies and mechanisms they operate within. This assumption allows economists to assert that an individual is made better off if he moves from one situation to another

[*] A previous shorter version of this paper was circulated under the title "Lost in Persuasion: Negative Reciprocity in Information Design". We thank Yair Antler, Ezra Einy, Roee Levy and Roi Zultan for helpful comments.

[†] School of Economics, Tel Aviv University, eilatr@tauex.tau.ac.il

[‡] School of Economics, Tel Aviv University, kfire@tauex.tau.ac.il and King's Business School

that he prefers under his current preferences. However, as Gintis (1972) pointed out in his critique of welfare economics,

> [This assertion is only] "intuitively obvious" if the process by which individuals are moved to preferred positions does not, in itself, alter their preference structures. The individual *would be* better off, we will agree, *if* he were the same individual; i.e., insofar as his preference structure had not changed in the process of movement." (Gintis, 1972, p. 578)

The notion that institutions and mechanisms shape the values and preferences of their participants is not merely theoretical; it is supported by extensive experimental evidence. Falk and Szech (2013) demonstrated that merely framing a mechanism as a market, where outcomes are determined by clearing prices, can influence how participants value harm inflicted on third parties. Similarly, the vast literature on intentions-based social preferences documents how individuals exhibit negative reciprocity, willingly forgoing material payoffs to punish behavior perceived as unfair or manipulative (for a survey, see Fehr and Gächter (1998)).

The economic literature on negative reciprocity has primarily focused on allocation mechanisms. However, if the tendency to exhibit negative reciprocity for manipulative intentions is deeply rooted in human nature, then we would expect it to be displayed also in other domains. One such domain is strategic information transmission, where an informed party seeks to influence an uninformed decision-maker's beliefs to induce an action favorable to the informed party.

Negative reciprocity in strategic information transmission can manifest in various ways. A decision-maker who suspects manipulation may choose to disregard the sender's information entirely, even when the information holds instrumental value. For example, a news consumer might stop reading a newspaper if he believes it is attempting to steer him toward more extreme political views, even if the newspaper generally aligns with his existing political orientation. Similarly, a driver using a navigation app might ignore its suggested route if he suspects the app is experimenting with traffic patterns rather than recommending the shortest or fastest route (Kremer et al., 2014).

Another way negative reciprocity may manifest is through actions that are, at least weakly, unfavorable to the manipulative sender. For example, a judge who perceives the prosecution as excessively strategic in presenting evidence may respond by raising the burden of proof required for conviction. Likewise, when a financial advisor who privately knows the riskiness of an investment, and who stands to benefit from recommending riskier options, communicates vaguely with a client, the client may respond to the advisor's manipulation by shifting toward safer investments than he would have chosen otherwise.

This paper incorporates negative reciprocity into the framework of strategic information transmission, allowing the sender's strategy (and in particular how "manipulative" it is) to directly influence the receiver's preferences and behavior, beyond its indirect effect through belief updating. We consider both environments with commitment, as in information-design, and environments without commitment as in cheap-talk. In each environment, we represent the degree of manipulation in the sender's strategy by an appropriate measure of inaccuracy. Our goal is to study the effect of negative reciprocity on equilibrium outcomes.

We begin by analyzing a "pure" persuasion setting, where a sender with state-*independent* preferences *commits* ex-ante to a Blackwell experiment. To demonstrate that negative reciprocity is not wedded to one particular form, we explore two distinct manifestations of negative reciprocity and analyze their implications. We first consider a receiver who ignores the sender's signal with a probability that increases proportionally with the signal's inaccuracy. We then study a receiver whose preferences over actions shift adversely for the sender as the signal's inaccuracy increases. Specifically, for each posterior belief, the receiver's preferred action becomes weakly worse for the sender when the signal is more inaccurate.

Next, we consider a canonical cheap-talk setting where a sender with state-*dependent* preferences *cannot* commit ex-ante to a message strategy. We introduce a novel feature that allows the receiver's preferences to depend directly on the sender's *strategy.* Specifically, we assume that when the sender deviates from full disclosure of the state, the receiver develops a bias in the direction opposite to the sender's, with a magnitude that depends on the inaccuracy of the sender's strategy.

Finally, we use an example to demonstrate how negative reciprocity can affect the outcome of designing information for multiple interactive receivers.

The contribution of this paper is threefold. First, we propose a tractable approach to incorporating the motive of negative reciprocity into the canonical models of strategic communication in economics. Second, we show how to analyze and solve the model with this additional new motive, even when it gives rise to endogenous preferences. Finally, we characterize the novel ways in which classical results change in the presence of negative reciprocity.

Related Literature This paper relates to the extensive literature on reciprocity in economics, which explores how behavior depends not only on material outcomes but also on perceived intentions. Rabin (1993) formalizes fairness by allowing utility to depend on beliefs about others' kindness. Charness and Rabin (2002) propose a model combining outcome-based and intention-based preferences, capturing how distributional concerns and reciprocity jointly shape behavior. Dufwenberg and Kirchsteiger (2004) extend this framework to extensive games, emphasizing the role of the sequential structure. Fehr and Gächter (2000) provide experimental evidence of reciprocity, demonstrating that individuals willingly forgo material payoffs to punish unfair behavior.

A relatively small literature examines mechanism design environments in which the design itself influences participants' preferences. Falk and Szech (2013) show that merely framing an interaction as a market, where outcomes are determined by clearing prices, can alter how participants value harm inflicted on third parties. More recently, Antler (2015) studied two-sided matching where agents' rankings of potential partners depend on how they are ranked by others. Bierbrauer and Netzer (2016) examined what social choice functions are implementable under intention-based social preferences. Au et al. (2023) study Bayesian persuasion in the lab and find evidence consistent with receivers penalizing senders they perceive as hostile in information provision.

Since the receiver in our framework exhibits psychologically motivated behavior, our research is also related to Caplin and Eliaz (2003) who analyze the design of information structure for receivers who prefer late resolution of uncertainty; Caplin and Leahy (2004), who study the problem of a benevolent sender facing a receiver with belief-based utility; Lipnowski and Mathevet (2018), who analyze a broader setting of a receiver with intrinsic preferences over beliefs; and, more recently, Hagenbach and Saucet (2025), who experimentally study how individuals read strategically-transmitted information when they have preferences over what they will learn.

2 Pure Persuasion

We begin by presenting a benchmark setup for a persuasion problem, in which the receiver has "standard material preferences". In subsequent sections, we augment this framework by introducing two forms of negative reciprocity on the part of the receiver.

The Environment. A state is drawn from a finite set $\Omega = \{\omega_1, \dots, \omega_L\}$, with prior probabilities $q \in \Delta(\Omega)$. There are two players: Sender and Receiver. Sender designs a signal that conveys information about the state to Receiver. After observing the signal's outcome, Receiver chooses an action from a finite set $A = \{a_1, \dots, a_K\}$.

Let $u_R(a, \omega)$ denote Receiver's material payoff when he takes action $a \in A$ in state $\omega \in \Omega$. Assume that in each state $\omega \in \Omega$ Receiver has a unique optimal action.[1] Denote Sender's state-independent utility from action a by $u_S(a)$. For simplicity, we assume that the actions in A can be strictly ordered according to the utility they generate for Sender, and without further loss of generality, let $u_S(a_1) > u_S(a_2) > \cdots > u_S(a_K)$.

Sender commits to a signal, formally defined as a function σ that maps each state in Ω to a distribution over possible realizations in some finite set S.[2] Upon observing a realization $s \in S$, Receiver updates his posterior belief. As is customary in the literature, a signal can equivalently be described by a Bayes-plausible distribution over posterior beliefs.

Let $M_\sigma = \{\mu_1, \dots, \mu_N\}$ denote the set posterior belief distributions induced by a signal σ, where $\mu_i \in \Delta(\Omega)$. Let $p_\sigma(\mu)$ denote the probability that a posterior $\mu \in M_\sigma$ is realized under σ. Bayes plausibility requires:

$$\sum_{\mu \in M_\sigma} p_\sigma(\mu)\mu = q. \tag{BP}$$

We denote by e_i the *pure posterior belief* that assigns probability 1 to state ω_i. The *truthful signal*, denoted σ_T, corresponds to the distribution $(q_1, \dots, q_L)$ over the pure posteriors $(e_1, \dots, e_L)$. At the other extreme, not providing any

[1] This assumption implies that for posterior beliefs that are "close to" a "pure" posterior that assigns probability 1 to one of the states, the optimal action remains unchanged; we use this property in the proof of Proposition 6.

[2] For ease of exposition, we assume that S is finite, but this can be generalized at the cost of more cumbersome notations.

information corresponds to the degenerate signal that assigns probability 1 to the prior belief q.

Given a posterior belief μ, let $\hat{u}_S(\mu)$ denote Sender's indirect utility associated with μ, that is, Sender's utility from Receiver's best response to belief μ. If Receiver has multiple best responses, we assume that he selects the one most preferred by Sender. Define

$$W(\sigma) \;=\; \sum_{\mu \in M_\sigma} p_\sigma(\mu)\, \hat{u}_S(\mu), \tag{1}$$

and assume that, absent any reciprocity motives, Sender prefers full revelation of the state to providing no information:

$$W_\varnothing \equiv \hat{u}_S(q) \;<\; W(\sigma_T) \equiv W_T. \tag{2}$$

Extreme Beliefs. Denote by $B^*(a) \subset \Delta(\Omega)$ the set of posterior beliefs for which action $a \in A$ is optimal for Receiver. That is, $\mu \in B^*(a)$ if and only if

$$\sum_{\omega \in \Omega} \mu(\omega)\,(u_R(a,\omega) - u_R(a',\omega)) \geq 0, \quad \forall a' \in A. \tag{3}$$

where $\mu(\omega)$ denotes the probability assigned to state ω by the belief μ.

Each inequality in (3) defines a closed half-space in $\mathbb{R}^{|\Omega|}$, so $B^*(a)$ is the intersection of the simplex $\Delta(\Omega)$ with finitely many half-spaces. Since the simplex is a bounded convex set and each half-space is convex, it follows that $B^*(a)$ is a convex polytope.

Denote by $\text{Ext}(B^*(a))$ the set of extreme points of $B^*(a)$, and note that this set is finite.[3] Let

$$\mathcal{E} = \bigcup_{a \in A} \text{Ext}(B^*(a)) \tag{4}$$

denote the union of extreme points of $B^*(a)$ across all actions. We refer to elements in $\mathcal{E}$ as *extreme beliefs* of Receiver. Note that they are determined only by Receiver's indifferences between optimal actions, and by the probability simplex.

Measuring Inaccuracy. The building block of our inaccuracy measure in this section is the information-theoretic notion of *mutual information*, which

[3] See also Proposition 2 below.

quantifies the reduction in uncertainty about one random variable (the state) that results from observing the realization of another random variable (the signal). Specifically, given a signal function σ, let $\mathcal{S}_\sigma$ denote the random variable associated with the signal realizations induced by σ, and let $\mathcal{Q}$ denote the random variable associated with the state. The mutual information between $\mathcal{S}_\sigma$ and $\mathcal{Q}$ is given by:

$$I(\mathcal{Q};\mathcal{S}_\sigma) = H(\mathcal{Q}) - \mathbb{E}_{\mathcal{S}_\sigma}[H(\mathcal{Q}|\mathcal{S}_\sigma)], \tag{5}$$

where $H(\mathcal{Q})$ denotes the Shannon entropy of the prior belief, and $H(\mathcal{Q}|\mathcal{S}_\sigma)$ is the conditional Shannon entropy of the posterior belief about the state given the signal realization. With a slight abuse of notation, we identify a random variable with its associated probability distribution, and rewrite the right-hand side of Equation (5) as:

$$H(q) - \sum_{\mu \in M_\sigma} p_\sigma(\mu) H(\mu),$$

where $H(q) = -\sum_{\omega\in\Omega} q(\omega)\ln(q(\omega))$ is the entropy of the prior q and $H(\mu) = -\sum_{\omega\in\Omega} \mu(\omega)\ln(\mu(\omega))$ is the entropy of the posterior μ. Hence, greater mutual information indicates that the signal conveys more information about the state, implying that it is more accurate.

To normalize this measure and express *inaccuracy* rather than accuracy, we define

$$\rho(\sigma) = 1 - \frac{I(\mathcal{Q};\mathcal{S}_\sigma)}{H(q)} = \frac{\sum_{\mu\in M_\sigma} p_\sigma(\mu)H(\mu)}{H(q)}. \tag{INAC}$$

This transformation expresses inaccuracy as the proportion of the initial uncertainty that remains after observing the signal. The value of ρ ranges from 0 to 1, where $\rho = 1$ corresponds to providing no information (i.e., the posterior entropy remains equal to the prior entropy), while $\rho = 0$ corresponds to the truthful signal that eliminates uncertainty entirely (i.e., the posterior entropy is zero).

Finally, note that the inaccuracy measure ρ is consistent with Blackwell informativeness in the following sense: if a signal σ is Blackwell more informative than a signal σ', then σ' is more inaccurate than σ. This follows directly from the Data Processing Inequality, since the outcome of σ' is conditionally independent of the state ω, given the outcome of σ.[4]

We now proceed to explore two alternative channels through which Receiver

[4] Given three random variables, X, Y, Z, that form a Markov chain $X \to Y \to Z$, the Data Processing Inequality states that $I(X,Y) \geq I(X,Z)$, with equality if and only if X and Y are conditionally independent given Z (see Theorem 2.8.1 in Cover and Thomas (2012)).

may negatively reciprocate Sender who is perceived as manipulative.

2.1 Ignoring a Manipulative Sender

Our first departure from the standard communication model is to assume that Receiver ignores the signal with probability $\gamma(\sigma)$ that is propositional to the signal's inaccuracy $\rho(\sigma)$:

$$\gamma(\sigma) = \lambda\rho(\sigma),$$

where $\lambda \in [0, 1]$ is a scaling parameter.

Under this interpretation, Sender conducts his experiment, and Receiver may decide not to observe its realization with probability $\gamma(\sigma)$. A second interpretation is that Sender offers Receiver a test or an experiment, and Receiver does not carry it out with probability $\gamma(\sigma)$. A third interpretation is that there is a continuum of Receivers, each characterized by a threshold tolerance for inaccuracy, where threshold are uniformly distributed over $[0, 1]$ in the population. If $\gamma(\sigma)$ exceeds the Receiver's threshold, then Receiver refuses to interact with Sender.

Under this specification, Sender's expected payoff from employing a signal σ, denoted by $V(\sigma)$, is given by:

$$V(\sigma) = \gamma(\sigma)W_{\varnothing} + (1 - \gamma(\sigma))\,W(\sigma) \tag{6}$$

where $W(\sigma)$ is defined in Equation (1). The first term on the right-hand side represents Sender's utility from the action Receiver takes if he ignores the information. The second term captures Sender's expected payoff when Receiver optimally responds to the signal.

Sender's objective is therefore to find:[5]

$$\sigma^* \in \arg\max_{\sigma} V(\sigma).$$

2.1.1 Warm-up: a binary environment

Suppose there are two actions, $A = \{a_1, a_2\}$, and two states of the world, $\Omega = \{\omega_1, \omega_2\}$. With a slight abuse of notation we identify a belief with the probability

[5] Representing a signal as a distribution over posterior beliefs, the set of signals corresponds to the set of Bayes-plausible distributions over beliefs, which is compact. Hence, the function V, being upper semicontinuous and defined on a compact set, attains a maximum.

that the state is ω_1. Suppose that the prior belief is $q < 0.5$.

Sender's utility is 1 if Receiver chooses a_1, and 0 otherwise. Receiver's material payoff is 1 if his action matches the state, and 0 otherwise, i.e. $u_R(a_i, \omega_j) = \mathbb{1}_{(i=j)}$. Hence, conditional on Receiver not ignoring the information, he chooses a_1 if and only if his posterior belief that the state is ω_1 exceeds 0.5.

Suppose, for now, that Sender is restricted to recommending actions to Receiver, or equivalently, to employing only binary signals that generate two posterior beliefs, i.e., $M_\sigma = \{\mu_1, \mu_2\}$. By Equation (BP), if the signal is informative, then one of these posterior beliefs must be above q and the other below q. Without loss of generality, assume $\mu_2 < q < \mu_1$. Furthermore, given M_σ, Equation (BP) uniquely determines the probabilities $p_\sigma(\mu_1)$ and $p_\sigma(\mu_2)$.

Since in the current example Sender's utility is zero when Receiver ignores the information, i.e. $W_\varnothing = 0$, Equation (6) simplifies in the case of a binary signal to:[6]

$$V(\sigma) = \big[1 - \gamma(\sigma)\big] \frac{q - \mu_2}{\mu_1 - \mu_2}, \tag{7}$$

The expression on the right-hand side of Equation (7) consists of a product of two terms. The second term, $\frac{q-\mu_2}{\mu_1-\mu_2}$, is decreasing in μ_2 by direct computation. The first term, $1 - \gamma(\sigma)$, is also decreasing in μ_2. This is an implication of the following result:

Lemma 1 *For a fixed $\mu_1 > q$, the inaccuracy measure $\gamma(\sigma)$ is strictly increasing in μ_2 for all $\mu_2 \in (0, q)$.*

Thus, as in the "standard" two-actions, two-states Bayesian persuasion problem, the optimal signal for Sender satisfies $\mu_2 = 0$. Indeed, decreasing μ_2 serves two purposes simultaneously: it reduces inaccuracy and increases the probability that Receiver takes the action preferred by Sender.

The following result characterizes the optimal binary signal.

Lemma 2 *Sender's expected payoff, V, is maximized when $\mu_2 = 0$ and either $\mu_1 = 0.5$ or $\mu_1 = 1$.*

In words, when restricted to binary signals, it is optimal for Sender to either fully reveal the state ($\mu_2 = 0, \mu_1 = 1$) or to use the same signal as in the "standard"

[6] Here, $\gamma(\sigma) = \frac{\lambda}{H(q)}\big(p_\sigma(\mu_1)H(\mu_1) + p_\sigma(\mu_2)H(\mu_2)\big)$.

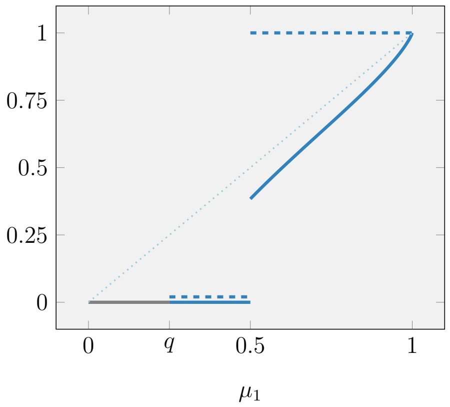

Figure 1: Sender's (per-posterior) indirect utility in a binary environment with $q = 0.25$ and $\lambda = 1$, from a signal that induces posterior beliefs $(\mu_1, \mu_2 = 0)$, plotted as a function of μ_1. The dashed line shows Sender's utility conditional on Receiver best-responding to the information ($\hat{u}_S$); The solid line shows Sender's utility accounting for the probability that Receiver ignores the information.

Bayesian persuasion problem $(\mu_2 = 0, \mu_1 = 0.5)$ (although this yields a lower expected payoff compared to the standard case because, with some probability, Receiver ignores the information).

In the "standard" persuasion problem, the optimal signal can be derived using a concavification method applied to Sender's indirect utility function (Kamenica and Gentzkow, 2011). In an environment with two states and two actions, this method can be depicted graphically. In our augmented setup with negative reciprocity, the concavification argument becomes more involved, because Sender's payoff is indirectly affected by the signal he constructs via the probability that Receiver ignores the signal. In the current example, because $\mu_2 = 0$, the posterior μ_1 determines both Receiver's best response *and* the inaccuracy measure, and therefore we can graphically depict Sender's expected payoff as a function of each posterior μ_1.

This is shown in Figure 1, where the dashed line represents Sender's indirect utility for each high posterior $\mu_1 > q$, *assuming Receiver responds to the signal.* The solid line accounts for the probability that Receiver ignores the information. The dotted line represents a "concavification" of Sender's "net indirect utility". Figure 1 depicts the case where the truthful signal is optimal.

2.1.2 Failure of the Revelation Principle

We now show that, in searching for the optimal signal, one cannot restrict attention only to signals that recommend actions. Specifically, in the two-states, two actions environment presented in Section 2.1.1, we demonstrate that a signal with three realizations can outperform the optimal binary signal. To illustrate this, we consider the case that $q = 0.25$ and $\lambda = \frac{\frac{1}{2}\ln 2 + \frac{3}{4}\ln\frac{4}{3}}{\ln 2} \approx 0.81$.

By Lemma 2, when restricted to binary signals, Sender either employs a truthful signal σ_T, or a signal σ_{BP}, corresponding to the set of posteriors $M_{\sigma_{BP}} = \{0, 0.5\}$, which is optimal in the standard Bayesian persuasion setting. Computing Sender's utility, we obtain that $V(\sigma_T) = V(\sigma_{BP}) = 0.25$. Hence, both σ_T and σ_{BP} are optimal among the set of binary signals, and Sender is indifferent between employing either one.

Consider instead a non-binary signal σ' that induces the set of posteriors $M_{\sigma'} = \{0, 0.5, 1\}$, realized with probabilities $p_{\sigma'}(1) = \frac{1}{6}$, $p_{\sigma'}(0.5) = \frac{1}{6}$, and $p_{\sigma'}(0) = \frac{2}{3}$. Since $H(1) = H(0) = 0$, the inaccuracy measure of this signal, computed according to Equation (INAC), simplifies to:

$$\rho(\sigma') = \frac{p_{\sigma'}(0.5)H(0.5)}{H(0.25)} \approx 0.2054.$$

Since Sender's utility is zero when Receiver ignores the information, Equation (6) implies that Sender's expected payoff from employing the signal σ' is given by:

$$V(\sigma') = (p_{\sigma'}(0.5) + p_{\sigma'}(1))\Big[1 - \lambda \cdot \rho(\sigma')\Big] \approx 0.2778.$$

It follows that Sender's expected payoff from the signal σ' exceeds her payoff under the *optimal* binary signal.

2.1.3 The structure of optimal signals

When applicable, the Revelation Principle provides structure to the optimal signal. For instance, in settings with two actions, it implies that only two posteriors are sufficient. Since the Revelation Principle does not hold in the presence of negative reciprocity, a question arises: does any discipline remain on the the structure of optimal signals?

Our next result shows that the optimal signal is supported on a set of extreme

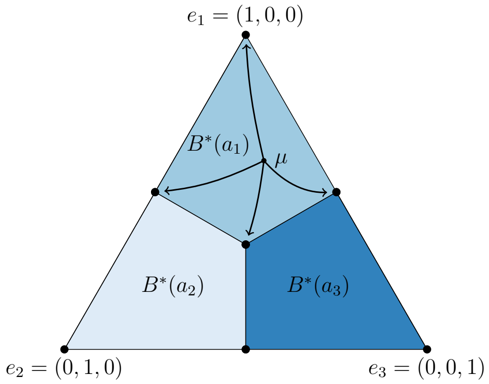

Figure 2: A probability simplex in a three-actions, three-states example, where each region represents a set of beliefs for which a different action is optimal. Extreme beliefs are indicated by the black dots.

beliefs, as defined in Equation (4). In Figure 1, which corresponds to the two-states two-actions example in Section 2.1.1, the set of extreme beliefs is $\mathcal{E} = \{0, 0.5, 1\}$. Figure 2 illustrates a setting with three states and three actions. Each of the three colored regions represent a set of posterior beliefs for which one of the actions is optimal, projected on the probability simplex. The extreme beliefs in this case are indicated by the black dots.

Proposition 1 *If σ^* is an optimal signal, then $M(\sigma^*) \subseteq \mathcal{E}$.*

This result is closely related to the Method of Posterior Covers introduced by Lipnowski and Mathevet (2017, 2018). In terms of this method, the collection $\{B^*(a)\}_{a\in A}$ constitutes a $\hat{u}_S$-posterior cover. Therefore, by Theorem 1 in Lipnowski and Mathevet (2017), conditional on Receiver responding to information, there exists an optimal signal supported only on the extreme points of $\mathcal{E}$. However, in our setting, Sender's payoff is not additive across induced posterior beliefs, and his preference for more informative signals does not stem from the convexity of $\hat{u}_S$ over elements of the posterior cover. Instead, it arises from the effect of increased informativeness on the likelihood that Receiver responds to information. This implies that the support of an optimal signal *must* lie within $\mathcal{E}$. In light of these differences, we provide in the Appendix a direct proof of Proposition 1.

The fact that, in an optimal signal structure, posterior beliefs are supported only on extreme beliefs in $\mathcal{E}$ allows us to derive an upper bound on the number of posterior beliefs employed by optimal signals. To state the result, we impose the additional assumption that, for each action $a \in A$, the set of beliefs where a is optimal, $B^*(a)$, has full dimension in $\mathbb{R}^{L-1}$, where L is the number of states. We then obtain the following result:

Proposition 2 *Suppose that* $\dim(B^*(a)) = L - 1$ *for every* $a \in A$. *Then the number of posterior beliefs induced by any optimal signal is bounded from above by* $K \cdot \Phi$, *where* Φ *is the following sum of binomial coefficients:*

$$\Phi = \binom{K + \lfloor \frac{L-1}{2} \rfloor}{K} + \binom{K + \lfloor \frac{L-2}{2} \rfloor}{K}. \tag{8}$$

The result follows from the *Upper Bound Theorem* for convex polytopes.[7] Note, however, that in general the bound characterized in Proposition 2 is not tight because the intersection of the set of extreme beliefs for different actions may not be empty. For instance, in the example presented in Section 2.1.1, with $A = \{a_1, a_2\}$ and $\Omega = \{\omega_1, \omega_2\}$, we have $L = K = 2$ and therefore $K \cdot \Phi = 2 \cdot (1+1) = 4$. While each set of beliefs $B^*(a_1) = [0.5, 1]$ and $B^*(a_2) = [0, 0.5]$ has two extreme points, the belief 0.5 is an extreme point of both sets, hence, the number of distinct extreme beliefs in this example is three.

2.1.4 Monotonicity in the sensitivity to manipulation

We now turn to explore how the inaccuracy of the optimal signal changes with Receiver's sensitivity to Sender's manipulation, as captured by the parameter λ.

Our first observation is that when Receiver becomes more sensitive to manipulation, Sender's optimal signal becomes weakly more accurate. Formally,

Proposition 3 *Suppose that* σ *and* σ' *are Sender–optimal signals for reciprocity parameters* λ *and* λ', *respectively, with* $\lambda' > \lambda$. *Then* $\rho(\sigma') \leq \rho(\sigma)$.

The proof follows from a revealed preference argument. Intuitively, as Receiver becomes more sensitive to manipulation, inaccurate signals are penalized more heavily in Sender's payoff, so any signal that is optimal at a higher sensitivity must be at least as accurate as one that is optimal at a lower sensitivity, or else it would also be preferred at the lower sensitivity, contradicting optimality.

Our second observation is that, at extreme levels of sensitivity to manipulation, the optimal signal exhibits "rigidity". To present this result, suppose that we allow

[7] See, for example, Theorem 18.1 in Brøndsted (1983).

λ to exceed 1, but to keep the model well-defined, we redefine the probability of ignoring the signal to be $\gamma(\sigma) = \min\{1, \lambda\rho(\sigma)\}$.

Proposition 3 shows that as λ increases, the inaccuracy of Sender's optimal signal weakly decreases. A natural question is whether this inaccuracy remains bounded away from zero. Our next result shows that it does not: for sufficiently large values of λ, the optimal signal fully reveals the state.[8]

At the other extreme, when $\lambda = 0$ the model coincides with the Bayesian persuasion benchmark, and the optimal signal is exactly the benchmark solution. A natural question is how the optimal signal behaves for values of λ close to zero. Our next result shows that the benchmark signal remains optimal for positive values of λ that are sufficiently small. Formally,

Proposition 4 *There exists $\overline{\lambda} > 0$ such that for any $\lambda > \overline{\lambda}$, the truthful signal σ_T is Sender-optimal. Conversely, suppose that under "standard" Bayesian persuasion with utilities u_R and u_S, and prior belief q, the signal σ_{BP} is uniquely optimal.*[9] *Then, there exists $\underline{\lambda} > 0$ such that for any $\lambda < \underline{\lambda}$ the signal σ_{BP} remains optimal.*

The proof of the first part of Proposition 4 proceeds by demonstrating that when λ is sufficiently large, any biased signal can be improved by "moving slightly" towards truth-telling. The crux of the proof is finding a "uniform threshold" $\overline{\lambda}$ that depends only on the primitives.

The second part of the result relies on the following reasoning. When $\lambda = 0$, the problem of finding the optimal signal reduces to maximizing a linear objective over a polytope. An optimal solution is therefore attained at an extreme point of the feasible set. When $\lambda > 0$, the objective function is perturbed by the addition of a nonlinear term. This term is continuously differentiable, and thus the gain it generates from small deviations around the original optimizer is uniformly bounded above as a function of the size of the deviation. By choosing $\underline{\lambda}$ sufficiently small, we guarantee that the loss induced by the linear term dominates any po-

[8] This result is not trivial, since even for arbitrarily large values of λ Sender can always design a signal that induces Receiver to respond nontrivially to it with positive probability. The result therefore shows that, even in this case, truth telling remains Sender's optimal strategy. Note, however, that this conclusion depends on the specific manifestation of negative reciprocity considered here. As we will see in the next section, under a different manifestation of negative reciprocity, truth telling may *never* be optimal.

[9] The restriction to a unique solution in the Bayesian persuasion benchmark simplifies the proof but is not essential for the substance of the result.

tential gain from the nonlinear term, so that the original extreme point remains optimal.

2.2 Disutility from helping a manipulative sender

We now turn to explore a different manifestation of negative reciprocity toward a manipulative sender. Specifically, we assume that Receiver experiences disutility from choosing Sender's preferred actions.

Suppose that Receiver's utility function $u_R(\omega, a, \rho)$ is parameterized by the inaccuracy measure ρ in a continuous way, where for $\rho = 0$ it coincides with Receiver's material payoff. Accordingly, denote by $B^*_\rho(a)$ the set of beliefs for which action $a \in A$ is optimal when the inaccuracy level is ρ.

We assume that, as ρ increases, Receiver's optimal action for each posterior belief becomes weakly less favorable for Sender. Formally, if $\rho > \rho'$, then

$$\bigcup_{i=1}^{k} B^*_\rho(a_i) \subseteq \bigcup_{i=1}^{k} B^*_{\rho'}(a_i) \quad \text{for all } k = 1, \ldots, K. \tag{9}$$

Following the convention of the preceding section, we let $\mathrm{Ext}(B^*_\rho(a))$ denote the set of extreme points of $B^*_\rho(a)$, and refer to its elements as the *extreme beliefs* for action a. We define

$$\mathcal{E}_\rho = \bigcup_{a \in A} \mathrm{Ext}(B^*_\rho(a))$$

as the union of extreme beliefs across all actions.

As in the specification explored in Section 2.1, the present manifestation of negative reciprocity also leads to a failure of the Revelation Principle. We defer the example illustrating this failure to Section 2.2.1 and turn instead to the following question: can some simplification of the domain of optimal signals still be recovered? Our next result shows that the answer is positive. Receiver's extreme beliefs continue to play a central role in characterizing the optimal signal in this environment.

The key difference, however, is that these beliefs are now *endogenous*. In particular, attempting to replicate the proof strategy of Proposition 1 and to "split" a posterior into extreme beliefs reveals a new complication: because Receiver's preferences depend on ρ, the extreme beliefs themselves may shift with the signal's inaccuracy. Nevertheless, we can establish a result analogous to Proposition 1.

To state our next result, we introduce the following definition. For any $\varepsilon > 0$, a signal σ_ε is said to be ε-sender-optimal if

$$V(\sigma_\varepsilon) \geq \sup_\sigma V(\sigma) - \varepsilon.$$

We then have:

Proposition 5 *For any $\varepsilon > 0$, there exists an ε-sender-optimal signal σ^*_ε that is supported on its corresponding set of extreme beliefs; that is,*

$$M_{\sigma^*_\varepsilon} \subseteq \mathcal{E}_{\rho(\sigma^*_\varepsilon)} \equiv \bigcup_a \mathrm{Ext}(B^*_{\rho(\sigma^*_\varepsilon)}(a)).$$

Moreover, if a Sender-optimal signal exists, then there also exists a Sender-optimal signal that is supported on its corresponding set of extreme beliefs.

The proof shows that for any $\varepsilon > 0$ there exists a minimal inaccuracy level $\underline{\rho}$ that is attained by some ε-sender-optimal signal. Starting from such a signal, applying the splitting argument of Proposition 1 weakly decreases the signal's inaccuracy and weakly improves Sender's payoff. The latter follows because, for any posterior belief that is split, Equation (9) implies that the actions Receiver takes at the corresponding extreme beliefs can only be more favorable to Sender. As a result, the modified signal remains ε-sender-optimal. By the minimality of $\underline{\rho}$, the inaccuracy of the modified signal must equal $\underline{\rho}$, implying that Receiver's preferences do not change and, therefore, that the set of extreme beliefs remains unchanged. Consequently, the modified signal is supported entirely on extreme beliefs. The argument for optimal signals is analogous.

Unlike Proposition 5, which highlights an analogy in the properties of optimal signals under the current specification of negative reciprocity and the one explored in Section 2.1, our next result shows that the two manifestations lead to different conclusions regarding whether truth telling can be an optimal strategy for Sender. While under the specification in which Receiver's negative reciprocity is manifested through ignoring a manipulative signal, truth telling may be optimal for sufficiently manipulation-averse Receivers (Proposition 4), under the current specification with endogenous preferences truth telling is *never* optimal.

Proposition 6 *The truthful signal σ_T is never optimal for Sender.*

The idea of the proof is as follows. Starting from a truthful signal (which assigns positive probabilities only to pure posteriors) Sender can slightly shift probability mass from one pure posterior to a “nearby” posterior, introducing only negligible inaccuracy to the signal. This small adjustment doesn’t change Receiver’s actions (due to the continuity of Receiver’s preferences in the inaccuracy measure) but it reallocates probability mass in favor of Sender’s most preferred action, thereby increasing Sender’s expected payoff.

2.2.1 Example: optimal signal with two states

Suppose there are two states, $\Omega = \{\omega_1, \omega_2\}$, with a prior probability $q = 0.25$ on the state ω_1. There are two actions for Receiver, $A = \{a_1, a_2\}$. Sender’s utility depends solely on Receiver’s action, with $u_S(a_1) = 1$ and $u_S(a_2) = 0$.

Receiver’s utility depends on the state, the chosen action and the signal’s inaccuracy. Specifically, Receiver earns a material payoff of 1 if his action matches the state (i.e., a_i in state ω_i), and 0 otherwise. However, his overall utility, which is influenced by feelings of negative reciprocity, is equal to his material payoff multiplied by $1 - \delta\rho(\sigma)$ when he chooses Sender’s optimal action a_1, where σ is the signal chosen by Sender.

Under this specification, Receiver’s utility from Sender’s preferred action a_1 decreases with the signal’s inaccuracy. The “textbook” persuasion problem corresponds to the case where $\delta = 0$. In this example we set $\delta = 5$.

By proposition 5, Sender’s optimal signal is supported on at most three posterior beliefs (probabilities on ω_1): $\{0, \mu, 1\}$, where μ is an extreme posterior at which Receiver is indifferent between the two actions. Therefore, Sender’s problem reduces to finding the posterior μ, and a distribution (p_0, p_μ, p_1) over posteriors that maximizes the probability that Receiver selects a_1. Formally, Sender’s problem is given by:

$$\max_{\sigma=(\mu,p_0,p_\mu,p_1)} p_\mu + p_1$$

$$s.t. \quad p_0 + p_\mu + p_1 = 1 \tag{PR}$$

$$\mu p_\mu + p_1 = 0.25 \tag{BP}$$

$$\mu = \frac{1}{2 - 5\rho(\sigma)} \tag{EXT}$$

The (BP) constraint represents the Bayes plausibility condition. The (EXT) constraint is implied by the indifference condition at μ which makes it an extreme posterior.

From (BP) we can substitute $p_1 = 0.25 - \mu p_\mu$ into the objective function. From (INAC) it follows that $\rho(\sigma) = p_\mu H(\mu)/H(0.25)$, hence we can use constraint (EXT) to express p_μ as a function of μ, which we can substitute into the objective function to obtain the following expression:

$$0.25 + (1-\mu) \cdot \left(\frac{0.2\left(2 - \frac{1}{\mu}\right) H(0.25)}{-\mu \log \mu - (1-\mu)\log(1-\mu)} \right)$$

Maximizing this expression yields that $\mu \approx 0.8$, and hence the optimal signal is (approximately) given by the distribution $(0.71, 0.17, 0.12)$ over the posteriors $(0, 0.8, 1)$.

It is useful to compare this optimal signal to two benchmarks. First, the optimal signal in "textbook" persuasion model with $\delta = 0$ is given by the distribution $(0.5, 0.5)$ over the posteriors $(0, 0.5)$. Second, the optimal signal when Sender is restricted to only give action recommendations (i.e. a binary signal) is given by the distribution $(0.72, 0.28)$ over the posteriors $(0, 0.888)$. This signal is *worse* for Sender than the optimal signal characterzied above, which demonstrates again the failure of the revelation principle. However, it is better for Sender than truth-telling.

2.3 A Comment on Measuring Inaccuracy

A key element of our framework is that negative reciprocity depends on the "distance" between the signal and truth-telling. To quantify this distance, an appropriate index is required. A natural choice is one that aligns with the partial Blackwell ordering. Since information theory provides well-established measures for this purpose, we adopt the canonical index of mutual information. However, Propositions 1 and 5 remain valid under any alternative index consistent with Blackwell ordering. Naturally, the specific form of optimal signals will depend on the chosen index.

Depending on the context, a potential drawback of the inaccuracy measure ρ is that for a pair of signals that are not ordered according to Blackwell informativeness, one signal may be more inaccurate according to ρ than another yet still yield a higher payoff for Receiver. An alternative approach would be to assume that negative reciprocity intensifies with the distance between Receiver's material payoff under truth-telling and his payoff under the signal. By definition, this utility-based measure is consistent with the Blackwell ordering. However, adopting a utility-based measure presents a conceptual challenge when Receiver's utility is endogenous, as in Section 2.2. The difficulty lies in determining the appropriate utility function to measure the deviation from Receiver's first-best payoff. In the next Section we use an inaccuracy measure that is both consistent with Blackwell ordering and has the property that a more inaccurate Sender strategy is strictly worse for Receiver.

3 Cheap-Talk

To demonstrate a different environment in which Receiver's adversarial response to manipulated information may matter, we now consider a variation of the model of Crawford and Sobel (1982). We begin by presenting the modified framework. In Section 3.1 we demonstrate how equilibria can be characterized when Receiver's preference are endogenous and we highlight the properties that are analogous to the CS benchmark. In Sections 3.2 and 3.3 we show that negative reciprocity can generate equilibria with novel features that cannot arise in the CS benchmark.

A state ω is drawn from $\Omega = [0, 1]$ according to a distribution F with density f. Sender observes the state and sends a message to Receiver. After observing the message, Receiver chooses an action $a \in \mathbb{R}$. Sender's utility is given by $u_S(a, \omega) = -(a - \omega - b)^2$, where $b > 0$ captures Sender's bias toward higher actions relative to the state. Sender's strategy is a mapping from states to messages in some set M. We focus on Sender's pure strategies, and denote such a strategy by $\sigma : \Omega \to M$.

In this setting, we adopt a different inaccuracy measure. Specifically, given a strategy σ, we define its *inaccuracy* as the expected posterior variance of the induced distribution over states:

$$\rho(\sigma) = \mathbb{E}_\sigma[\mathrm{Var}_\sigma[\omega \mid m]] \,, \tag{10}$$

This measure captures the idea that a more informative signal induces, on average, posterior beliefs with lower variance. Accordingly, $\rho(\sigma)$ quantifies the residual uncertainty that remains after the message is observed. The expectation and variance on the right-hand side are taken with respect to the strategy σ and the underlying type distribution F. For brevity, we suppress the dependence on F in the notation.

Departing from the specification of Crawford and Sobel (1982) (hereafter the *CS benchmark*), we assume that Receiver's utility depends on the (in)accuracy of σ. Maintaining the quadratic-loss form, let:

$$u_R(a, \omega; \rho) = -(a - \omega + d(\rho(\sigma)))^2, \tag{11}$$

where $d : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing, continuous *reciprocity function* satisfying $d(0) = 0$. Thus, the less informative (i.e., higher ρ) the communication strategy, the further Receiver's preferred action diverges from Sender's ideal point. Receiver's strategy is denoted by $\alpha : M \to \mathbb{R}$. This formulation introduces a *strategic feedback loop*: Sender's strategy affects Receiver's bias, which in turn shapes Sender's optimal signal choice.

The solution concept is Perfect Bayesian Equilibrium (PBE).[10] Given an equilibrium strategy σ^* for Sender, standard arguments imply that Receiver's best response is given by:

$$\alpha^*(m) = \mathbb{E}_{\sigma^*}[\omega \mid m] - d(\rho(\sigma^*)). \tag{12}$$

Thus, Receiver chooses the expected state but offsets the action downward (opposite to Sender's preferred direction) by an amount that increases with the inaccuracy of Sender's strategy. Sender's expected utility from employing σ^*, and Receiver's expected utility from best responding to σ^* in equilibrium, are given by

$$\mathbb{E}u_S(\sigma^*) = -\rho(\sigma^*) - \big(b + d(\rho(\sigma^*))\big)^2, \tag{13}$$

$$\mathbb{E}u_R(\sigma^*) = -\rho(\sigma^*). \tag{14}$$

Thus, the inaccuracy measure is aligned with Receiver's disutility from Sender's manipulation.

[10] Our focus on pure strategies is made for simplicity. This means that there are messages that may not be sent with positive probability. However, one could always replace the pure strategy with a behavioral strategy with the same inaccuracy that induces the same best response by Receiver, where each message is sent by some positive measure of types. Because of this, we will not specify Receiver's out-of-equilibrium beliefs and focus only on the players' strategies.

3.1 Equilibria with an Endogenous Receiver Bias

We call a strategy σ an *interval strategy* if it partitions $\Omega = [0,1]$ into intervals such that all types within each interval send the same message. An equilibrium is an *interval equilibrium* if Sender employs an interval strategy. Our next result shows that there is no loss of generality in focusing on interval equilibria. Its proof relies on the specific inaccuracy measure in Eq. (10), and follows from the law of total variance.

Proposition 7 *For any Perfect Bayesian Equilibrium (σ, α), there exists an interval equilibrium (σ', α') such that (i) $\alpha(\sigma(\omega)) = \alpha'(\sigma'(\omega))$ for all $\omega \in \Omega$, and (ii) $\rho(\sigma') = \rho(\sigma)$.*

A trivial interval equilibrium always exists: the *babbling equilibrium*, in which all Sender types pool on a single message. The existence of other equilibria, particularly Sender-optimal ones, is more subtle. To establish their existence, it is useful to relate equilibria under reciprocity to those in the CS benchmark.

Let $\phi(s)$ denote the minimal expected posterior variance in the CS benchmark with bias s. By Crawford and Sobel (1982), ϕ is well-defined and nondecreasing. For bias b and reciprocity function $d(\cdot)$, define

$$T_{b,d}(\rho) = \phi\big(b + d(\rho)\big), \quad \rho \in [0, \mathrm{Var}(\omega)]. \tag{15}$$

Observation 1 *Any fixed point $\hat{\rho}$ satisfying $\hat{\rho} = T_{b,d}(\hat{\rho})$ corresponds to the inaccuracy of some equilibrium in our model.*

The reason is as follows. The same interval partition that supports an equilibrium in the CS benchmark with effective bias $b+d(\hat{\rho})$ also supports an equilibrium in our model with bias b and a reciprocity penalty equal to $d(\hat{\rho})$. In both environments, each Sender type ω obtains the same utility from revealing the interval to which she belongs.

Equation (15) is useful because the existence of fixed points of $T_{b,d}$ follows from the Knaster-Tarski fixed-point theorem, since both $\phi(\cdot)$ and $d(\cdot)$ are monotone.[11] Moreover, a least fixed point exists, and this fixed point corresponds to the expected posterior variance in the Sender-optimal equilibrium:

[11] See, e.g., Theorem 2.35 in Davey and Priestley (2002).

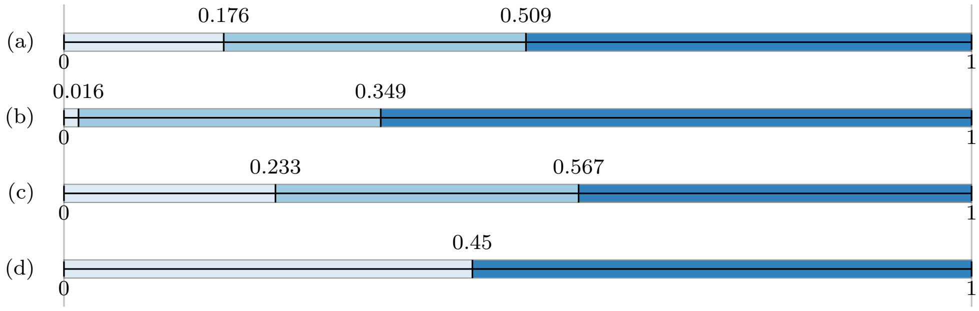

Figure 3: Interval equilibria with and without negative reciprocity.

Proposition 8 *A Sender-optimal equilibrium exists. In particular, Sender employs finitely many messages, and the equilibrium's expected posterior variance equals the least fixed point of $T_{b,d}$.*

In equilibrium, Receiver's reciprocity has two opposing effects: it rewards informativeness but amplifies the conflict of interest. Nonetheless, the fixed-point characterization delivers a clean comparative statics result: *Stronger* reciprocity concerns lead Sender to reveal *less*. Let $\rho^*_{(b,d)}$ denote the expected posterior variance in the Sender-optimal equilibrium with bias b and reciprocity $d(\cdot)$. Then,

Proposition 9 *Fix $b > 0$. Let $d_1, d_2 : \mathbb{R}_+ \to \mathbb{R}_+$ be reciprocity functions with $d_1(0) = d_2(0) = 0$ and $d_1(\rho) \geq d_2(\rho)$ for all $\rho > 0$. Then $\rho^*_{(b,d_1)} \geq \rho^*_{(b,d_2)}$. Consequently, Sender's expected utility is greater under d_2 than under d_1.*

3.2 Novel Implications of Negative Reciprocity

The endogeneity of Receiver's preferences generates equilibrium patterns that differ from those in the CS benchmark. We highlight two such differences in the observations below.

Observation 2 (Multiplicity of Equilibria with the same number of cells) *Unlike in the CS benchmark, Receiver reciprocity permits multiple equilibria with the same number of cells, which differ in the expected posterior variance they induce.*

Example 3. To illustrate this multiplicity, suppose the state is uniformly distributed on $[0, 1]$, with Sender bias $b = 1/40$ and reciprocity function $d(\rho) = 80\rho^2$. Under these parameters, both partitions in Figures 3(a) and 3(b) constitute 3-cell equilibria. The former yields lower expected variance ($\rho_{(a)} \approx 0.013$) and more balanced cells; the latter, a higher variance ($\rho_{(b)} \approx 0.026$) and more asymmetric cells. For comparison, Figure 3(c) shows the unique 3-cell equilibrium in the CS benchmark, corresponding to $d(\rho) = 0$.

Observation 3 (non-monotonicity of equilibrium existence) *Unlike in the CS benchmark, Under reciprocity the existence of an N-cell equilibrium need not imply existence of an $(N-1)$-cell equilibrium.*

In the parameters of Example 3, the CS benchmark admits both 2- and 3-cell equilibria (Figures 3(c)–3(d)), whereas under reciprocity, the 3-cell partitions in Figures 3(a)–3(b) are equilibria, but no 2-cell equilibrium exists.

3.3 A Comment on Intrinsic Preferences for Information

Proposition 7 relies on the fact that the inaccuracy measure we employ, which is based on the expected variance of Receiver's information structure, is only *weakly* monotone with respect to Blackwell informativeness. In particular, two information structures may have the same expected variance even if one strictly Blackwell dominates the other. For example, suppose the state is uniformly distributed on $[0,1]$. Compare the information structure under which Receiver knows whether the state is in $[\frac{1}{3}, \frac{2}{3}]$ or not, with the information structure under which he learns nothing. The former strictly Blackwell dominates the latter, yet both induce the same expected variance. This follows from the Law of Total Variance.

However, if Receiver has intrinsic preferences for more informative signals (as in Grant et al., 1998), he may be less upset when Sender uses a more precise strategy, even if the additional precision is non-instrumental. As we show next, this may fundamentally change the properties of the equilibrium.

Consider an inaccuracy measure ρ that strictly decreases with the Blackwell informativeness of Sender's strategy. In addition, assume that ρ is continuous under small refinements of the induced information structure, and that when the partition becomes sufficiently fine the inaccuracy measure decreases towards zero.[12]

[12] One way to formalize such continuity is through collision probabilities, where two information structures (partitions) are considered "close" if they fail to distinguish types with similar probabilities. Let ν_F denote the probability measure induced by the distribution F. For any finite partition $\mathcal{P}$ induced by a strategy σ, define $\kappa(\mathcal{P}) = \sum_{P \in \mathcal{P}} \nu_F(P)^2$. Thus, $\kappa(\mathcal{P})$ represents the probability that two independent draws from F fall into the same partition element (i.e.,

Indeed, in such a case, Sender has an incentive to cater to Receiver's demand for information, even if the additional information is not more instrumental. This leads to the following result:

Proposition 10 *Suppose that the reciprocity function $d(\cdot)$ is strictly increasing and continuous, and that the inaccuracy measure ρ satisfies (i) strict Blackwell monotonicity, (ii) continuity under small refinements, and (iii) it vanishes as the partition gets sufficiently fine. Then, there exists no Sender-optimal equilibrium in which Sender employs a finite number of messages.*

This result is in sharp contrast to one of the key features in the CS benchmark, where any equilibrium is supported by finitely many messages.

To see the intuition for the proof, recall the example from the opening paragraph of this section. In this example, Sender "splits" the uninformative signal into a Blackwell dominating signal that preserves posterior means (and thus the action in the CS benchmark), while lowering the inaccuracy and thereby reducing the reciprocity-induced bias. This "procedure" of splitting a signal in a non-instrumental allows us to construct an equilibrium that is more favorable for Sender.

4 Bayes Correlated Equilibrium

We now turn to a setting in which a single sender interacts with multiple receivers who respond adversarially to signals they perceive as manipulative. Our objective is to show that, in the presence of multiple receivers, equilibrium forces can generate a counterintuitive effect: even when receivers' reactions to perceived information manipulation make the sender's preferred action profiles strictly less attractive to them, such reactions may nonetheless *increase* the ex-ante probability that the sender's preferred action profile is played (relative to a benchmark in which receivers are indifferent to information manipulation). As a result, the sender's ex-ante expected payoff may increase. This phenomenon stands in sharp contrast to the outcomes obtained in the single-receiver settings analyzed in the previous sections. We illustrate the effect through a simple example using the

they "collide"). Continuity under small refinements requires that for every partition $\mathcal{P}$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that for every refinement $\mathcal{Q}$ of $\mathcal{P}$, if $\kappa(\mathcal{P}) - \kappa(\mathcal{Q}) < \delta$ then $|\rho(\mathcal{Q}) - \rho(\mathcal{P})| < \varepsilon$. In addition, the property that ρ vanishes as the partition gets sufficiently fine implies that, given a sequence $\{P_i\}_{i\in\mathbb{N}}$, if $\kappa(P_i) \to 0$ then $\rho(P_i) \to 0$.

solution concept of Bayes correlated equilibrium (BCE; Bergemann and Morris, 2016).[13]

There are two players (receivers), indexed by $i \in \{1, 2\}$. Each player i chooses an action $a_i \in \{0, 1\}$. There are two states of the world, $\omega \in \Omega = \{0, 1\}$, which occur with equal prior probability.

A mediator (sender) designs a signal and sends private, potentially correlated, action recommendations to the players. Her objective is to maximize the probability that the action profile $(0, 0)$ is played at state 0 and that $(1, 1)$ is played at state 1. Formally, she aims to maximize $V = \frac{1}{2}\Pr[(0, 0) \mid \omega = 0] + \frac{1}{2}\Pr[(1, 1) \mid \omega = 1]$.

The players dislike manipulative signals. Let ρ_i denote the inaccuracy of the *private* signal received by player i, defined as in Equation (INAC) above. Specifically, $\rho_i = 1 - \mathrm{I}(S_i, 0.5)/H(0.5)$, where $\mathrm{I}(S_i, 0.5)$ denotes the mutual information between the random variable representing player i's private signal and the random variable representing the state, and $H(0.5)$ is the entropy of the prior.

The players' payoffs in each state are as follows.

	$\boldsymbol{a_2 = 0}$	$\boldsymbol{a_2 = 1}$
$\boldsymbol{a_1 = 0}$	$(1 - \lambda\rho_1,\ 4 - \lambda\rho_2)$	$(2,\ 4)$
$\boldsymbol{a_1 = 1}$	$(3,\ 1)$	$(4,\ 0)$

Payoffs at $\omega = 0$

	$\boldsymbol{a_2 = 0}$	$\boldsymbol{a_2 = 1}$
$\boldsymbol{a_1 = 0}$	$(4,\ 2)$	$(2,\ 5)$
$\boldsymbol{a_1 = 1}$	$(3,\ 3)$	$(2 - \lambda\rho_1,\ 1 - \lambda\rho_2)$

Payoffs at $\omega = 1$

Note that the players' payoffs in the mediator-preferred action profiles, that is $(0, 0)$ at $\omega = 0$ and $(1, 1)$ at $\omega = 1$, are reduced by $\lambda\rho_i$, where $\lambda \geq 0$ captures the intensity of the preferences response to manipulation.

Benchmark: Players indifferent to information manipulation. Suppose first that $\lambda = 0$, so that the adverse response to manipulation is inactive. Consider a mediator who fully reveals the state. When $\omega = 0$, the only recommendation that the players will follow is $(a_1, a_2) = (1, 0)$; in particular, the mediator cannot persuade them to play her preferred profile $(0, 0)$. When $\omega = 1$, the mediator recommends Player 1 to play $a_1 = 0$ with probability 0.4 and $a_1 = 1$ with probability 0.6. This mixture makes Player 2 indifferent between his two actions, so he is willing to follow the recommendation $a_2 = 1$. In this case, the action profile $(1, 1)$ is played with probability 0.6. Thus, the mediator's expected payoff is $V_{\lambda=0,\text{reveal}} = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0.6 = 0.3$.

[13] For related work on the information designer's optimization problem in general Bayesian games, see Mathevet et al. (2020).

In fact, it can be shown that 0.3 is the highest expected payoff that the mediator can achieve in any Bayes correlated equilibrium of this game.

Lemma 3 *If $\lambda = 0$, then in any BCE of the game, $V \leq 0.3$. Hence, full revelation of the state is optimal.*

The proof works by taking an appropriate linear combination of the obedience constraints and the prior constraint to construct an inequality that directly upper-bounds the mediator's payoff.

Players who dislike manipulated information. Suppose now that $\lambda = 1.1$.[14] We will construct a Bayes correlated equilibrium of this game.

Consider the following signal-and-recommendation scheme: If $\omega = 0$, the mediator recommends $(a_1, a_2) = (1, 0)$ with probability 5/6 and $(a_1, a_2) = (0, 0)$ with probability 1/6; if $\omega = 1$, she recommends $(a_1, a_2) = (0, 1)$ with probability 1/2 and $(a_1, a_2) = (1, 1)$ with probability 1/2.

Under this signal-and-recommendation scheme, Player 2 perfectly infers the state, and so $\lambda\rho_2 = 0$. By contrast, Player 1 receives a noisy signal about the state, and computation shows that $\lambda\rho_1 = 1$.

For Player 1, obeying the recommendation is a best response. When recommended $a_1 = 1$, she assigns probability 5/8 to $(\omega = 0, a_2 = 0)$ and the remaining probability 3/8 to $(\omega = 1, a_2 = 1)$, under which $a_1 = 1$ is a best response. When recommended $a_1 = 0$, she assigns probability 1/4 to $(\omega = 0, a_2 = 0)$ and the remaining probability 3/4 to $(\omega = 1, a_2 = 1)$, under which $a_1 = 0$ is (weakly) a best response. Hence, following the recommendation is incentive compatible for Player 1. By similar reasoning, Player 2 also finds it optimal to obey. Therefore, this information structure and recommendations indeed constitute an equilibrium.

Under this equilibrium, the probability that $(0, 0)$ is played in state 0 is 1/6, whereas the probability that $(1, 1)$ is played in state 1 is 1/2. Consequently, $V_{\lambda,\text{noisy}} = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{3}$, which exceeds the mediator's maximal expected payoff under $\lambda = 0$, which was 0.3.

The example therefore shows that the mediator may prefer to interact manipulatively with players who respond adversarially to manipulation, such that their

[14] More precisely, for the computation in this example we require $\lambda = \Big(- \frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} \Big) \cdot \Big(\frac{2}{3} \big(- \frac{5}{8} \log \frac{5}{8} - \frac{3}{8} \log \frac{3}{8} \big) + \frac{1}{3} \big(- \frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} \big) \Big)^{-1} \approx 1.1029$.

preferences shift in ways that render the mediator's favored action profiles less appealing, rather than with players who remain indifferent to manipulation.

A Proofs

Proof of Lemma 1

Define:

$$g(\mu_2) = \frac{(q-\mu_2)H(\mu_1) + (\mu_1 - q)H(\mu_2)}{\mu_1 - \mu_2}.$$

Differentiating $g(\mu_2)$ gives:

$$g'(\mu_2) = \frac{(\mu_1 - q)[(\mu_1 - \mu_2)H'(\mu_2) + H(\mu_2) - H(\mu_1)]}{(\mu_1 - \mu_2)^2}.$$

Since $\mu_1 > q$, the sign of $g'(\mu_2)$ depends on $\Delta := (\mu_1 - \mu_2)H'(\mu_2) + H(\mu_2) - H(\mu_1)$. Because $H(\cdot)$ is strictly concave, its tangent at μ_2 lies above the graph:

$$H(\mu_1) < H(\mu_2) + H'(\mu_2)(\mu_1 - \mu_2).$$

Rearranging gives $\Delta > 0$, implying $g'(\mu_2) > 0$, which completes the proof. □

Proof of Lemma 2

Restricting attention to signals with $\mu_2 = 0$, the expected payoff in Equation (7) can be written as follows:

$$V(\sigma) = \frac{q}{\mu_1}\left(1 - \lambda \frac{q}{H(q)} \frac{H(\mu_1)}{\mu_1}\right). \tag{16}$$

Accordingly, to prove the result it suffices to show that the maximum of the function

$$f(\mu) = \frac{1}{\mu}\left(1 - K \cdot \frac{H(\mu)}{\mu}\right), \quad K > 0,$$

on the interval $\mu \in [0.5, 1]$ is attained at one of the endpoints, namely $\mu = 0.5$ or $\mu = 1$.

Since f is continuous on the compact interval $[0.5, 1]$, it must attain a maximum. Suppose, towards a contradiction, that this maximum occurs at an interior point $\mu^* \in (0.5, 1)$. Then the derivative of f must vanish at μ^*, that is,

$$f'(\mu^*) = \frac{-\mu^* - K\mu^* H'(\mu^*) + 2KH(\mu^*)}{(\mu^*)^3} = 0.$$

Solving for K, we obtain

$$K = \frac{\mu^*}{-\mu^* H'(\mu^*) + 2H(\mu^*)}.$$

Substituting this expression into $f(\mu^*)$, we find

$$f(\mu^*) = \frac{\mu^* - KH(\mu^*)}{(\mu^*)^2} = \frac{H(\mu^*) - \mu^* H'(\mu^*)}{2\mu^* H(\mu^*) - (\mu^*)^2 H'(\mu^*)}.$$

Because $\mu^* \in (0.5, 1)$, we have:[15]

$$2\mu^* H(\mu^*) - (\mu^*)^2 H'(\mu^*) > H(\mu^*) - \mu^* H'(\mu^*) = -\ln(1 - \mu^*) > 0.$$

It follows that

$$f(\mu^*) < \frac{H(\mu^*) - \mu^* H'(\mu^*)}{H(\mu^*) - \mu^* H'(\mu^*)} = 1 = f(1).$$

Thus, the value of f at the endpoint $\mu = 1$ is strictly greater than its value at the interior point μ^* where the derivative of f vanishes, contradicting the assumption that μ^* is a maximizer of f on $[0.5, 1]$. We conclude that the maximum of f on $[0.5, 1]$ is attained at either $\mu = 0.5$ or $\mu = 1$.

Proof of Proposition 1

Proof. Suppose that a signal σ generates a posterior belief μ_1 with probability $p_1 = p_\sigma(\mu_1)$, where $\mu_1 \in B^*(a_1)$ but $\mu_1 \notin \text{Ext}(B^*(a_1))$. Then, μ_1 can be expressed as a convex combination of beliefs $\{\mu_1^E, \ldots, \mu_J^E\} \subseteq B^*(a_1)$. That is, $\mu_1 = \sum_{j=1}^J \alpha_j \mu_j^E$, for some $\alpha_j \geq 0$ satisfying $\sum_{j=1}^J \alpha_j = 1$.

[15] The first inequality follows from a direct computation. Subtracting the right-hand side from the left-hand side yields

$$g(\mu) = \left[2\mu H(\mu) - \mu^2 H'(\mu)\right] - \left[H(\mu) - \mu H'(\mu)\right] = (1-\mu)^2 \ln(1-\mu) - \mu^2 \ln \mu.$$

The function $g(\mu)$ approaches zero as $\mu \to 0.5$ and as $\mu \to 1$. Moreover, g is concave, since

$$g''(\mu) = 2\ln\left(\frac{1-\mu}{\mu}\right) < 0 \quad \text{for all } \mu \in (0.5, 1).$$

Therefore, $g(\mu) > 0$ for all $\mu \in (0.5, 1)$.

Consider now an alternative signal σ' that "splits" the belief μ_1 into the beliefs in $B^*(a_1)$. That is, instead of generating the posterior belief μ_1 with probability p_1, the signal σ' generates the posterior beliefs $(\mu_1^E, \ldots, \mu_J^E)$ with probabilities $(\alpha_1 p_1, \ldots, \alpha_J p_1)$, respectively, and is otherwise identical to σ. Bayes plausibility ensures the existence of such a signal. Because a_1 is optimal for Receiver under every belief μ_j^E, the distribution over Receiver's actions (assuming he does not disregard the signal) becomes weakly more favorable to Sender (and strictly so if Receiver changes his action, breaking an indifference in favor of Sender).

However, the probability that Receiver disregards the information under the new signal σ' is weakly smaller than under σ:

$$\begin{aligned}\gamma(\sigma) &= \frac{\lambda}{H(q)}\left(p_1 H(\mu_1) + \sum_{\mu \in M(\sigma)\setminus\{\mu_1\}} p_\sigma(\mu) H(\mu)\right)\\ &\geq \frac{\lambda}{H(q)}\left(\sum_{j=1}^{J}(\alpha_j p_1) H(\mu_j^E) + \sum_{\mu \in M(\sigma)\setminus\{\mu_1\}} p_\sigma(\mu) H(\mu)\right)\\ &= \gamma(\sigma'),\end{aligned}$$

where the inequality follows from the concavity of the Shannon entropy. ■

Proof of Proposition 2

The result follows directly from the *Upper Bound Theorem* for convex polytopes (see, e.g., Theorem 18.1 in Brøndsted (1983), p. 113), which provides an upper bound on the number of vertices of a d-dimensional convex polytope as a function of the number of its facets.

In the present setting, viewing beliefs as elements of the probability simplex $\Delta(\Omega) \subset \mathbb{R}^L$, for each action a, the set $B^*(a)$ is defined by $L + K - 1$ linear inequalities: (i) $L-1$ inequalities that ensure the non-negativity of each coordinate μ_i for any $\mu \in B^*(a)$; (ii) one inequality that ensures that $\sum_{i=1}^{L-1} \mu_i \leq 1$; and (iii) $K - 1$ inequalities ensuring that action a is preferred to any other available action. Accordingly, $B^*(a)$ is an $(L-1)$-dimensional convex polytope with *at most* $L + K - 1$ facets. The Upper Bound Theorem then implies that the number of vertices of $B^*(a)$ is bounded from above by Φ, as defined in Equation (8). Since there are K actions, the total number of induced posteriors is at most $K \cdot \Phi$.

Proof of Proposition 3

Throughout the proof, we assume that $W_\varnothing \geq 0$. The argument for the case $W_\varnothing < 0$ is analogous and therefore omitted.

Since σ and σ' are optimal signals, we have $W(\sigma) \geq W_\varnothing$ and $W(\sigma') \geq W_\varnothing$, reflecting that optimal communication is weakly preferred to no communication.

For the proof, it will be useful to slightly update the notation in Equation (6) to reflect the fact that the Sender's expected payoff from employing signal σ depends on the parameter λ. Accordingly, for any λ we write

$$V_\lambda(\sigma) = \lambda\rho(\sigma)W_\varnothing + \big(1 - \lambda\rho(\sigma)\big)W(\sigma).$$

Since σ' is optimal for λ', we have $V_{\lambda'}(\sigma') \geq V_{\lambda'}(\sigma)$.

Assume, toward a contradiction, that $\rho(\sigma') > \rho(\sigma)$. The payoff $V_{\lambda'}(\sigma)$ is a weighted average of $W_\varnothing$ and the higher payoff $W(\sigma)$. Similarly, the payoff $V_{\lambda'}(\sigma')$ is a weighted average of $W_\varnothing$ and the higher payoff $W(\sigma')$. Because, by assumption, $\lambda'\rho(\sigma') > \lambda'\rho(\sigma)$, the inequality $V_{\lambda'}(\sigma') \geq V_{\lambda'}(\sigma)$ requires that $W(\sigma') \geq W(\sigma)$.

Define the ratio $R(\lambda) \equiv \frac{V_\lambda(\sigma')}{V_\lambda(\sigma)}$, so $R(\lambda') \geq 1$. Differentiating $R(\lambda)$ with respect to λ and substituting the expressions for $V_\lambda(\cdot)$ and V'_λ yields $\frac{N}{(V_\lambda(\sigma))^2}$, where the numerator N is an expression independent of λ:

$$\begin{aligned}
N &= W_\varnothing\left[\rho(\sigma')W(\sigma) - \rho(\sigma)W(\sigma')\right] + W(\sigma)W(\sigma')\left[\rho(\sigma) - \rho(\sigma')\right] \\
&\leq W_\varnothing\left[\rho(\sigma')W(\sigma') - \rho(\sigma)W(\sigma')\right] + W(\sigma)W(\sigma')\left[\rho(\sigma) - \rho(\sigma')\right] \\
&= \left(W_\varnothing W(\sigma') - W(\sigma)W(\sigma')\right)\left(\rho(\sigma') - \rho(\sigma)\right) \\
&= W(\sigma')\left(W_\varnothing - W(\sigma)\right)\left(\rho(\sigma') - \rho(\sigma)\right) \\
&\leq 0.
\end{aligned}$$

The first inequality holds because $W(\sigma) \leq W(\sigma')$, as established above. The final inequality follows because $W_\varnothing \leq W(\sigma)$ and the assumption $\rho(\sigma') > \rho(\sigma)$.

Since $N \leq 0$, the ratio $R(\lambda)$ is a decreasing function of λ. Given $\lambda < \lambda'$, the decreasing property of $R(\lambda)$ implies:

$$R(\lambda) = \frac{V_\lambda(\sigma')}{V_\lambda(\sigma)} > R(\lambda') \geq 1.$$

The resulting inequality, $V_\lambda(\sigma') > V_\lambda(\sigma)$, contradicts the premise that σ is Sender-optimal under the parameter λ. ■

Proof of Proposition 4

Part I. By Proposition 1, we restrict attention to signals that induce posterior beliefs in $\mathcal{E}$. Let $\underline{H}$ denote the minimal entropy among all *non-pure* posteriors in this set normalized by $/H(q)$, i.e.,

$$\underline{H} = \min_{\mu \in \mathcal{E} \setminus \{e_1, \ldots, e_L\}} \frac{H(\mu)}{H(q)}.$$

Since $\mathcal{E}$ is finite, this minimum exists and is strictly positive.

Consider an inaccurate signal σ that yields a higher expected payoff for Sender than truth-telling. By Equation (2) we have:

$$W(\sigma) > W_T > W_{\varnothing}. \tag{17}$$

It follows that $(1 - \lambda\rho(\sigma)) > 0$. Additionally, because σ is inaccurate, it induces at least one non-pure posterior, which we denote by $\mu = (\mu_1, \ldots, \mu_L)$, with positive probability $p > 0$.

Now, consider an alternative signal that is identical to σ except that it induces the posterior μ with probability $p - \varepsilon$ and increases the probability of each pure posterior e_i by $\varepsilon\mu_i$.

On the one hand, the modified signal is more accurate than σ. The difference in inaccuracy, as measured by γ, is bounded from below by $\varepsilon\lambda\underline{H}$. This follows from the fact that the contribution of the pure posteriors $(e_1, \ldots, e_L)$ to the total inaccuracy is zero, while the reduction in inaccuracy resulting from the decreased probability assigned to the posterior μ is at least $\varepsilon\lambda H(\mu)/H(q) \geq \varepsilon\lambda\underline{H}$.

On the other hand, the modified signal induces a probability distribution over actions that may be less favorable than the action distribution induced by σ. The difference in Sender's expected payoff, conditional on Receiver not ignoring the signal, is bounded from above by $\varepsilon\Delta u$, where Δu denotes the difference in utilities between Sender's most and least preferred actions.

Therefore, the difference in Sender's expected payoff from switching from σ to the modified signal is bounded from below by

$$\begin{aligned}&\overbrace{\left((1-\lambda\rho(\sigma)+\varepsilon\lambda\underline{H}\right)\left(W(\sigma)-\varepsilon\Delta u\right)+(\lambda\rho(\sigma)-\varepsilon\lambda\underline{H})W_{\varnothing})}^{\text{lower bound on expected payoff from employing the modified signal}}\\ &-\underbrace{((1-\lambda\rho(\sigma))W(\sigma)+(\lambda\rho(\sigma))W_{\varnothing})}_{\text{expected payoff from employing }\sigma}\end{aligned}$$

Simplifying this expression and omitting terms of order $\mathcal{O}(\varepsilon^2)$, we obtain:

$$\begin{aligned}-\varepsilon(1-\lambda\rho(\sigma))\Delta u+\varepsilon\lambda\underline{H}W(\sigma)-\varepsilon\lambda\underline{H}W_{\varnothing} &> \varepsilon\left(\lambda\underline{H}\left(W(\sigma)-W_{\varnothing}\right)-\Delta u\right)\\ &> \varepsilon\left(\lambda\underline{H}\left(W_T-W_{\varnothing}\right)-\Delta u\right) \qquad (18)\end{aligned}$$

where the last inequality follows from Equation (17).

By Equation (17), we also have that $(W_T - W_{\varnothing}) > 0$. Hence, for any $\lambda > \overline{\lambda}$, where

$$\overline{\lambda} \equiv \frac{\Delta u}{\underline{H}\left(W_T - W_{\varnothing}\right)},$$

the expression on the right-hand side of Equation (18) is strictly positive. Note that $\overline{\lambda}$ is independent of the signal σ and depends only on the problem's primitives. Thus, when $\lambda > \overline{\lambda}$, any attempt to construct a signal that is better than truth-telling leads to a contradiction.

Part II. By Proposition 1, any optimal signal can be represented as a probability distribution over the extreme posterior beliefs in $\mathcal{E}$. Let $\mathcal{P}$ denote the compact polytope of all Bayes-plausible distributions over $\mathcal{E}$. It is therefore convenient to identify a signal with a vector $p = (p_1, \ldots, p_{|\mathcal{E}|}) \in \mathcal{P}$ which specifies the probabilities assigned to the extreme beliefs $(\mu_1, \ldots, \mu_{|\mathcal{E}|})$.

Sender's expected payoff (Equation (6)) can then be written as a function of p:

$$V_\lambda(p) \;=\; f(p) - \lambda g(p),$$

where $f(p) = c^\top p$ with $c = \Big(\hat{u}_S(\mu_1),\ \ldots,\ \hat{u}_S(\mu_{|\mathcal{E}|})\Big)$, and $g : \mathcal{P} \to \mathbb{R}$ is continuously differentiable, with the dependence on the fixed extreme beliefs in $\mathcal{E}$ suppressed in the notation.

When $\lambda = 0$, the problem reduces to the standard Bayesian persuasion benchmark. Let $p^* \in \mathcal{P}$ denote the unique maximizer of $V_0(p)$. Our goal is to show that for sufficiently small $\lambda > 0$, the point p^* remains the maximizer of V_λ:

$$p^* = \arg\max_{p \in \mathcal{P}} V_\lambda(p) \qquad \forall \lambda < \underline{\lambda}.$$

For notational convenience, we translate coordinates so that the unique maximizer p^* becomes the origin, i.e. $p^* = 0$.[16] Thus we need to prove that for "sufficiently small" λ:

$$V_\lambda(0) - V_\lambda(p) = \big(f(0) - f(p)\big) - \lambda\big(g(0) - g(p)\big) \ > \ 0 \qquad \forall\, p \in \mathcal{P} \setminus \{0\}. \tag{19}$$

Our proof strategy is to find a linear (in $\|p\|$) upper bound on $\big|g(0) - g(p)\big|$, and a linear (in $\|p\|$) lower bound on $f(0) - f(p)$, and then to choose a constant $\underline{\lambda} > 0$ that ensures that Equation 19 is positive for all $p \in \mathcal{P}$ when $\lambda < \underline{\lambda}$.

To establish the upper bound on $\big|g(0) - g(p)\big|$, note that because g is continuously differentiable on the compact polytope $\mathcal{P}$, its gradient is bounded. Hence, g is globally Lipschitz on $\mathcal{P}$. Thus, there exists $\delta_g > 0$ such that

$$|g(0) - g(p)| \ \leq \ \delta_g \, \|p\| \qquad \text{for all } p \in \mathcal{P}. \tag{20}$$

To establish the lower bound on $f(0) - f(p)$, note that because f is linear and 0 is its unique maximizer over $\mathcal{P}$, we have

$$f(0) - f(p) = \langle -c, p \rangle > 0 \qquad \forall\, p \in \mathcal{P} \setminus \{0\}. \tag{21}$$

Let C denote the tangent cone of $\mathcal{P}$ at 0, and let K denote its intersection with the unit sphere:

$$K = C \cap S^{d-1}, \qquad d = |\mathcal{E}| - 1.$$

For any $p \in \mathcal{P} \backslash \{0\}$, the direction $v = p/\|p\|$ lies in K. Furthermore, since Equation (21) implies that $\langle -c, v \rangle > 0$ for all feasible directions $v \in K$, and because K is compact, the continuous function $v \mapsto \langle -c, v \rangle$ attains a strictly positive minimum on K. Denote this minimum by $\delta_f \equiv \min_{v \in K} \langle -c, v \rangle > 0$. Thus, for any

[16] Formally, we replace each $p \in \mathcal{P}$ by $\tilde{p} = p - p^*$, which shifts the polytope to $\tilde{\mathcal{P}} = \mathcal{P} - p^*$. In these translated coordinates we write $\tilde{p}$ simply as p, so that p^* corresponds to 0.

$p \in \mathcal{P} \setminus \{0\}$,

$$f(0) - f(p) = \langle -c, p \rangle = \|p\| \, \langle -c, p/\|p\| \rangle \geq \delta_f \, \|p\|. \tag{22}$$

Combining Equations (20) and (22), for any $p \in \mathcal{P} \setminus \{0\}$ we have

$$V_\lambda(0) - V_\lambda(p) = \big(f(0) - f(p)\big) - \lambda\big(g(0) - g(p)\big) \geq (\delta_f - \lambda\delta_g) \, \|p\|.$$

Choose $\underline{\lambda} \equiv \delta_f/\delta_g$. Then for any $0 < \lambda < \underline{\lambda}$, the right-hand side is strictly positive for all $p \in \mathcal{P} \setminus \{0\}$. Thus 0 remains the unique maximizer of V_λ on $\mathcal{P}$.

Undoing the translation of coordinates shows that for sufficiently small $\lambda > 0$, the original vector p^* remains the unique maximizer of V_λ, which completes the proof. ■

Proof of Proposition 5

Fix $\varepsilon > 0$ and define the minimal inaccuracy level among ε-sender-optimal signals by

$$\underline{\rho} \; := \; \inf\Big\{\rho(\sigma) \; : \; \sigma \text{ is } \varepsilon\text{-Sender-optimal}\Big\}.$$

Step 1: $\underline{\rho}$ is attainable. Choose a sequence $\{\sigma_n\}$ of ε-sender-optimal signals such that $\rho(\sigma_n) \searrow \underline{\rho}$. Let P_n denote the distribution over posterior beliefs induced by σ_n, and let $M_n = \text{supp}(P_n)$ denote its (finite) support.

Since the belief simplex is compact and each P_n has finite support, there exists a subsequence (relabeled $\{\sigma_n\}$) such that P_n converges to some distribution P_∞. Bayes plausibility is preserved in the limit, so there exists a signal σ_∞ that induces P_∞. By continuity of ρ in the distribution over posteriors, $\rho(\sigma_\infty) = \lim_{n\to\infty} \rho(\sigma_n) = \underline{\rho}$.

We claim that σ_∞ is ε-sender-optimal. Observe first that if Receiver's preferences were held fixed, then Sender's expected payoff could not decrease under convergence of the induced posterior distributions. In our setting, Receiver's preferences vary with ρ, but along the minimizing sequence we have $\rho(\sigma_n) \searrow \rho(\sigma_\infty)$. By Equation (9), lowering ρ can only induce Receiver to choose actions that are weakly more favorable for sender at each belief, that is $\hat{u}_S(\mu; \rho)$ is non-increasing in ρ (the change in Receiver's preferences can only increase sender's payoff). Combining with the fact that $\mu \mapsto \hat{u}_S(\mu; \rho)$ is upper semicontinuous for each fixed ρ, we obtain that $V(\sigma_\infty) \geq \sup_\sigma V(\sigma) - \varepsilon$, that is, the limit signal remains ε-sender-optimal. In particular, the infimum $\underline{\rho}$ is attained.

Step 2: Support on extreme beliefs. If $M_\infty \subseteq \mathcal{E}_{\underline{\rho}}$, we are done. Otherwise, there exist posterior beliefs in M_∞ which are not extreme (with respect to Receiver's preferences determined by $\underline{\rho}$). Apply the splitting procedure described in the proof of Proposition 1 to obtain a new signal $\tilde{\sigma}$ satisfying

$$\rho(\tilde{\sigma}) \leq \rho(\sigma_\infty) = \underline{\rho}, \qquad \text{and} \qquad U_S(\tilde{\sigma}) \geq U_S(\sigma_\infty).$$

Thus $\tilde{\sigma}$ is also ε-sender-optimal. By definition of $\underline{\rho}$, this implies $\rho(\tilde{\sigma}) = \underline{\rho}$, and hence $\mathcal{E}_{\rho(\tilde{\sigma})} = \mathcal{E}_{\underline{\rho}}$. Therefore, Receiver's preferences are unchanged when moving from σ_∞ to $\tilde{\sigma}$, and the splitting procedure ensures that $\tilde{\sigma}$ is supported on extreme beliefs in $\mathcal{E}_{\rho(\tilde{\sigma})}$. This completes the proof for the case of ε-sender-optimal signals.

If a sender-optimal signal exists, the same argument applies with $\varepsilon = 0$, yielding an optimal signal supported on $\mathcal{E}_{\rho(\sigma)}$. ■

Proof of Proposition 6

The truthful signal corresponds to the distribution $(q_1, \ldots, q_L)$ over the pure posteriors $(e_1, \ldots, e_L)$. Without loss of generality, suppose that action a_i is uniquely optimal in state ω_i.

Define $\mu = \left(1-\delta, \frac{\delta}{L-1}, \ldots, \frac{\delta}{L-1}\right)$. Notice that when δ is small, μ is a posterior belief that is "close to," but not equal to, the pure posterior e_1. Now, consider a signal that induces the posteriors $(\mu, e_2, \ldots, e_L)$ with probabilities $(p_1, \ldots, p_L)$, where

$$p_1 = \frac{q_1}{1-\delta}, \quad p_i = q_i - \frac{q_1\delta}{(L-1)(1-\delta)} \quad \text{for } i \geq 2.$$

For δ sufficiently small, all these probabilities are positive. It is straightforward to verify that $\sum_{i=1}^{L} p_i = 1$ and that Bayes plausibility holds, i.e., $\mu p_1 + \sum_{i=2}^{L} e_i p_i = q$. This modification can be interpreted as transferring a small amount of probability from each pure posterior e_i (for $i \geq 2$) to the posterior μ.

When δ is small, the inaccuracy ρ of the new signal is close to zero. Because Receiver's utility is continuous in ρ, the optimal action at the posterior μ, which is close to e_1, remains a_1. Moreover, Receiver's optimal action at each pure posterior e_i remains unchanged.

Thus, relative to the truthful signal, the modified signal reallocates probability mass in a way that increases the likelihood that Receiver selects action a_1, rather than other actions that are less preferred by Sender. Consequently, the modified signal yields a higher expected payoff for Sender than the truthful signal, completing the proof. $\blacksquare$

Proof of Proposition 7

We begin by noting that two properties of equilibria in the CS benchmark continue to hold. First, if two Sender types θ, θ' induce the same *action* in equilibrium, then every type $\theta'' \in [\theta, \theta']$ must also induce that action. This follows from the single-crossing property of Sender's preferences.

Second, in equilibrium no positive-measure set of Sender types can reveal themselves by sending distinct, type-specific messages. It follows that, without loss of generality, we may restrict attention to equilibria with only countably many messages (each sent by a positive measure of types), and therefore also countably many induced actions.

For each action a induced in (σ, α), let I_a denote the set of types that induce a. By the two properties above, I_a is an interval. Let M_a be the set of messages sent by types in I_a. Since all these messages induce the same action, they must also yield the same posterior mean:

$$\mathbb{E}_\sigma[\omega \mid m] = \mathbb{E}[\omega \mid \omega \in I_a] \quad \forall m \in M_a.$$

Now define (σ', α') as follows: for each a, all types in I_a send a single message m'_a, and Receiver responds with $\alpha'(m'_a) = a$. This construction preserves the type–action mapping of (σ, α), and hence satisfies the first property in the statement of the proposition.

It remains to show that $\rho(\sigma) = \rho(\sigma')$. By Equation 10, and by the Law of Total Variance, for any equilibrium strategy σ,

$$\rho(\sigma) = \mathbb{E}_\sigma[\mathrm{Var}_\sigma[\omega \mid m]] = \mathrm{Var}[\omega] - \mathrm{Var}_\sigma\big(\mathbb{E}_\sigma[\omega \mid m]\big).$$

Since σ' preserves the distribution of posterior means generated by σ, we have $\mathrm{Var}_{\sigma'}\big(\mathbb{E}[\omega \mid m']\big) = \mathrm{Var}_\sigma\big(\mathbb{E}[\omega \mid m]\big)$, and hence $\rho(\sigma') = \rho(\sigma)$.

Proof of Proposition 8

Because $d(\cdot)$ is nondecreasing by assumption and ϕ is nondecreasing (see Crawford and Sobel (1982)), the operator $T_{b,d}$ defined in Equation (15) is an order-preserving self-map.

Suppose that ρ is the expected posterior variance induced by some equilibrium in our model. Sender then faces the same incentives as in the Crawford-Sobel benchmark with effective bias $b + d(\rho)$. Hence, ρ corresponds to an equilibrium variance of the Crawford-Sobel model at that effective bias. Since ϕ selects the minimal expected posterior variance among such equilibria, it follows that ρ is a pre-fixpoint of $T_{b,d}$, that is,

$$\rho \geq \phi\big(b + d(\rho)\big) = T_{b,d}(\rho).$$

By the Knaster-Tarski fixed-point theorem, $T_{b,d}$ admits a least fixed point. Denote it by $\underline{\rho}$. Moreover, by Observation 1, any fixed point of $T_{b,d}$ corresponds to an equilibrium of our model. Since the least fixed point coincides with the infimum of all pre-fixpoints of $T_{b,d}$, it follows that $\underline{\rho}$ corresponds to the equilibrium with the smallest achievable inaccuracy.

Finally, note that Sender's expected utility, given by Equation (13), is strictly decreasing in ρ. Hence, Sender strictly prefers the equilibrium associated with the least fixed point $\underline{\rho}$. In the Crawford-Sobel benchmark, the equilibrium with effective bias $b + d(\underline{\rho})$ and variance $\underline{\rho}$ employs only finitely many messages. Since the same partition also supports an equilibrium in our model with bias b, it follows that a Sender-optimal equilibrium exists in which Sender employs only finitely many messages. ■

Proof of Proposition 9

To simplify notation, let $\rho_1^* \equiv \rho^*_{(b,d_1)}$ and $\rho_2^* \equiv \rho^*_{(b,d_2)}$. Recall that ρ_1^* and ρ_2^* are the least fixed points of T_{b,d_1} and T_{b,d_2}, respectively. Then

$$\rho_1^* = T_{b,d_1}(\rho_1^*) = \phi\big(b + d_1(\rho_1^*)\big) \;\geq\; \phi\big(b + d_2(\rho_1^*)\big) = T_{b,d_2}(\rho_1^*),$$

where the inequality follows from $d_1(\cdot) \geq d_2(\cdot)$ pointwise and the fact that ϕ is nondecreasing.

Therefore, $T_{b,d_2}(\rho_1^*) \leq \rho_1^*$; that is, ρ_1^* is a pre-fixpoint of T_{b,d_2}. By the Knaster-Tarski fixed-point theorem, the least fixed point of T_{b,d_2} coincides with the infimum of its pre-fixpoints. Hence, $\rho_2^* \leq \rho_1^*$, as desired. ■

Proof of Proposition 10

Suppose, toward a contradiction, that a Sender-optimal equilibrium exists with a finite number of messages. Let $\rho_0 > 0$ be the inaccuracy of this equilibrium. Sender's payoff in this equilibrium is identical to the Sender-optimal equilibrium in the CS benchmark with effective bias $b + d(\rho_0)$.

For any $\varepsilon \in (0, d(\rho_0))$, consider the Sender-optimal equilibrium in the CS benchmark with effective bias $b + \varepsilon$. Let $\mathcal{P}^\varepsilon = \{I_1, \ldots, I_n\}$ be the partition corresponding to this equilibrium, and denote $\mu_j = \mathbb{E}[\omega \mid \omega \in I_j]$. Since Sender's equilibrium payoff in the CS benchmark is strictly decreasing in the effective bias, the partition $\mathcal{P}^\varepsilon$ yields a strictly higher payoff to Sender (under the CS benchmark) than the initial equilibrium.

Because $\phi(b) > 0$, and since d is continuous, strictly increasing and satisfies $d(0) = 0$, it follows that for sufficiently small $\varepsilon > 0$ we have

$$\phi(b + d(\rho_\varepsilon)) > \rho_\varepsilon$$

where $\varepsilon = d(\rho_\varepsilon)$. Fix such an ε. For ease of notation, in what follows we omit the superscript ε from the partition $\mathcal{P}^\varepsilon$.

Next, we define a global refinement of the partition $\mathcal{P}$, denoted $\mathcal{P}_{\gamma,M}$ for $\gamma \in [0, 1]$ and $M \in \mathbb{N}$, by applying the following construction to every interval $I_j \in \mathcal{P}$ simultaneously (see Figure 4):

1. **The Residual Mass:** Within each I_j, we define a cell $S_{j,0}$ containing a $(1-\gamma)$ fraction of the interval's mass, balanced such that $\mathbb{E}[\omega \mid \omega \in S_{j,0}] = \mu_j$.
2. **The Refined Mass:** The remaining γ fraction of mass in each I_j is partitioned into M disjoint cells $\{S_{j,k}\}_{k=1}^M$ of equal mass. Each $S_{j,k}$ consists of a pair of segments balanced around μ_j so that $\mathbb{E}[\omega \mid \omega \in S_{j,k}] = \mu_j$.

By construction, every element in the refined partition $\mathcal{P}_{\gamma,M}$ associated with I_j has the mean μ_j. For $\varepsilon > 0$ sufficiently small, there exists $\overline{M}$ such that $\rho(\mathcal{P}_{1,\overline{M}}) < \rho_\varepsilon$. At $\gamma = 0$, the refined partition coincides with $\mathcal{P}$, so $\rho(\mathcal{P}_{0,\overline{M}}) = \rho(\mathcal{P}) > \rho_\varepsilon$.

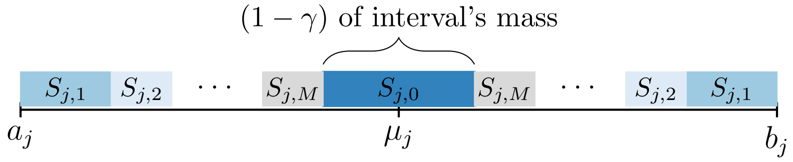

Figure 4: Construction of the partition $\mathcal{P}_{\gamma,M}$ for a representative interval I_j.

By the continuity of ρ under small refinements, there exists $\gamma^* \in (0,1)$ such that $\rho(\mathcal{P}_{\gamma^*,\overline{M}}) = \rho_\varepsilon$.

The refined partition $\mathcal{P}_{\gamma^*,\overline{M}}$ constitutes a valid equilibrium in our model (with the interpretation that all Sender types in the same cell send the same message). Because the refined partition's inaccuracy, as measured by ρ, is exactly ρ_ε, Receiver's reciprocity induced bias is $d(\rho_\varepsilon)$, making the strategic interaction equivalent to that in a CS benchmark with effective bias $b + d(\rho_\varepsilon) = b + \varepsilon$. Since $\mathcal{P}$ was the equilibrium partition for this bias and our refinement preserves the mapping from states to posterior means, Sender's incentive compatibility is satisfied. This equilibrium provides Sender with a strictly higher payoff than the original equilibrium, contradicting the optimality of the original equilibrium. ■

Proof of Lemma 3

The obedience constraints for Player i require that the expected payoff from following the mediator's recommendation weakly dominates the expected payoff from deviating. Formally, they are given by

$$\sum_{\omega,a_{-i}} \pi^{\omega}_{1,a_{-i}} \big[u_i(\omega,1,a_{-i}) - u_i(\omega,0,a_{-i})\big] \geq 0,$$
$$\sum_{\omega,a_{-i}} \pi^{\omega}_{0,a_{-i}} \big[u_i(\omega,0,a_{-i}) - u_i(\omega,1,a_{-i})\big] \geq 0.$$

where $\pi^{\omega}_{a_i,a_{-i}}$ denotes the ex ante joint probability that the mediator recommends action a_i to Player i and action a_{-i} to the other player when the state is ω, and $u_i(\omega,a_i,a_{-i})$ denotes Player i's payoff when she plays action a_i, the other player plays a_{-i}, and the state is ω.

Since the prior probability of each state is 0.5, the probabilities in state 1 satisfy

$$\pi^1_{00} + \pi^1_{01} + \pi^1_{10} + \pi^1_{11} = 0.5. \tag{23}$$

The obedience constraint ensuring that Player 1 prefers action 0 to action 1 when recommended action 0 is $-2\pi^0_{00} - 2\pi^0_{01} + \pi^1_{00} \geq 0$ which can be equivalently written as

$$\pi^0_{00} + \pi^0_{01} - 0.5\pi^1_{00} \leq 0. \tag{24}$$

Similarly, the obedience constraint ensuring that Player 2 prefers action 1 to action 0 when recommended action 1 is $-\pi^0_{11} + 3\pi^1_{01} - 2\pi^1_{11} \geq 0$, or equivalently:

$$0.2\pi^0_{11} - 0.6\pi^1_{01} + 0.4\pi^1_{11} \leq 0. \tag{25}$$

Multiplying (23) by 0.6, adding (24) and (25) and rearranging, we obtain:

$$\pi^0_{00} + \pi^1_{11} + \pi^0_{01} + 0.2\pi^0_{11} + 0.1\pi^1_{00} + 0.6\pi^1_{10} \leq 0.3.$$

Since all probabilities on the left-hand side are nonnegative, we obtain an upper bound on the mediator's expected payoff:

$$V = \pi^0_{00} + \pi^1_{11} \leq 0.3.$$

This completes the proof. ■

Bibliography

Antler, Yair. 2015. "Two-Sided Matching with Endogenous Preferences." *American Economic Journal: Microeconomics* 7 (3): 241–58.

Au, Pak Hung, OSub Kwon, and King King Li. 2023. "A Simple Experiment on Simple Bayesian Persuasion." *Working paper.*

Bergemann, Dirk, and Stephen Morris. 2016. "Bayes correlated equilibrium and the comparison of information structures in games." *Theoretical Economics* 11 (2): 487–522.

Bierbrauer, Felix, and Nick Netzer. 2016. "Mechanism Design and Intentions." *Journal of Economic Theory* 163 557–603.

Brøndsted, Arne. 1983. *An Introduction to Convex Polytopes.* New York: Springer-Verlag.

Caplin, Andrew, and Kfir Eliaz. 2003. "AIDS Policy and Psychology: A Mechanism-Design Approach." *The RAND Journal of Economics* 34 (4): 631–646.

Caplin, Andrew, and John Leahy. 2004. "The supply of information by a concerned expert." *The Economic Journal* 114 (497): 487–505.

Charness, Gary, and Matthew Rabin. 2002. “Understanding Social Preferences with Simple Tests.” *The Quarterly Journal of Economics* 117 (3): 817–869.

Cover, T.M., and J.A. Thomas. 2012. *Elements of Information Theory.* Wiley.

Crawford, Vincent P., and Joel Sobel. 1982. “Strategic Information Transmission.” *Econometrica* 50 (6): 1431–1451.

Davey, B. A., and H. A. Priestley. 2002. *Introduction to Lattices and Order.* Cambridge University Press, , 2nd edition.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. “A theory of sequential reciprocity.” *Games and Economic Behavior* 47 (2): 268–298. https://doi.org/10.1016/j.geb.2003.06.003.

Falk, Armin, and Nora Szech. 2013. “Morals and Markets.” *Science* 340 (6133): 707–711.

Fehr, Ernst, and Simon Gächter. 1998. “Reciprocity and Economics: The Economic Implications of Homo Reciprocans.” *European Economic Review* 42 (3-5): 845–859.

Fehr, Ernst, and Simon Gächter. 2000. “Cooperation and Punishment in Public Goods Experiments.” *The American Economic Review* 90 (4): 980–994.

Gintis, Herbert. 1972. “A Radical Analysis of Welfare Economics and Individual Development.” *The Quarterly Journal of Economics* 86 (4): 572–599.

Grant, Simon, Atsushi Kajii, and Ben Polak. 1998. “Intrinsic Preference for Information.” *Journal of Economic Theory* 83 (2): 233–259.

Hagenbach, Jeanne, and Charlotte Saucet. 2025. “Motivated Skepticism.” *The Review of Economic Studies* 92 (3): 1882–1919.

Kamenica, Emir, and Matthew Gentzkow. 2011. “Bayesian Persuasion.” *American Economic Review* 101 (6): 2590–2615.

Kremer, Ilan, Yishay Mansour, and Motty Perry. 2014. “Implementing the “Wisdom of the Crowd”.” *Journal of Political Economy* 122 (5): 988–1012.

Lipnowski, Elliot, and Laurent Mathevet. 2017. “Simplifying Bayesian Persuasion.” *Working paper.*

Lipnowski, Elliot, and Laurent Mathevet. 2018. “Disclosure to a Psychological Audience.” *American Economic Journal: Microeconomics* 10 (4): 67–93.

Mathevet, Laurent, Jacopo Perego, and Ina Taneva. 2020. “On Information Design in Games.” *Journal of Political Economy* 128 (4): 1370–1404.

Rabin, Matthew. 1993. “Incorporating Fairness into Game Theory and Economics.” *The American Economic Review* 83 (5): 1281–1302.