Image-Building Persuasion

Carl Heese and Shuo Liu*

April 2023

Abstract

In many economic situations, a sender communicates strategically with a receiver not only to influence his decision-making but also to influence how certain unobserved characteristics of herself (e.g. loyalty, integrity, or unselfishness) are perceived. To study such strategic interactions, we introduce image-building motives into the canonical framework of Bayesian persuasion. We characterize how the sender optimally sacrifices her persuasive influence on the decision in order to boost her reputation, by manipulating the communication about a payoff-relevant state. We describe when this harms and benefits the receiver unambiguously, revealing also that effects often depend on the selected equilibrium and are non-monotone in the sender's characteristics. We illustrate our findings within various standard preference settings, for instance with quadratic losses or state-independent sender preferences. Finally, the model provides insights into a wide range of applications, such as how reputational concerns of politicians can explain why reforms deemed welfare-enhancing by economists fail to be adopted, or how hierarchical structures in organizations may engender harmful intransparencies.

Keywords: information design, persuasion, image concerns, signaling

JEL Classification: C72, D72, D82, M50

*Heese: Department of Economics, University of Vienna. Liu: Guanghua School of Management, Peking University. Emails: carl.heese@univie.ac.at and shuo.liu@sm.pku.edu.cn. We are grateful for very helpful comments to Si Chen, Navin Kartik, Daniel Krähmer, Stephan Lauermann, and seminar participants at Peking University, University of International Business and Economics, University of Vienna, and the SAET Annual Conference 2022. This draft is preliminary and incomplete. Any comments are welcome.

1 Introduction

This paper develops and applies a novel model of strategic communication and information control. In many economic situations, a sender strategically communicates with a receiver with a two-fold goal. There is a *persuasion motive*. The sender likes to influence the decision-making of the receiver. There is a *signaling motive*. The sender cares about how the communication reflects certain unobserved characteristics of herself, e.g., loyalty, unselfishness, or integrity.

The existing work on strategic communication focuses on the first goal, the sender's ability to influence the receiver's action (see, e.g., Grossman, 1981; Kamenica and Gentzkow, 2011; Milgrom, 1981) and signaling is considered a means to this end (see, e.g., Crawford and Sobel, 1982; Green and Stokey, 2007). We present a model in which there is a strategic tension between the two goals. This tension arises since the type inference is (directly) pay-off relevant to the sender and does not advance persuasion. This captures a variety of unexplored situations: First, the sender may care about the type inference as a social signal (Bénabou and Tirole, 2006). Consider a manager of an organization who controls the information flow to the subordinate about the requirements of tasks, and by doing so, may take influence on her effort level. She *also* cares about the signal that her communication behavior sends to the higher-ups or other employees. Second, the sender may care about the type inference as a signal to herself; that is, she has self-image concerns (Bodner and Prelec, 2003). Consider a consumer who collects information about the sustainability of a cheap fast-fashion product before deciding if to buy it. By doing so, she signals to herself that she is moral; however, the acquired information will also affect her decision-making. Third, closer to existing work, the payoff from the type inference may also be viewed as the reduced form of the value of reputation in a continuation game. Thus, there is an intertemporal trade-off between the persuasion of the current receiver and influence in the future through reputation.

To model the strategic tension between the two goals, we introduce a signaling motive into the canonical framework of Bayesian persuasion (Kamenica and Gentzkow, 2011). The key assumption of the Bayesian persuasion framework is that the sender cannot distort information or conceal it from the receiver once a signal realization is known. The Bayesian persuasion framework is versatile: It is natural in situations in which observations are public[1],

[1]Note the equivalence of the static Bayesian persuasion framework to dynamic sequential sampling models when information is about a binary state (Chen and Heese, 2021; Henry and Ottaviani, 2019; Morris and Strack, 2019).

or in situations where sender and receiver are two selves of the same agent. It also captures situations in which a sender privately acquires verifiable information and can then decide on disclosure. Here, unraveling arguments typically enforce full disclosure, qualifying this as a reduced form assumption for a broad set of settings, see, e.g., Gentzkow and Kamenica (2017) for a discussion of this point. Finally, the framework has a literal interpretation in terms of a sender who holds private information and has commitment power over his communication strategy.

As in the standard persuasion framework, there is a pay-off relevant state, a sender (he) with a one-dimensional private preference type, and a receiver. The sender chooses how to communicate with the receiver by choosing a distribution of a recommended action for each realization of a pay-off relevant state. The receiver chooses an action upon observing the recommendation. The sender has the standard persuasion motive; that is, he derives state-dependent utility from the action of the receiver. Additionally, in our model, he has a signaling motive. He derives an increasing utility from the belief implied by his communication strategy about her private preference type.

The model presents challenges. Even without the signaling motive, the equilibrium sender strategy is often intractable. Further, as typical in signaling games, choices of off-equilibrium beliefs generate a large multiplicity of equilibria. We advance in two ways. First, to rule out unreasonable equilibria, we invoke a standard equilibrium refinement, the D1 criterion by Cho and Kreps (1987).[2] Second, instead of studying the sender's communication strategies directly, we study an auxiliary game, at the interim stage. In this game, the sender chooses over bundles of an expected interim utility from the receiver's action and the belief over her type. We maintain a simple condition on the sender's preferences that implies a single-crossing condition of the indifference conditions over such bundles. Loosely speaking, the single-crossing condition means that higher types care relatively more about the utility from the belief.

We show that all equilibria are *semi-separating*: Higher types are willing to sacrifice their persuasive influence on the receiver's action in order to separate from the lower types. There is a unique cutoff so that all types below the cutoff separate by using different strategies, and all types above pool on the same strategy. We derive the explicit formula for the (opportunity) cost of signaling the own preference type for the types on the separating interval. Our

[2] Alternative criteria such as Universal Divinity (Banks and Sobel, 1987) and Never-a-Weak-Best-Response (Kohlberg and Mertens, 1986) are equivalent to D1 in our setting, as we will show formally.

characterization will imply that all equilibria are payoff-equivalent for all sender types. Thus, they are Pareto ranked, only varying in the receiver's payoff from her decision.[3]

The implications of signaling concerns on the receiver's welfare are intricate. Clearly, if the sender and receiver share the same preferences over the receiver's action, when the sender sacrifices her material utility, this harms the receiver as well. If the utility of the sender and receiver from the receiver's action is zero-sum, clearly, when the sender sacrifices his influence on the receiver's action, this benefits the receiver. Intuitively, what seems to matter is if and how the preferences of the sender and receiver are aligned. Consider an intermediate case. There is a binary state, and preferences are aligned in one state but not in the other. The sender obtains a utility of 1 only if the receiver chooses action $a = 1$. The receiver obtains a utility of 1 only if his action matches the state. In all other scenarios, utilities are zero. Without the signaling motive, there is a maximal probability of recommending the action $a = 1$ to the receiver in an incentive-compatible way.[4] The signaling motive now causes the higher types to separate by recommending the action $a = 1$ less often. However, they do not pin down in which state this happens. Recommending $a = 1$ less often in state 0 benefits the receiver. Recommending $a = 1$ less often in state 1 harms the receiver. Thus, the separation incentives do not pin down the direction of the welfare effect. In this example, there are multiple non-payoff equivalent equilibria, and receiver welfare is non-monotone in some of them, typically including the Pareto-optimal one.

Allowing for a general state space, action space, and preferences, we characterize the settings in which the receiver is unambiguously worse (better) off relative to the setting without signaling motive. These include simple examples like the ones above but also classical settings, such as those with quadratic losses in Crawford and Sobel (1982) and Melumad and Shibano (1991). For a large set of environments, the welfare effect is ambiguous. Its direction depends on the selected equilibrium and on the distribution of sender types. The multiple equilibria differ in the strategies chosen by the sender types, reflecting different *cultures of communication* (compare with Schelling, 1980).

We characterize the Pareto frontier of the equilibrium set and show that information transmission about the state is often non-monotone in the sender type. Further, information

[3]Since all equilibria share the same cutoff, the information released about the sender type is the same. Hence, if this information is pay-off relevant for a third party or within a continuation game, the pay-off consequences are the same in all equilibria. Thus, even when taking into account such additional payoffs, the equilibria are Pareto-ranked.

[4]See Kamenica and Gentzkow (2011).

transmission is often non-monotone when varying the level of signaling concerns uniformly for all types. Low levels improve information transmission. Yet, in the extreme of high signaling concerns, there is "conformity" (Bernheim, 1994). All types send the same message with probability one.

In the last part of the paper, we present applications of the theory. First, we revisit the classic question of why the political system often fails to adopt policies that experts consider welfare-enhancing. Fernandez and Rodrik (1991) show that such "resistance to reform" may arise when a majority of a democratic public is skeptical about the consequences of reform, in that they expect to be impacted negatively themselves, although ex-post they would benefit. We apply our theory to investigate when and why such ex-ante beliefs may exist in the first place. Within a model of an expert, a politician, and the public, we show that such skeptical beliefs may sustain in equilibrium when the politician has reputational concerns, e.g. due to future elections. The politician distorts the communication of expert knowledge to the public to conform with the public's prior skepticism, a communication strategy observed in modern populists. Second, we revisit the question of intransparency in organizations, as in Jehiel (2015). This line of work asks when a manager of an organization chooses, or *should* choose, to be intransparent in a moral hazard interaction with an agent. We formalize the novel argument that a manager's communication and information management towards subordinates may be governed by incentives to impress and signal to higher-ups, and we show that such signaling may involve the use of strategies that are non-trivial and only selectively reveal information. Last but not least, we present several self-signaling applications, and relate to a companion paper with a laboratory experiment about moral self-signaling (Chen and Heese, 2021).

The paper contributes to the literature on strategic communication. Below, we discuss relations and contributions to specific streams of the literature.

We contribute to the information design literature by introducing a novel consideration next to the standard persuasion motive. The sender trades off the motive to signal about her type and her persuasive influence on a receiver's action. This trade-off speaks to a wide array of applications, as will be showcased in Section 4. In particular, the signaling concerns may be given a behavioral interpretation in terms of psychological preferences (Geanakoplos, Pearce and Stacchetti, 1989), such as conformity (Bernheim, 1994), social image concerns (Bénabou and Tirole, 2006), or self-image concerns (Baumeister, 1998; Bodner and Prelec,

2003; Köszegi, 2006). Complementary to the interpretation of the sender's signaling concerns as psychological preferences, there is work on Bayesian persuasion to a receiver with psychological preferences (Lipnowski and Mathevet, 2018; Schweizer and Szech, 2018). There is also related literature on information design with signaling, which has however either maintained that all senders share a common preference type (Hedlund, 2017) or that the pay-off relevant state is identical to the sender's type (Koessler and Skreta, 2021; Perez-Richet, 2014). Consequently, the existing work has not studied the trade-off between the motive to signal about a preference type and the persuasion motive, and the analysis in it relies on different techniques.

We contribute to the literature on signaling games. Unlike the previous literature, we provide a framework for situations in which there is a strategic tension between signaling and influencing a receiver's action. This is interesting in terms of new applications but also since there are non-trivial theoretical implications. Non-standard results about the receiver's welfare arise. For example, receiver welfare if often non-monotone in the sender's type despite an appropriate single-crossing condition, and it varies with the selected equilibrium. Other results, such as the incomplete separation at the top, have been established also in other contexts (see, e.g., Bernheim, 1994; Cho and Sobel, 1990; Kartik, 2009).

Finally, the paper contributes to a broader literature on reputational concerns. Most closely related is the work on "bad reputation" (Ely, Fudenberg and Levine, 2008; Ely and Välimäki, 2003; Morris, 2001) which identifies conditions when reputational concerns harm a long-lived player who repeatedly interacts with short-lived players. Relatedly, we provide characterizations about when reputational or image concerns reduce receiver welfare. However, our results are derived in a much different context and setting, in which, for example, there are no commitment types.

The paper's structure as follows: Section 2 presents the formal model. Section 3 contains all results about the equilibria and welfare, and applies the results to numerous standard preference settings. Section 4 presents applications to various streams of literature. Section 5 contains concluding remarks.

2 Model

We study a communication game between a sender (she) and a receiver (he). There is a state space Ω, with a typical state denoted by ω, and an action space A, with a typical action denoted by a. Both A and Ω are compact metric spaces. The players are uncertain about the state at the outset of the game and share a prior $\mu_0 \in \text{int}(\Delta(\Omega))$. The sender moves first by choosing an information structure $\pi : \Omega \to \Delta(A)$, where each signal realization $s \in supp(\pi)$ is interpreted as an action recommended to the receiver. The set of all possible information structures is denoted by Π. The receiver observes the sender's choice of information structure and the signal realization, and finally chooses an action $a \in A$ (which need not coincide with what is recommended by the sender).

Preferences. The receiver has a continuous utility function $u^R(a, \omega)$ that depends on both his action and the state of the world. The sender is endowed with a private type $\theta \in \Theta \equiv [0, 1]$, which is commonly known to be distributed according to an absolutely continuous distribution function Γ with full support. The sender also has a continuous utility function

$$u^S(a, \omega, \theta, \eta) = v(a, \omega) + \phi \cdot w(p(\eta), \theta),$$

where $\eta \in \Delta(\Theta)$ denotes the receiver's belief about the sender's type, and $p(\eta) \equiv \mathbb{E}_\eta[\tilde{\theta}]$ is interpreted as the sender's *image*. Naturally, $\phi > 0$ measures how much the sender cares about the image payoff $w(p, \theta)$ relative to the material payoff $v(a, \omega)$. Further, the function $w(\cdot)$ is continuously differentiable with

$$\frac{\partial w(p, \theta)}{\partial p} > 0 \text{ and } \frac{\partial^2 w(p, \theta)}{\partial p \partial \theta} > 0, \tag{1}$$

meaning that, while all types prefer to be perceived as a high type, such a desire is stronger for higher types.

Strategies and equilibrium. A pure strategy of the sender is a mapping $\sigma : \Theta \to \Pi$ that specifies for each type an information structure. A pure strategy of the receiver is a mapping that specifies an action for every possible information structure and signal realization. We analyze the perfect Bayesian equilibria in pure sender strategies (Fudenberg and Tirole, 1991, p. 333; henceforth equilibrium). In equilibrium, given the sender's choice of information

structure and the subsequent signal realization, the receiver forms posteriors beliefs about the state using Bayes' rule, and then takes an action to maximize his expected payoff.[5] At the same time, the receiver may also update his beliefs about the sender's type. In other words, the sender's strategy influences not only the material outcome of the game but also her image in the eyes of the receiver.

An application of the revelation principle reveals that it is without loss to focus on equilibria in which the receiver follows the sender's recommendation. Therefore, we can identify an equilibrium with an *incentive compatible* sender strategy $\sigma = \{\pi_\theta\}_{\theta\in\Theta}$ and a belief system $H = \{\eta(\pi)\}_{\pi\in\Pi}$ such that each $\eta(\pi) \in \Delta(\Theta)$ is consistent with Bayes' rule given σ. Here, incentive compatibility requires that for every $\theta \in \Theta$, the associated information structure π_θ is a solution to the following utility maximization problem:

$$\max_{\pi\in\Pi^*} U^S(\pi, \theta; \sigma) \equiv \mathbb{E}_\pi[v(s, \omega)] + \phi \cdot w(p(\eta(\pi), \theta), \tag{2}$$

where

$$\Pi^* \equiv \left\{ \pi \in \Pi : s \in \arg\max_{a\in A} \mathbb{E}\left[u^R(a, \omega)|s; \pi\right] \ \forall s \in supp(\pi) \right\}.$$

That is, given the receiver's system of beliefs and the constraint that following the sender's recommendation is indeed optimal for the receiver, no sender type can be strictly better off by deviating from the strategy σ.[6]

Equilibrium refinement. Since Bayes' rule does not put any restriction on the receiver's out-of-equilibrium beliefs about the sender's type, the usual equilibrium multiplicity of signaling games also arises in our model. We follow the literature and invoke a standard equilibrium refinement, the D1 criterion due to Cho and Kreps (1987) and Banks and Sobel (1987). The core idea is to restrict the receiver's out-of-equilibrium beliefs to the sender types that are "most likely" to benefit from the deviation to the off-path choice. Formally, given a sender

[5] We assume that whenever the receiver is indifferent between multiple actions and one of them is recommended by the sender, he will choose that action.

[6] Note that restricting the sender's choice to the set $\Pi^* \subsetneq \Pi$ (i.e., the set of information structures with which the receiver would find it optimal to always follow the sender's recommendation) is without loss of generality. To see this, take any $\pi \in \Pi \setminus \Pi^*$ and suppose that it induces the receiver to use some (sequentially rational) decision rule $\hat{a} : supp(\pi) \to A$. By relabeling every signal realization $s \in supp(\pi)$ as $\hat{s} = \hat{a}(s)$, we obtain an information structure $\hat{\pi} \in \Pi^*$. Then, it is clear that we can always set $\eta(\pi) = \eta(\hat{\pi})$ to make sure that π is sub-optimal for the sender whenever (2) holds.

strategy σ and an associated belief system of the receiver, we define for any $(\pi, \theta) \in \Pi^* \times \Theta$ the sets

$$D^0(\pi, \theta) \equiv \{\tilde{\eta} \in \Delta(\Theta) : \mathbb{E}_\pi[v(s, \omega)] + \phi \cdot w(p(\tilde{\eta}), \theta) \geq U^S(\pi_\theta, \theta; \sigma)\}$$

and

$$D(\pi, \theta) \equiv \{\tilde{\eta} \in \Delta(\Theta) : \mathbb{E}_\pi[v(s, \omega)] + \phi \cdot w(p(\tilde{\eta}), \theta) > U^S(\pi_\theta, \theta; \sigma)\}.$$

Then, an equilibrium (σ, H) is selected by the D1 criterion if for any $\pi \in \Pi^*$ that is not used by any sender type under σ, and for any sender types θ and θ',

$$D^0(\pi, \theta) \subsetneq D(\pi, \theta') \implies \theta \notin supp(\eta(\pi)). \tag{3}$$

In words, condition (3) requires that if, for a type θ, there is another type θ' that has a strict incentive to deviate to the off-path choice $\pi \in \Pi^*$ whenever θ has a weak incentive to do so, then the receiver's out-of-equilibrium beliefs upon observing this choice of the sender shall not put any weight on θ.[7] An equilibrium that passes this test is a *D1 equilibrium*; henceforth, often simply called equilibrium if no misunderstanding is possible.

3 Analysis

3.1 A Reduced-Form Characterization of Equilibria

Kamenica and Gentzkow (2011) analyze the benchmark scenario in which the sender does not have image concerns ($\phi = 0$), that is, she is purely guided by the persuasion motive. It is known that, even in that special setting, the equilibrium information structure is often intractable. This problem does not get any easier, if not more difficult, in our model, because the sender's persuasion motive may be entangled with her signaling motive. To make progress, we simplify the infinite-dimensional maximization problem (2) of the sender by moving the

[7] Considering also the off-path choices $\pi \in \Pi \setminus \Pi^*$ will not change the set of equilibria selected by the D1 criterion. To see this, fix an equilibrium and a sender strategy σ. Note that for a given belief $\tilde{\eta} \in \Delta(\Theta)$ and any $\pi \in \Pi \setminus \Pi^*$, there is $\hat{\pi} \in \Pi^*$ that yields the same interim expected payoff to the sender. Specifically, $\hat{\pi}$ can be constructed by relabeling the support of π (see footnote 6) . Hence, in the spirit of Banks and Sobel (1987) and Cho and Kreps (1987), a sender strategy passes the test required by the D1 criterion at π if and only if it passes the test at $\hat{\pi}$.

analysis to the interim stage. In particular, instead of information structures directly, we focus on the expected material payoff that the sender obtains by the choice of an information structure. Similar "reduced-form approaches" have proven useful in a variety of mechanism design settings (e.g., Ben-Porath, Dekel and Lipman, 2014; Che, Kim and Mierendorff, 2013).

We start by observing that, when viewing the game at the interim stage, it exhibits a number of useful properties. First, the interim game is monotonic in the sense of Cho and Sobel (1990), because, holding the expected material payoff fixed, all sender types share the same ordinal preferences over their images in the eyes of the receiver.[8] Second, the set of expected material payoffs that the sender can implement, formally defined by $\mathcal{V} \equiv \{V \subset \mathbb{R} : \exists\, \pi \in \Pi^* \text{ such that } V = \mathbb{E}_\pi[v(s,\omega)]\}$, is a compact interval. To see this, consider the payoffs

$$\bar{V} \equiv \max_{\pi \in \Pi^*} \mathbb{E}_\pi[v(s,\omega)] \quad \text{and} \quad \underline{V} \equiv \min_{\pi \in \Pi^*} \mathbb{E}_\pi[v(s,\omega)], \tag{4}$$

and let $\bar{\pi}$ and $\underline{\pi}$ be the information structures that give rise to $\bar{V}$ and $\underline{V}$, respectively.[9] It is clear that $\mathcal{V} \subseteq [\underline{V}, \bar{V}]$. Since any $V \in [\underline{V}, \bar{V}]$ can be implemented by appropriately mixing the information structures $\bar{\pi}$ and $\underline{\pi}$, the converse relationship $\mathcal{V} \supseteq [\underline{V}, \bar{V}]$ also holds.[10] Hence, $\mathcal{V} = [\underline{V}, \bar{V}]$. Third, as we formally show in the Appendix (Lemma A1), the sender's interim payoff has the following single-crossing property: for any $(V, \eta), (V', \eta') \in \mathcal{V} \times \Delta(\Theta)$ with $V < V'$, if type θ weakly prefers (V, η) over (V', η'), then (V, η) will be strictly preferred over (V', η') by all types $\theta' > \theta$.

The above properties allow us to apply techniques from the costly signaling literature (e.g. Cho and Sobel, 1990; Mailath, 1987; Ramey, 1996) to partially characterize the set of D1 equilibria. Given a sender strategy σ, we define $V(\theta;\sigma) \equiv \mathbb{E}_{\pi_\theta}[v(s,\omega)]$ and $p(\theta;\sigma) \equiv \mathbb{E}[\tilde{\theta}|\tilde{\theta} : \sigma(\tilde{\theta}) = \sigma(\theta)]$, i.e., the expected material payoff and the perceived image that the strategy induces for each type θ, respectively. We say that a type θ is *separating* under the strategy

[8] This property implies that our equilibrium selection is robust to alternative criteria such as Universal Divinity (Banks and Sobel, 1987) and Never-a-Weak-Best-Response (Kohlberg and Mertens, 1986), as they are equivalent to D1 in monotonic games (see Proposition 1 of Cho and Sobel, 1990)

[9] Since we can always break the tie by selecting the action that gives the highest payoff to the sender, the value function of the maximization problem in (4) is upper semi-continuous, which guarantees that its solution is well-defined. Similarly, since we can always break the tie by selecting the action that gives the lowest payoff to the sender, the value function of the minimization problem in (4) is lower semi-continuous, so a solution is guaranteed to exist as well.

[10] Take any solutions $\bar{\pi}$ and $\underline{\pi}$ of the sender's max- and minimization problems in (4), respectively. Then, to implement the payoff $V = \lambda \underline{V} + (1-\lambda)\bar{V}$ for some $\lambda \in [0,1]$, we may use the following "grand" information structure $\hat{\pi}$: with probability λ, the sender draws a signal s according to $\underline{\pi}$, and with probability $1-\lambda$, according to $\bar{\pi}$. It is straightforward to check that $\hat{\pi} \in \Pi^*$ and $\mathbb{E}_{\hat{\pi}}[v(s,\omega)] = V$, i.e., $\hat{\pi}$ indeed implements V.

σ if $\sigma(\theta') \neq \sigma(\theta)$ for all $\theta' \neq \theta$ (in which case we necessarily have $p(\theta;\sigma) = \theta$). Otherwise, we say that θ is *pooling*. Our main result shows that there exists a *unique* cutoff $\hat{\theta}$ such that all types $\theta < \hat{\theta}$ ($\theta \geq \hat{\theta}$) will be separating (pooling) in any equilibrium that satisfies D1.[11] Moreover, although the D1 criterion may not select a unique equilibrium, it fully pins down the equilibrium payoff of the sender.

Theorem 1. *There is a unique cutoff $\hat{\theta} \in [0,1) \cup +\infty$ such that any strategy $\sigma = \{\pi_\theta\}_{\theta\in\Theta}$ of the sender with $\pi_\theta \in \Pi^*$ for all $\theta \in [0,1]$ is part of a D1 equilibrium if and only if the following two conditions are both satisfied:*

(i) All types $\theta < \hat{\theta}$ are separating, with

$$V(\theta;\sigma) = \bar{V} - \phi \cdot \int_0^\theta \frac{\partial w(x,x)}{\partial p} dx; \tag{5}$$

(ii) All types $\theta \geq \hat{\theta}$ are pooling, with $V(\theta;\sigma) = \underline{V}$ and $p(\theta;\sigma) = \mathbb{E}[\tilde{\theta}|\tilde{\theta} \geq \hat{\theta}]$.

Theorem 1 implies that there exist D1 equilibria — one can always construct a strategy that satisfies (i) and (ii) (recall the property that $\mathcal{V} = [\underline{V}, \bar{V}]$). Exactly which strategy is chosen among the qualified ones is immaterial for the sender, because from her perspective all of them are equivalent in terms of payoffs. As a consequence, the set of D1 equilibria can be Pareto-ranked according to the welfare of the receiver. In later analysis, we will provide examples and applications which also feature payoff equivalence for the receiver, or which permit an analytical description of the equilibria that are extremal in the Pareto ranking.

In what follows, we prove the only-if part of Theorem 1, i.e., that all D1 equilibria necessarily satisfy conditions (i) and (ii), which is instructive as it highlights how the equilibrium outcome is shaped by the tension between the conflicting motives of the sender. The proof of the if-part of the theorem, i.e., that all strategies satisfying (i) and (ii) are part of a D1 equilibrium, is relegated to the Appendix as it is rather mechanical: Plainly, types would not want to mimic each other because conditions (i) and (ii) will be derived (among others) from the on-path incentive compatibility constraints. With attention to detail, one can further construct the appropriate out-of-equilibrium beliefs that prevent off-path deviations and satisfy D1.

[11] We assume that if a type (e.g., the cut-off type $\hat{\theta}$ when $\hat{\theta} \in (0,1)$) is indifferent between separating herself or pooling with some higher types, she would break the tie in favor of the latter. With a continuous type distribution, this tie-breaking rule is inconsequential.

Monotone strategies and incomplete separation. To begin with, we derive some qualitative features of the sender's strategy from her equilibrium incentives. Recall that the sender's central trade-off is between the material persuasion motive and her image concerns. In particular, it is clear that the sender will be willing to sacrifice her material payoff only if that can boost her reputation. Further, since the image concern $w(\cdot)$ satisfies the increasing difference condition in (1), such a "money-burning" incentive will be (strictly) higher for higher types. Lemma 1 below exploits this property and shows that any equilibrium must be monotone in the sense that the interim material payoff of higher types is lower, while their interim image is higher.

Lemma 1. *In any equilibrium, $V(\theta;\sigma)$ is decreasing in θ and $p(\theta;\sigma)$ is increasing in θ.*

Next, we show that a type cannot be pooling unless its associated material payoff is already minimal.

Lemma 2. *In any equilibrium that satisfies the D1 criterion, $\forall \theta \neq \theta'$, if $\sigma(\theta) = \sigma(\theta')$, then $V(\theta;\sigma) = V(\theta';\sigma) = \underline{V}$.*

The intuition behind Lemma 2 is as follows. Given the single-crossing property of the sender's interim preferences, a higher type in a pooling set will be more likely to benefit from an off-path choice that slightly reduces her material payoff than any type lower than her. To be consistent with the D1 criterion, such an unexpected move must convince the receiver that the sender's type is higher than anyone in that pooling set. As a consequence, a pooling type can obtain a discrete gain in image payoff by sacrificing an arbitrarily small amount of material payoff. Plainly, this kind of deviation is not a threat to the equilibrium if and only if the material payoff is already "used up" by the pooling types: they are receiving $\underline{V}$, the lowest possible material payoff, so undercutting is simply not feasible.

Lemmas 1 and 2 jointly imply that, in any D1 equilibrium, when choosing the interim allocation the sender must use a strategy where all types below a cutoff $\hat{\theta}$ separate by monotonically decreasing their material payoff, while all types above $\hat{\theta}$ pool at the lower bound of the material-payoff space. Similar incomplete separation at the top has been established in other contexts (Bernheim, 1994; Kartik, 2009). Indeed, Cho and Sobel (1990) show that this semi-separating structure is inherent to the equilibria selected by D1 in a large class of

costly signaling games where the sender faces a compact signal space. In this regard, the interim game in our setting is similar to a game in which the sender is endowed with money $\bar{V}$ and can "'burn" at most $\bar{V} - \underline{V}$ of it. However, what is quite different is that there are no exogenous costs of actions in our setting. The sender engages in information control and this can be costly for him since it affects the receiver's behavior. Importantly, information control *also* affects the receiver's payoff, unlike in costly signaling games.

The cost of reputation. We now proceed to characterize the endogenous cost of signaling (i.e., the extent to which one's material payoff needs to be sacrificed) for types in the separating interval $[0, \hat{\theta})$. Here, the central idea is to leverage that the sender's utility function is quasi-linear with respect to her reputation. This payoff structure reminds of the standard mechanism design setting with transfers. Thus, naturally, we advance the analysis by applying a classical envelope theorem argument to the (local) incentive compatibility constraints of the sender.[12]

Take any $\theta \in [0, \hat{\theta})$. Note that for sufficiently small $\epsilon > 0$, we have $\theta + \epsilon \in [0, \hat{\theta})$ as well. Incentive compatibility for the type-θ sender implies that

$$\phi \cdot [w(\theta + \epsilon, \theta) - w(\theta, \theta)] \leq V(\theta; \sigma) - V(\theta + \epsilon; \sigma). \tag{6}$$

That is, the image gain for type θ from mimicking $\theta + \epsilon$ is weakly smaller than the associated loss in material utility. Similarly, incentive compatibility for the type $\theta + \epsilon$ implies that

$$\phi \cdot [w(\theta + \epsilon, \theta + \epsilon) - w(\theta, \theta + \epsilon)] \geq V(\theta; \sigma) - V(\theta + \epsilon; \sigma). \tag{7}$$

Combining (6) and (7) and dividing them by ϵ, we have

$$\frac{\phi \cdot [w(\theta + \epsilon, \theta) - w(\theta, \theta)]}{\epsilon} \leq \frac{V(\theta; \sigma) - V(\theta + \epsilon; \sigma)}{\epsilon} \leq \frac{\phi \cdot [w(\theta + \epsilon, \theta + \epsilon) - w(\theta, \theta + \epsilon)]}{\epsilon}.$$

Since $w(\cdot)$ is continuously differentiable, it follows from the squeeze theorem that

$$V'(\theta; \sigma) \equiv \lim_{\epsilon \to 0} \frac{V(\theta + \epsilon; \sigma) - V(\theta; \sigma)}{\epsilon} = -\phi \cdot \frac{\partial w(\theta, \theta)}{\partial p}. \tag{8}$$

[12] See, e.g., Proposition 23.D.2 in Mas-Colell, Whinston and Green (1995). Note that the (exogenous) image concerns play here a role similar to that of "quasi-money" in other design settings (e.g., Kolotilin, Mylovanov, Zapechelnyuk and Li, 2017).

Hence, $V(\cdot;\sigma)$ is also continuously differentiable.

Further, whenever $\hat{\theta} > 0$, the type $\theta = 0$ is in the separating interval and gets the lowest possible image payoff. Thus, incentive compatibility also requires that this type must be earning the highest possible material payoff, i.e., $V(0;\sigma) = \bar{V}$. Combining this boundary condition and the differentiable equation (8), we immediately obtain the payoff formula (5) and conclude that it must hold for all $\theta \in [0, \hat{\theta})$ in any D1 equilibrium.

Uniqueness of the equilibrium cutoff. To complete the proof of Theorem 1, it remains to show that the cutoff $\hat{\theta}$ is unique across all D1 equilibria. The characterization of the equilibrium payoffs on $[0, \hat{\theta})$ implies that the following indifference condition must hold for an *interior* cutoff type $\hat{\theta} \in (0, 1)$:

$$\left(\bar{V} - \phi \cdot \int_0^{\hat{\theta}} \frac{\partial w(x,x)}{\partial p} dx\right) + \phi \cdot w(\hat{\theta}, \hat{\theta}) = \underline{V} + \phi \cdot w\left(\mathbb{E}[\tilde{\theta}|\tilde{\theta} > \hat{\theta}], \hat{\theta}\right). \tag{9}$$

Intuitively, if condition (9) does not hold, then by continuity either some pooling type $\hat{\theta} + \epsilon$ would have a strict incentive to mimic, e.g., the separating type $\hat{\theta} - \epsilon$, where $\epsilon > 0$ is sufficiently small, or vice versa. We rewrite (9) as

$$\frac{\bar{V} - \underline{V}}{\phi} = I(\hat{\theta}) \tag{10}$$

where the mapping $I(\cdot)$ is given by

$$I(\theta) = \int_0^{\theta} \frac{\partial w(x,x)}{\partial p} dx + w(\mathbb{E}[\tilde{\theta}|\tilde{\theta} > \theta], \theta) - w(\theta, \theta)$$

for all $\theta \in [0,1]$. Note that I is strictly increasing.[13] Also, I is continuous in $\hat{\theta}$ because $w(\cdot)$ is continuously differentiable and because the type distribution Γ is absolutely continuous.

We distinguish three cases. First, if

$$I(0) < \frac{\bar{V} - \underline{V}}{\phi} < I(1), \tag{11}$$

an application of the intermediate value theorem implies that (10) admits an interior solution $\hat{\theta} \in (0,1)$, and this solution is unique due to the strict monotonicity.

Second, consider the case $(\bar{V} - \underline{V})/\phi \geq I(1)$ or, equivalently, $\phi \leq \underline{\phi} \equiv (\bar{V} - \underline{V})/I(1)$. Suppose that there would be an equilibrium with cutoff $\hat{\theta} < 1$. Then, all types $\theta < 1$ would strictly prefer separating to pooling with higher types, which contradicts with $\hat{\theta} < 1$. As a result, any equilibrium selected by D1 must be fully separating and we can write $\hat{\theta} = +\infty$ without loss of generality.

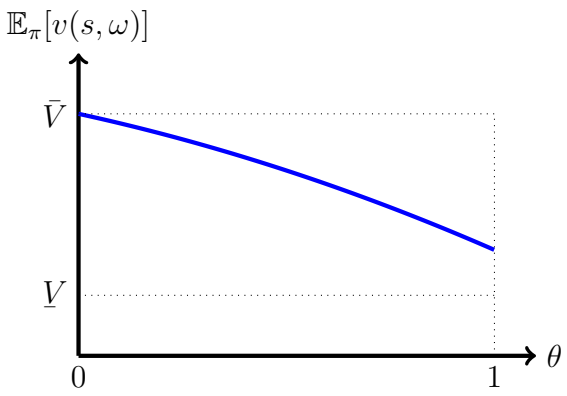
Third, consider the case when $(\bar{V} - \underline{V})/\phi \leq I(0)$ or, equivalently, $\phi \geq \bar{\phi} \equiv (\bar{V} - \underline{V})/I(0)$. Suppose that there would be an equilibrium with cutoff $\hat{\theta} > 0$. Then, all types $\theta < 1$ would strictly prefer pooling with higher types than separating (except that type 0 may be indifferent), which contradicts with $\hat{\theta} > 0$. This implies that all types must be pooling in any equilibrium, and consequently, we have $\hat{\theta} = 0$ as the unique cutoff. □

We close this section with a graphical illustration of the main findings of Theorem 1. Figure 1 presents all three types of the sender's strategy that could emerge in an equilibrium satisfying the D1 criterion. That is, the cases $\phi \leq \bar{\phi}$ (Panel a), $\underline{\phi} < \phi < \bar{\phi}$ (Panel b), and $\phi \geq \bar{\phi}$ (Panel c).

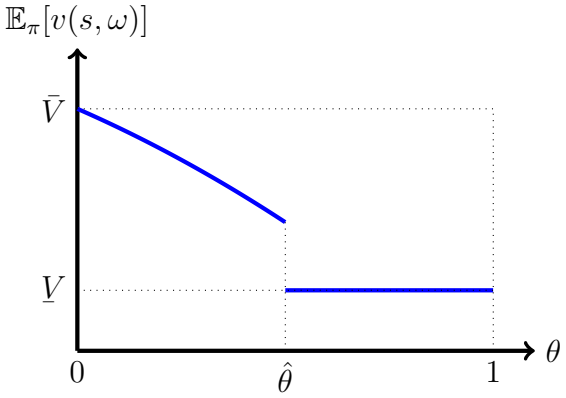
[13] For all $\theta, \theta' \in [0,1]$ with $\theta' < \theta$, we have

$$\begin{aligned} I(\theta) - I(\theta') &= \int_{\theta'}^{\theta} \frac{\partial w(x,x)}{\partial p} dx + \int_{\theta}^{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta]} \frac{\partial w(x,\theta)}{\partial p} dx - \int_{\theta'}^{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta']} \frac{\partial w(x,\theta')}{\partial p} dx \\ &> \int_{\theta'}^{\theta} \frac{\partial w(x,\theta')}{\partial p} dx + \int_{\theta}^{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta]} \frac{\partial w(x,\theta')}{\partial p} dx - \int_{\theta'}^{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta']} \frac{\partial w(x,\theta')}{\partial p} dx \\ &= \int_{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta']}^{\mathbb{E}[\tilde{\theta}|\tilde{\theta}>\theta]} \frac{\partial w(x,\theta')}{\partial p} dx \\ &\geq 0, \end{aligned}$$

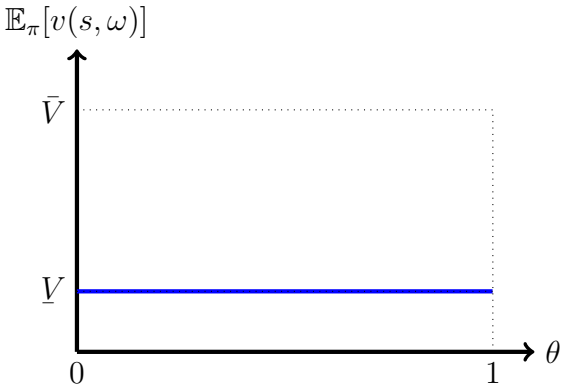
where the strict inequality follows since $w(\cdot)$ has strictly increasing differences, and the weak inequality holds because $w(p,\theta')$ is strictly increasing in p and because $\mathbb{E}[\tilde{\theta}|\tilde{\theta} > \theta] \geq \mathbb{E}[\tilde{\theta}|\tilde{\theta} > \theta']$.



(a) $\phi = 0.3$, fully separating



(b) $\phi = 0.6$, semi-separating



(c) $\phi = 1.2$, fully pooling

Figure 1: Sender's expected material payoff as a function of her type in a D1 equilibrium, with $\theta \sim \mathcal{U}[0,1]$, $w(p,\theta) = p \cdot (\theta + 1)$, $\bar{V} - \underline{V} = 0.6$, and different ϕ.

3.2 Equilibrium Multiplicity and Pareto (In)efficiency

As mentioned, Theorem 1 implies that the sender's interim payoffs are equivalent across all D1 equilibria. In particular, the theorem specifies explicitly the level of material payoff that each sender type will give up in order to separate herself from lower types. Given the abundance of possible information structures, there are, however, manifold ways how types can make such sacrifices. In other words, Theorem 1 does not give a very sharp prediction on which information structure will be used by the sender, which also means that the payoff of the receiver is not necessarily pinned down by the D1 criterion. Since there is no a priori reason to restrict attention to a specific class of information structures, we proceed by (i) providing simple sufficient conditions under which the implications of the sender's image concerns for the receiver's welfare will (or will not) be robust to equilibrium selection, and (ii) analyzing

the Pareto-frontier of the equilibrium set.

To simplify the discussion, we make two additional mild assumptions. First, information is valuable to the receiver: $\bar{U} \equiv \mathbb{E}_{\mu_0}\left[\max_{a\in A} u^R(a,\omega)\right] > \underline{U} \equiv \max_{a\in A} \mathbb{E}_{\mu_0}\left[u^R(a,\omega)\right]$. Second, in the benchmark scenario in which the sender has *no* image concern, the receiver's equilibrium payoff – which we denote by U^* – is uniquely defined.

3.2.1 When will the sender's image concerns be harmful?

When will the presence of the sender's image concerns only do harm to the receiver's welfare, irrespective of which D1 equilibrium is selected? An obvious sufficient condition is that the receiver would be earning his full-information payoff when the sender does not have any image concern. Our next result summarizes this simple observation and goes beyond it by identifying what the best- and worst-case scenarios for the receiver may look like within the set of D1 equilibria.

Theorem 2. *If $U^* = \bar{U}$, the receiver can never benefit from the presence of the sender's image concerns. Moreover, if $\phi < \bar{\phi}$ (so that we have a cutoff type $\hat{\theta} > 0$), then*

(i) there exists a D1 equilibrium in which the receiver is strictly worse off compared to the case without image concerns;

(ii) in any Pareto-optimal D1 equilibrium, the receiver's expected payoff is strictly decreasing with respect to the sender's type θ on the separating interval $[0, \hat{\theta})$;

(iii) in any Pareto-worst D1 equilibrium, the receiver's expected payoff is quasi-convex with respect to θ on the separating interval $[0, \hat{\theta})$.

Intuitively, the conditions of Theorem 2 imply that a no-disclosure protocol is suboptimal for the sender when she is purely guided by material interests since otherwise, the receiver would not have been able to enjoy his full-information payoff. Therefore, an image-concerned sender can always separate herself from those very low types by occasionally sending a completely uninformative signal to the receiver, which obviously engenders a negative "side-effect" on the receiver's payoff. As for the properties of the Pareto-extremal equilibria, our proof mainly exploits the convexity of the set of payoff profiles that can be implemented via information design: For instance, suppose, within the separating interval of a D1 equilibrium, the

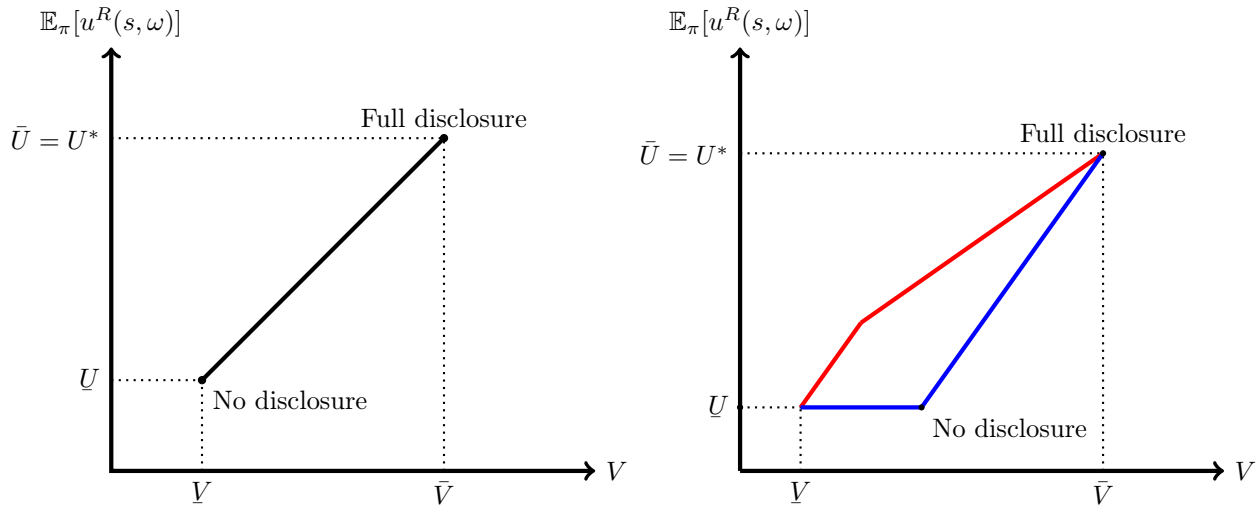


Figure 2: The equilibrium set of implementable payoffs in settings with $U^* = \bar{U}$. Panel (a) represents a game where the preferences of the players are perfectly aligned, with $u^R(a,\omega) = v(a,\omega)$, $\bar{V} = 0.5$ and $\underline{V} = 0.1$. Panel (b) represents a game where the preferences of the players are almost perfectly aligned, with $A = \Omega = \{-1, 0, 1\}$, $\Pr(\omega = 1) = \Pr(\omega = 0) = 0.4$, $\Pr(\omega = -1) = 0.2$ and the payoff functions given by (13). The upper curve (colored in red) in the graph depicts the utility-frontier of the Pareto-optimal D1 equilibria, while the lower curve (colored in blue) corresponds to the utility-frontier of the Pareto-worst D1 equilibria.

receiver's payoff implied by the strategy of a type θ is lower than that of a higher type $\theta' > \theta$. This equilibrium cannot be Pareto-optimal, because of the following argument: Replacing the communication protocol that type θ initially chooses with an appropriate mix of those used by types 0 and θ' will not change the sender's payoff, but will strictly improve the payoff of the receiver. A similar constructive argument (which involves the no-disclosure protocol instead of the one used by type 0) shows that any Pareto-worst D1 equilibrium must be either decreasing or U-shaped with respect to the sender's type. Otherwise, it would have been feasible to further reduce the receiver's payoff without altering the sender's.

In what follows, we exemplify the insights of Theorem 2 within various classic settings from the literature on sender-receiver games.

Example 1: Congruent preferences. Suppose that the preferences of the players are *congruent* with each other in the sense that they agree on the ex-post optimal actions in every state $\omega \in \Omega$:

$$a^* \in \max_{a \in A} u^R(a,\omega) \iff a^* \in \max_{a \in A} v(a,\omega). \tag{12}$$

When (12) holds, it is clear that the material payoff of the sender is maximized when she provides full information to the receiver. Hence, we have $U^* = \bar{U}$, and Theorem 2 applies.

An obvious setting with congruent preferences is when players' material interests are *perfectly aligned*. Namely, when there exists a strictly increasing function $\Psi : \mathbb{R} \to \mathbb{R}$, such that $u^R(a, \omega) = \Psi(v(a, \omega))$ for all $(a, \omega) \in A \times \Omega$. Panel (a) in Figure 2 depicts the set of implementable material payoff profiles for the case of $\Psi(\cdot)$ being a linear function. In this case, the mapping between the *expected payoffs* of the two players is also a linear one. Consequently, the D1 equilibria are not only payoff-equivalent to the sender (as already asserted by Theorem 1), but also to the receiver. A particularly simple equilibrium is one in which the sender always commits to an information structure that either reveals everything (i.e., recommending an ex-post optimal action) or reveals nothing (i.e., recommending an ex-ante optimal action) to the receiver, with the frequency of the former action decreasing in the sender's type. Restricting to this class of equilibrium information structures, a more image-concerned sender (captured by either a higher θ or ϕ) will transmit less information to the receiver, therefore intensifying the negative impact on the latter's welfare.

Perfect alignment of material interests is by far not the only setting that implicates congruent preferences. We illustrate this point by considering a special case with $A = \Omega = \{-1, 0, 1\}$ and the following material payoff functions of the players:

$$u^R(a, \omega) = \begin{cases} 1 & \text{if } a = \omega, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad v(a, \omega) = \begin{cases} 1 & \text{if } a = \omega, \\ 0 & \text{if } a \neq \omega \text{ and } a \neq -1, \\ -1 & \text{if } a \neq \omega \text{ and } a = -1. \end{cases} \tag{13}$$

The interpretation of the above payoff specification is that the material interests of the players are *almost* perfectly aligned. Both players would like to match the action to the true state. However, the action $a = -1$ is somewhat riskier than others for the sender, because she will be additionally punished when the receiver takes it by mistake. By contrast, the receiver is indifferent between different types of errors. Despite the discrepancy in the payoff functions, the congruency condition (12) is satisfied.

Panel (b) in Figure 2 identifies the set of implementable material payoff profiles under some specific prior distribution (see Appendix A.6.1 for the formal construction). Especially, the upper curve in red (the lower curve in blue) pins down, for any given level of the sender's

payoff $V \in [\underline{V}, \bar{V}]$, the maximal (minimal) payoff that the receiver can obtain. Thus, in any Pareto-extremal equilibrium, different sender types will "line up" along these curves to forgo their material utilities, giving rise to the patterns of monotonicity/quasi-convexity highlighted by Theorem 2.

Example 2: Quadratic loss. Let $A = \Omega = [0, 1]$, $u^R(a, \omega) = -(a-\omega)^2$, and $u^S(a, \omega, \theta, \eta) = -(a - a^*(\omega, \theta))^2 + \phi \cdot w(p(\eta), \theta)$. Specifically, the sender's bliss point is given by $a^*(\omega, \theta) = f(\theta) \cdot \omega + g(\theta)$. Communication games in which players' preferences take the form of such a quadratic loss function were popularized by the seminal work of Crawford and Sobel (1982), and they have received considerable attention in the information design literature (see, e.g., Galperti, 2019; Jehiel, 2015; Kamenica and Gentzkow, 2011; Smolin and Yamashita, 2022; Tamura, 2018). In the classic information design setting without image concerns, the players' incentives are purely governed by their disagreement over the optimal action plan: while the receiver wants to exactly match the state ($a = \omega$), the sender may have a systematically different target ($a = a^*(\omega, \theta)$). The current example, as well as Example 5 in the next subsection, examine the conditions under which introducing image concerns would mitigate or amplify the above misalignment of preferences and consequently lead to more or less information transmitted in equilibrium.

In Appendix A.6.2, we show that if $f(\theta) > 0.5 \ \forall \theta \in [0, 1]$ is satisfied, then the initial quadratic-loss game is equivalent to one in which the sender has the material payoff function $v(a, \omega) = u^R(a, \omega)$ and the image payoff function $\hat{w}(p(\eta), \theta) = w(p(\eta), \theta)/(2f(\theta) - 1)$. This transformation manifests that the players' interests are sufficiently aligned under the current specification, insomuch that a sender purely guided by material interests would be willing to share all information with the receiver. However, if function $\hat{w}(\cdot)$ satisfies the key condition (1) – which can be the case, for instance, if $f'(\cdot) < 0$, meaning that higher types put less weight on the state-dependent term relative to the state-independent target $g(\cdot)$ – then both Theorem 1 and Theorem 2 apply. They jointly imply that all types (except possibly type 0) will withhold information from the receiver for signaling purposes. Moreover, given that $v(a, \omega) = u^R(a, \omega)$, the equilibrium payoffs of both the sender and the receiver are uniquely pinned down by the D1 criterion.

Theorem 2 and the examples following it are related to the literature on "bad reputation" in repeated games (see, e.g., Ely *et al.*, 2008; Ely and Välimäki, 2003). An overarching finding of

this literature is that reputational concerns harm a long-lived player who repeatedly interacts with short-lived players if they are based on a desire to separate from a bad type rather than to mimic a good commitment type (see the discussion in Mailath, Samuelson *et al.*, 2006). The forces behind our results are quite different and more subtle: the sender tries to separate herself from the type that is *least image-concerned*, which requires her to avoid taking the strategy that would be *endogenously* chosen by the latter. In the current set-up, that strategy happens to be the one that maximizes the material payoffs of both players.

3.2.2 When will sender's image concerns be beneficial?

We now study when the presence of image concerns can only be beneficial. Analogous to the previous subsection, we focus on settings in which the following simple sufficient condition holds: a sender who acts out of pure material interest will implement the *no-information payoff* for the receiver. Theorem 3 below summarizes some key properties of the equilibrium set in such settings.

Theorem 3. *If $U^* = \underline{U}$, the receiver can never be harmed by the presence of the sender's image concerns. Moreover, if $\phi < \bar{\phi}$ (so that we have a cutoff type $\hat{\theta} > 0$), then*

(i) there exists a D1 equilibrium in which the receiver is strictly better off due to the presence of the sender's image concerns;

(ii) in any Pareto-optimal D1 equilibrium, the receiver's expected payoff is quasi-concave with respect to the sender's type θ on the separating interval $[0, \hat{\theta})$.

(iii) in any Pareto-worst D1 equilibrium, the receiver's expected payoff is increasing with respect to θ on the separating interval $[0, \hat{\theta})$;

Both the proof and the intuition of Theorem 3 are analogous to Theorem 2, and therefore omitted to avoid repetition. Below, we illustrate the main insights of the theorem through several examples.

Example 3: No gain from persuasion. Kamenica and Gentzkow (2011) characterize when a sender purely driven by material interests can benefit from persuasion. That is when

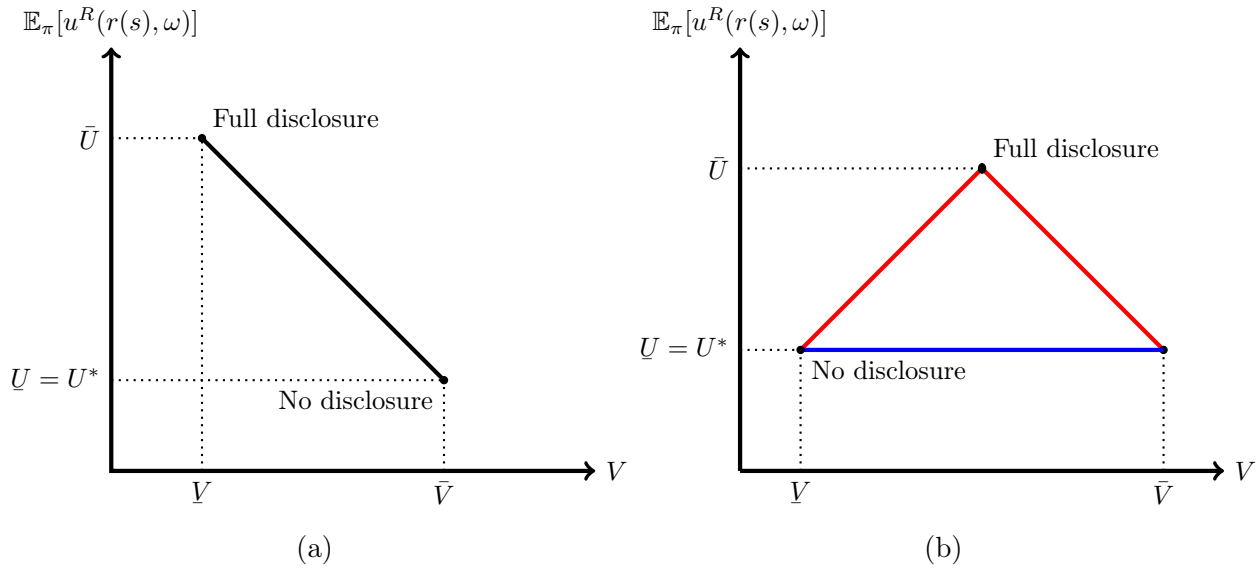


Figure 3: The equilibrium set of implementable payoffs in settings with $U^* = \underline{U}$. Panel (a) represents a game where the players have exactly opposite interests over the material outcomes, with $u^R(a,\omega) = -v(a,\omega)$, $\bar{V} = 0.5$ and $\underline{V} = 0.1$. In panel (b), we have a game with partially conflicting interests as described in Example 5. The upper curve (coloured in red) in the graph depicts the utility-frontier of the Pareto-optimal D1 equilibria, while the lower curve (coloured in blue) corresponds to utility frontier of the Pareto-worst D1 equilibria.

she can do *strictly* better than providing no information (or always recommending an ex-ante optimal action) to the receiver. When this is *not* the case, $U^* = \underline{U}$ obviously holds, so Theorem 3 applies.

A concrete setting where the sender would not want to share any information in the absence of image concerns is when players have *opposite* material interests. Namely, when there exists a strictly *decreasing* function $\Psi : \mathbb{R} \to \mathbb{R}$, such that $u^R(a,\omega) = \Psi(v(a,\omega))$ for all $(a,\omega) \in A \times \Omega$. Panel (a) in Figure 3 depicts the set of implementable material payoff profiles in such a game. Similar to Example 1, the function $\Psi(\cdot)$ is chosen to be linear, which gives rise to the linear mapping between the players' expected payoffs as we see from the figure. This property ensures that all D1 equilibria are payoff-equivalent to both the sender and the receiver. A particularly simple equilibrium is one in which the sender always commits to an information structure that reveals either everything or nothing about the true state, with the frequency of the former action increasing in the sender's type. In this case, a more image-concerned sender (captured by either a higher θ or ϕ) will transmit more information to the receiver, therefore boosting the positive impact on the latter's welfare.

Example 4: Quadratic loss (continued). Consider again the quadratic-loss games that we introduced in the previous subsection. In Appendix A.6.2, we show that under the condition $f(\theta) < 0.5$ $\theta \in [0,1]$, the initial game will be strategically equivalent to one in which the sender has the material function $v(a,\omega) = (a-\omega)^2$ and the image payoff function $\hat{w}(p(\eta),\theta) = \mathbb{E}_\eta[\tilde{\theta}]/(1-2f(\theta))$. Thus, $v = -u^R$ and we effectively have a game with strictly opposed interests. Provided that condition (1) holds for $\hat{w}(\cdot)$ – which can be the case, for instance, if the interests of higher types are more aligned with the receiver in the sense that $f'(\theta) > 0$ $\forall\theta \in [0,1]$ – then both Theorem 1 and Theorem 3 apply. Thus, the presence of image concerns will trigger all sender types (except possibly type 0) to share information with the receiver, which they would be reluctant to do otherwise. In addition, because $u^R(a,\omega) = -v(a,\omega)$, the payoff equivalence implied by the D1 criterion holds not only for the sender but also for the receiver.

Next, we introduce two widely-studied examples in which the sender would partially disclose the state if only the persuasion motive is present, yet the receiver's payoff remains to be minimal. This demonstrates that the applicability of Theorem 3 is *not* limited to settings in which the sender would not share any information in the absence of image concerns. Intuitively, the optimality of partial disclosure may be compatible with the premise $U^* = \underline{U}$ of Theorem 3, because having access to partial information does not guarantee that the receiver can do *strictly* better *on average* than taking his prior-optimal action. This observation is important and may prove useful beyond the examples because it is known that partial disclosure is optimal in many pure persuasion settings. For instance, Jehiel (2015) shows that this is typically the case when the information of the sender is higher dimensional than the action space of the receiver; Kolotilin and Wolitzky (2020) and Kolotilin, Corrao and Wolitzky (2022a) provide similar results in a setting that allows utilities of the sender and receiver to be non-linear in the state.[14]

Example 5: State-independent sender preferences, I. Suppose that $A = \Omega = \{0,1\}$, $v(a,\omega) = a$, and $u^R(a,\omega) = \mathbb{1}_{a=\omega}$. Thus, while the receiver wants to match the state, the sender's preference over material outcomes is state-independent: she always prefers the

[14]See, e.g., Theorem 2 in Kolotilin and Wolitzky (2020). Optimal partial disclosure has been shown to take the form of censorship (Kolotilin, Mylovanov and Zapechelnyuk, 2022b), nested intervals (Guo and Shmaya, 2019), (p)-pairwise signals (Kolotilin and Wolitzky, 2020; Terstiege and Wasser, 2022), or conjugate disclosure (Nikandrova and Pancs, 2017).

receiver to take the high action. This persuasion setting is most vividly embodied by the prosecutor-judge example in Kamenica and Gentzkow (2011). Since the state space is binary, we use μ_0 to denote the prior likelihood of the state being $\omega = 1$. We assume $\mu_0 \in (0, 0.5)$ so that $a = 0$ is the receiver's optimal action given the prior. Clearly, releasing no information minimizes the sender's material payoff. At the same time, Kamenica and Gentzkow (2011) show that partial information disclosure is optimal for the sender when she has no image concerns. Nevertheless, under the optimal disclosure policy, the receiver weakly prefers his prior-optimal action regardless of the signal realization, so her expected payoff is the same as under no information (i.e., $U^* = \underline{U}$). Hence, all results of Theorem 1 and Theorem 3 apply.

We present two simple classes of information structures that one may use to describe the Pareto-optimal and the Pareto-worst D1 equilibria in closed form, respectively. For every $q \in [0, 2\mu_0]$, define an information structure $\bar{\pi}^q$ as follows: Conditional on the true state, the signal $s = 1$ is drawn with probability

$$\bar{\rho}(\omega; q) = \begin{cases} \min\left\{\frac{q}{\mu_0}, 1\right\} & \text{if } \omega = 1, \\ \max\left\{\frac{q-\mu_0}{1-\mu_0}, 0\right\} & \text{if } \omega = 0. \end{cases} \tag{14}$$

With the remaining probability $1 - \bar{\rho}(\omega; q)$, the signal $s = 0$ is sent to the receiver. One can check that $\bar{\pi}^q$ is incentive-compatible, and it induces the receiver to choose the action $a = 1$ exactly with probability q. While there can be other information structures that induce the same marginal distribution of actions, all of them will be Pareto-dominated by $\bar{\pi}^q$ (see Appendix A.6.3 for a formal proof). For instance, consider the information structure $\underline{\pi}^q$ defined as follows: Conditional on the true state, the signal $s = 1$ is drawn with probability

$$\underline{\rho}(\omega; q) = \begin{cases} \frac{q}{2\mu_0} & \text{if } \omega = 1, \\ \frac{q}{2(1-\mu_0)} & \text{if } \omega = 0. \end{cases} \tag{15}$$

With the remaining probability $1 - \underline{\rho}(\omega; q)$, the signal $s = 0$ is sent to the receiver. With this information structure, the sender can also nudge the receiver to choose the high action with probability q. However, the probability that the receiver takes the *right* action is just $1 - \mu_0$ under $\underline{\pi}^q$, which he could also achieve by simply sticking to his prior-optimal action $a = 0$. This is clearly the worst possible outcome for the receiver, so he would clearly prefer $\bar{\pi}^q$ over $\underline{\pi}^q$. All things considered, there must exist a Pareto-optimal (Pareto-worst) equilibrium in

which each sender type θ uses the information structure $\bar{\pi}^{q(\theta)}$ ($\underline{\pi}^{q(\theta)}$), and in which $q(\theta)$, the total probability that the receiver would take the action $a = 1$, is decreasing in the sender's type. Panel (b) in Figure 3 depicts the receiver welfare in both equilibria, delineating the whole set of implementable payoff profiles for the receiver.

A salient feature of the Pareto-optimal equilibrium is that the receiver's welfare can be non-monotone in the sender's type. This non-monotonicity arises as follows: lower type can signal their type and separate by releasing more information about the state. However, the cost of separation for these low types may be so high that already an intermediate type is required to provide full information in order to separate. Then, even higher types can only signal their type by sacrificing further material utility in ways that also harm the receiver. By contrast, in the Pareto-worst equilibrium, all sender types minimize the receiver's payoff to his reservation utility $\underline{U}$.

Example 6: State-independent sender preferences, II. Let $A = \{0, 1\}$, $\Omega = [0, 1]$, $v(a, \omega) = a$ and $u^R(a, \omega) = a \cdot \omega + (1 - a) \cdot \underline{u}$, where $\underline{u} \in (0, 1)$ can be interpreted as the value of the receiver's outside option. We assume that $\underline{u} > \mathbb{E}_{\mu_0}[\omega]$. Thus, the receiver's default action is $a = 0$, and $\underline{u}$ will also be his expected payoff under no information. Further, in the absence of image concerns, the optimal strategy of the sender would extract all the surplus from the receiver (see Section V. B in Kamenica and Gentzkow (2011)). Taken together, we have $U^* = \underline{U} = \underline{u}$, so both Theorem 1 and Theorem 3 can be applied to study this example.

3.2.3 When will the welfare implications be ambiguous?

In general, the receiver's payoff may be strictly between his full- and no-information payoffs in the canonical setting without image concerns. Our last formal result confirms that in this case, whether the sender's image concerns will be beneficial or detrimental for the receiver is likely to be uncertain.

Theorem 4. *If $U^* \in (\underline{U}, \bar{U})$, it can depend on the selected equilibrium and the type distribution if the receiver benefits from the presence of the sender's image concerns or if he is harmed by it. In particular, provided that ϕ is sufficiently small, there always exist both (i) a D1 equilibrium in which the receiver is strictly better off and Blackwell-more information is transmitted and (ii) a D1 equilibrium in which the receiver is and strictly worse-off and Blackwell-less information is transmitted, relative to the setting without image concerns.*

As we alluded before, the ambiguous effect of image concerns is largely due to that standard refinements, including the D1 criterion, do not fully pin down the structure of the sender's equilibrium strategy, although they necessitate that the sender's interim payoffs are equivalent across all equilibria. The vital obstacle is that standard refinements depend on discerning unreasonable payoff incentives of the sender, e.g., the D1 criterion rules out equilibria with off-path beliefs that put mass on types who gain less from deviation. However, the abundance of possible information structures allows diverse choices that lead to the same payoff for the sender. Thus, these choices of information structures cannot be further differentiated by standard refinements, notwithstanding the possibility of having vastly different implications for the receiver's welfare.

Remark on optimal information structures. We view the multiplicity of equilibrium information structures in our model as a qualification of the information design approach rather than as a drawback. Following Schelling (1980), one may interpret the multiplicity as a manifestation of different cultures of communication. As Myerson (2009) emphasizes, selecting among multiple equilibria is a "fundamental social problem", and recognizing this problem "can help us to better understand the economic impact of culture".

Applying Schelling's approach to information design, external factors and details of a specific application can be used to qualify a class of information structures and thereby select an equilibrium. For example, in many settings, the manipulation of data is constrained by monitoring efforts, plausibility, or potential social and legal consequences. Therefore, outright fabrication of new data may be infeasible or prohibitively costly. However, partial omissions and deletions may not be detected easily and be feasible. In such settings, it might be natural to assume that the sender is restricted to strategies that *censor* the information, and we should therefore focus on the equilibrium in which the sender indeed uses such strategies. We will further elaborate on this point within the application in Section **??**, where a mid-manager of a firm can censor the information flow to subordinate employees.

On a related note, our reduced-form characterization of equilibria adds to the recent discussion of a common critique of the information design approach. The design approach distinguishes itself from other theories of sender-receiver games by allowing the sender to choose (and commit to) *any* information structure. The critique, as summarized by Kamenica, Kim and Zapechelnyuk (2021), is that "optimal information structures can be infeasible or difficult

to implement in practice". A strand of the information design literature has addressed the above issue by identifying sufficient conditions for simple information structures to be optimal among all information structures (e.g. Ivanov, 2021; Kolotilin *et al.*, 2022b; Kolotilin and Wolitzky, 2020). We show that a class of simple information structures (e.g. the censoring of available information) is consistent with equilibrium requirements in an extended game, under the condition that it can fully implement all possible material payoffs of the sender. Hence, this condition can serve as a formal justification for focusing on some specific class of information structures in applications.

4 Applications

4.1 Self-Signaling and Willful Ignorance

Since the sender and the receiver can be interpreted as two selves of the same agent, our model applies to situations of self-signaling (Bodner and Prelec, 2003). In a typical self-signaling situation, an individual forms beliefs about her own abilities (Köszegi, 2006; Schwardmann and Van der Weele, 2019), moralities (Bénabou and Tirole, 2006; Chen and Heese, 2021; Grossman and Van der Weele, 2017) or other inner characteristics such as self-control (Bénabou and Tirole, 2002, 2004) based on her past conduct, from which she may also derive a direct flow of utility.

Our model is similar to, e.g., Bénabou and Tirole (2002) and Grossman and Van der Weele (2017), in that the signaling is via the sender-self's information choice. The main difference is that we do not restrict the sender-self's choice to a prespecified class of information structures. The assumption that the sender can fully commit to any information structure, which plays a central role in the Bayesian persuasion literature and is often considered as somewhat extreme, can be quite natural in the dual-self setting. It simply captures that the information acquisition is public, that is, the sender-self cannot distort information or knowingly lie to the receiver-self. This point is most evident with a binary state because it has been shown that in such settings Bayesian persuasion is equivalent to a dynamic information acquisition game where information arrives according to a drift-diffusion process (e.g. Chen and Heese, 2021; Henry and Ottaviani, 2019; Morris and Strack, 2019).[15]

[15]The drift-diffusion model is a well-established model of information processing in neuroeconomics and psychology. See, e.g., Fehr and Rangel (2011); Fudenberg, Newey, Strack and Strzalecki (2020); Krajbich, Oud and Fehr (2014); Ratcliff, Smith, Brown and McKoon (2016) and the references therein.

To showcase the applicability of our model in situations of self-signaling, consider an agent who is faced with a mental task. Both selves of the agent share a state-dependent material payoff $v(a, \omega)$ from an action choice a. Ultimately, the receiver-self of the agent will decide which action a to take. Nevertheless, the sender-self can "cheat" by acquiring some information about the state. Formally, she can choose a joint distribution of the state and signal, and then make action recommendations to the receiver-self. The sender-self also has a private type θ that captures the agent's ability to figure out the solution to the task without any informational assistance. Specifically, the sender-self knows that, with probability $f(\theta)$, the receiver-self will be able to directly observe the true state in the action-taking stage, regardless of which information structure the sender-self has chosen. The probability $f(\theta)$ is strictly increasing in θ. So, higher types are associated with higher abilities. The agent further derives a "diagnostic utility" $\phi \cdot \mathbb{E}_\eta[\tilde{\theta}]$ from being perceived as a high type by her receiver-self. It is straightforward to check that this dual-self game is equivalent to one in which the sender has the utility function $u^S(a, \omega, \theta, \eta) = v(a, \omega) + \phi \cdot (\delta f(\theta) + \mathbb{E}_\eta[\tilde{\theta}])/(1 - f(\theta))$, where $\delta \equiv \mathbb{E}_{\mu_0}[\max_{a \in A} v(a, \omega)]/\phi$ is a constant, while the receiver has the utility function $u^R(a, \omega) = v(a, \omega)$.[16]

Our previous results for such common-value settings (see Example 1 in Section 3.2.1) are quite clear-cut. In equilibrium, the higher types will "self-handicap" by acquiring less accurate information, for the goal of boosting their egos. Such handicapping behavior, which was similarly found in Bénabou and Tirole (2002), unambiguously reduces the material payoff of the agent. This result contributes to the growing body of research on information avoidance, which studies the widely-documented phenomenon that decision makers may willfully abstain from obtaining free and useful information for, e.g., psychological or cognitive reasons. For an excellent survey on this topic, see Golman, Hagmann and Loewenstein (2017).[17]

4.2 On Transparency in Organizations

We revisit the question of transparency in organizations, as studied by Jehiel (2015). More specifically, the question is when a manager (sender) of an organization prefers being opaque

[16] A sender of type θ chooses $\pi \in \Pi^*$ to maximize $(1-f(\theta)) \cdot \mathbb{E}_\pi[v(s, \omega)] + f(\theta) \cdot \mathbb{E}_{\mu_0}[\max_{a \in A} v(a, \omega)] + \phi \cdot \mathbb{E}_\eta[\tilde{\theta}]$ (subject to the obedience constraints from the action recommendations s.) This is equivalent to maximizing $\mathbb{E}_\pi[v(s, \omega)] + \phi \cdot (\delta f(\theta) + \mathbb{E}_\eta[\tilde{\theta}])/(1 - f(\theta))$. Note that the function $w(p, \theta) = (\delta f(\theta) + p)/(1 - f(\theta))$ is continuously differentiable and satisfies our condition (1).

[17] Our result is also related to a strand of literature in social psychology, which documents that individuals exhibit a wide array of behavior that is factually bad for them but presumably useful for self-presentation; see, e.g, Crocker and Park (2004); Schlenker (2012).

about what she knows in a moral hazard interaction with a worker (receiver). In what follows, we identify a new force that drives intransparency in organizations, which rests on the reputational concerns of the manager.[18]

Reputational concerns in organizations might arise internally from the norms or guidelines of a company and the explicit or implicit incentives of employees to signal compliance. To make this point concrete, we follow Jehiel (2015)'s motivating example and formulate the moral hazard interaction through a preference setting with $A = \Omega = [0, 1]$ and quadratic losses à la Crawford and Sobel (1982). The worker's utility function is $u^R(a, \omega) = -(a - \omega)^2$, so his effort bliss point equals exactly to the state ($a = \omega$). However, from the viewpoint of the company's senior management, the effort bliss point is $\beta \cdot \omega$, where $\beta > 1$. Thus, the ideal level of effort is systematically higher for the senior management than for the worker. The (mid-level) manager's preferences over effort extrapolate between those of her boss and her subordinate, and this is captured by a material payoff function $v(a, \omega, \theta, \eta) = -(a - f(\theta) \cdot \omega)^2$, where $f(\theta) \equiv (\beta - 1) \cdot \theta + 1$. Note that $f(\cdot)$ is strictly increasing and satisfies $f(0) = 1$ and $f(1) = \beta$, reflecting the idea that higher types have internalized the senior management's point of view more strongly. Last, the manager likes to be perceived as a high type, that is, as being "compliant" to the preferences of the higher-ups. Formally, the manager receives an image payoff $\phi \cdot \theta \cdot \mathbb{E}_\eta[\tilde{\theta}]$, where η is interpreted as the senior management's belief about the manager's type. Overall, our payoff specification posits that higher types care more about the impression that they leave to the boss. This seems in line with these types internalizing the senior management's point of view less strongly. Presumably, these are the types that are more committed to a career in the current company.

What do the incentives of signaling compliance to the higher-ups imply in terms of transparency and organizational performance? Similar to Jehiel (2015), we have a fully transparent benchmark in the current quadratic-loss setting: If the manager has no reputational concerns ($\phi = 0$), then all manager types would fully communicate all information about the state to the worker.[19] However, Theorems 1 and 2 jointly imply that, when the manager worries about the (explicit or implicit) review of her compliance by the senior management, she will almost necessarily involve in strategies that hide information from the worker. Thus, the motive of "pleasing the boss" can be a compelling source of intransparency in organizations.

[18] Jehiel (2015) focuses on two distinct forces that make full transparency suboptimal, which concern either the sensitivity or the concavity of the players' utilities over actions in different states.

[19] See Kamenica and Gentzkow (2011). For details specific to our setting, see also Appendix A.6.2 and the analysis of Example 2 in Section 3.2.1.

This lack of transparency, in turn, harms organizational performance, because in expectation the worker's effort choice will be further away from the company's bliss point, i.e. the senior management's, compared to the fully transparent case.

It is perhaps unrealistic to think that the desire to establish a reputation among one's colleagues would always have an unequivocally negative effect on the transparency of the organization. For instance, instead of signaling compliance to the higher-ups, in some workplaces managers may want to signal altruism to their subordinates (Ellingsen and Johannesson, 2008). In those settings, it might seem natural to expect that the concern for reputation would encourage the manager to share more information with the workers, therefore enhancing the transparency of the organization. The caveat here is that one must consider what the manager would have done in the absence of such reputational concerns. It is possible that, with pure persuasion motives, the manager would disclose partial information about the state to the worker. Then, according to Theorem 4, whether the manager's reputational concerns will drive a more or less transparent organization may hinge on equilibrium selection, which, in turn, can be determined by factors such as social norms and/or the corporate culture of the organization (such informal factors of the organization are superbly surveyed and discussed in Hermalin, 2001; Kreps, 1990).

Taken together, the application in this section provides insights into a recent debate on the downsides of hierarchical structures in organizations. Specifically, there are concerns that since attention will naturally be directed up the hierarchy, performance in traditional hierarchical organizations may suffer from the managers focusing too much on "pleasing their bosses" rather than "helping their teams" (Dillon, 2017). To this end, our application provides a game-theoretic model in which pleasing-one's-boss schemes arise and are shown to harm the organization. Our model also offers a novel rationale for why many (but certainly not all) companies nowadays rely on committees to conduct performance evaluations instead of delegating these decisions solely to direct superiors.[20] Intuitively, such arrangements should mitigate the managers' signaling concerns, which, according to our theory, can potentially enhance transparency and improve the performance of the organization.

[20]In 2011, the Society for Human Resource Management surveyed 510 organizations with 2,500 or more employees and found that a majority (54%) of these organizations use formal committees as part of their performance evaluation process.

4.3 Populist Sentiments and Policy Stagnation

A classic question in political economy is why the political system often fails to adopt reforms that experts consider efficiency-enhancing. This is particularly puzzling since political leaders stress the importance of science and evidence-based politics for progress and growth.[21] Fernandez and Rodrik (1991) show that such "resistance to reform" can be explained by asymmetric information. Namely, it may arise when a majority of a democratic public is skeptical about the reform's consequences, in that they expect to be impacted negatively themselves, even though ex-post a majority benefits and the reform is welfare-enhancing. It is natural to ask: when and why do such beliefs exist in the first place?

To this end, we describe a (stylized) model of politics in which information may flow from experts/science to a politician, and further to the public who then accepts or rejects a reform, according to their beliefs. The public's prior opinion is marked by reform skepticism. That is, in the benchmark in which the politician provides no information to the public, the reform is rejected and policies stagnate, as in Fernandez and Rodrik (1991). These prior opinions may capture that the public suspects the reform initiative of being pushed by private interests. In our model, the politician will be concerned about not being perceived as a puppet of such private interests. One may think that such reputational concerns of the politician may facilitate the information flow to the public. On the contrary, we show that these concerns may sustain the skeptical beliefs of the public in equilibrium. In particular, we show that this may happen even when all agents benefit from reform implementation ex-post and when complete information about the reform can be shared.

The formal model is as follows. In the first stage, the politician can acquire information about a binary state $\omega \in \Omega = \{0, 1\}$, which is payoff-relevant for a reform that is currently debated in public media. To do so, she can commission a study. We think of a study π as a mapping that specifies a distribution of results for each of the two states. For instance, the politician can appoint an unbiased expert who truly knows the subject to lead the study, which would allow her to always uncover the true state. Alternatively, the politician can ask an expert who is known to be biased, e.g., towards the reform to investigate the matter, in which case a result supporting the reform probably would be less informative about the state than a result opposing it.

[21] See, e.g., Mallapaty (2022), Prillaman (2022), and the editorial note "Politics will be poorer without Angela Merkel's scientific approach" in *Nature*.

In the second stage, the politician observes the result of the study and then chooses between either keeping it private or disclosing it to the public in the form of a verifiable report. Implementing the reform ($a = 1$) enhances the citizens' welfare by 1 if $\omega = 1$ and otherwise reduces it by 1, relative to maintaining the status quo ($a = 0$). The public prior opinion is marked by reform skepticism, i.e., the common belief is that $\mu_0 \equiv \Pr(\omega = 1) \in (0, 0.5)$ ex-ante. Hence, only when the disclosed result is sufficiently compelling to overcome those initial predispositions, the politician's communication is effective and will lead to the adoption of the reform. The politician receives a state-independent payoff $w_1(\theta) > 0$ if the reform is adopted, where $w_1(\cdot)$ is strictly decreasing in her private type $\theta \in [0, 1]$; otherwise, her payoff at this stage is zero. The interpretation is that θ measures the personal interest that the politician has in the reform or more broadly how "corrupted" she is, presumably by the private interests that the public worries about, with higher types being less corrupted.

The third and last stage serves to create reputational concerns of the politician. In this stage, the politician runs for election.[22] Upon winning the election, more corrupt politicians will act less in the interest of the citizen. In particular, the citizens receive a payoff $\theta + \epsilon$ if a politician of type θ is elected to office, where ϵ is drawn according to a continuous cumulative distribution function G with full support. We fix the citizens' expected payoff from electing an alternative candidate as $\underline{u}$, which is assumed to satisfy $G(\underline{u}) \in (0, 1)$. Each citizen privately observes the preference shock ϵ but not θ when he makes the voting decision. Nevertheless, the citizens make a Bayesian inference about the politician's type from the latter's choices in the previous stages. As a result, a citizen would vote for the politician if and only if $p+\epsilon \geq \underline{u}$, where p is the former's posterior expectation regarding the latter's type. This implies that the likelihood of the politician winning the election will be $1 - G(\underline{u} - p)$. The politician's gains from being appointed to office are subsumed into one utility function $\phi \cdot w_2(\theta)$, which is strictly positive for all $\theta \in [0, 1]$. We assume that the ratio $w_2(\theta)/w_1(\theta)$ is continuously differentiable and strictly increasing in θ. That is, less corrupt types have higher incentives to be elected.[23] Thanks to the single-crossing property implied by the ratio $w_2(\cdot)/w_1(\cdot)$ being strictly increasing, the second-stage electoral incentives are effective: They create the desire

[22]The effect of concerns about future elections is studied in the literature on electoral accountability. Excellent surveys are Ashworth (2012) and Duggan and Martinelli (2017).

[23]Given that $w_1(\cdot)$ is strictly decreasing, the assumption $(w_2(\cdot)/w_1(\cdot))' > 0$ is satisfied if all types are purely office-motivated (i.e., $w_2(\theta)$ is constant), a setting that is widely studied in the literature on electoral competition (Persson and Tabellini, 2002). Further, as we formally show in Appendix A.8.2, this monotonicity assumption can also be derived from a setting where the politician is both office- and policy-motivated.

to impress the voters.[24]

In Appendix A.8, we show in detail how this model can be solved in reduced form with the previous analysis. Specifically, we show that the equilibrium problem maps into our setting by specializing Example 4 of Section 3.2.

Here, we sketch the logic of the equilibrium and that the equilibrium mechanism has a compelling interpretation in terms of "populist sentiments". The main equilibrium force is that higher electoral incentives of the third stage, as measured by ϕ, make communication more "populist", in the sense that the revealed study results more often lead the public to (weakly) update towards the politician being less corrupt. This interpretation follows the standard definition in political science and political philosophy that defines populism as a political strategy supporting the people in their struggle against the privileged elite [25]

What matters is that this populism effect on the politician's communication need not translate monotonically into the welfare of the public. Precisely, the results from Section 3.2 show that the public's welfare from the reform choice is *non-monotone* in ϕ. When ϕ is sufficiently small, welfare increases in ϕ in the Pareto optimal equilibrium; compare to panel (a) of Figure 1. However, when ϕ is sufficiently large, the reputational incentives "over-discipline" the politician. Then, all politician types conform to communicating in line with the prior skepticism of the reform; they pool on recommending the status quo with probability one in all states; compare to panel (c) of Figure 1.[26] Even when the reform is welfare-enhancing, the public's equilibrium beliefs equal the skeptical prior so that the status quo is kept and policies stagnate.

5 Conclusion

In many economic situations, a sender strategically communicates with a receiver with a two-fold goal. First, to influence the decision-making of the receiver. Second, to steer the perception of certain unobserved characteristics of herself, e.g., loyalty, integrity, or unselfishness. The existing work on strategic communication focuses on the first goal, the sender's ability to influence the receiver's action (see, e.g., Grossman, 1981; Kamenica and Gentzkow,

[24] One can show that, when more corrupt types have higher incentives to be elected, that is when $(w_2(\cdot)/w_1(\cdot))' < 0$, then, the election concerns have no effect: the unique equilibrium is so that all types play the equilibrium strategy of the benchmark game with $\phi = 0$.

[25] See the American heritage dictionary, and also Mudde (2004), Acemoglu, Egorov and Sonin (2013), and the recent Vox debate on populism.

[26] This resembles results about conformity in Bernheim (1994).

2011; Milgrom, 1981) and signaling is considered a means to this end (see, e.g., Crawford and Sobel, 1982; Green and Stokey, 2007).

We have provided and analyzed a model in which there is a strategic tension between the two goals. The model builds upon the canonical and general framework of Bayesian persuasion by Kamenica and Gentzkow (2011). Unlike in the standard framework, next to the persuasion motive, the sender has a second motive, namely to signal about her own type. These signaling concerns can take various interpretations, e.g., in terms of psychological preferences (Geanakoplos *et al.*, 1989), or in terms of reputational concerns in regards to a continuation game.

In view of the diverse interpretations of the model, the framework may prove versatile in future applied work. We have provided several applications to showcase this. Future work may relax some of the restrictions made on the environment in this paper, and thereby study additional effects. For example, when allowing for correlation between the state and the sender type, there is an additional channel of learning about the state through the equilibrium signal about the type. This additional channel may alter the interaction between persuasion and signaling.

Appendix

A.1 The Single-Crossing Property

Lemma A1. *Take any two expected material payoffs $V, V' \in \mathcal{V}$ with $V > V'$ and any two receiver beliefs $\eta, \eta' \in \Delta(\Theta)$. If $\theta \in [0,1]$ is indifferent between (V, η) and (V', η'). then*

(a) all types $\theta' < \theta$ strictly prefer (V, η) over (V', η),

(b) all types $\theta' > \theta$ strictly prefer (V', η') over (V, η).

Proof. Indifference of type θ means

$$V - V' = \phi \cdot [w(p(\eta'), \theta) - w(p(\eta), \theta)]. \tag{16}$$

Since $\partial w(p, \theta)/\partial p > 0$ and $V - V' > 0$, it is necessary that $p(\eta) < p(\eta')$. Then, given that $w(\cdot)$ has strictly increasing differences, the indifference condition (16) implies

$$V - V' > \phi \cdot [w(p(\eta'), \theta') - w(p(\eta), \theta')]$$

for all $\theta' < \theta$, and

$$V - V' < \phi \cdot [w(p(\eta'), \theta) - w(p(\eta), \theta)]$$

for all $\theta' > \theta$. □

A.2 Proof of Lemma 1

Let σ be an equilibrium strategy. To simplify a bit the expression, we further denote the sender's interim image as $p(\theta; \sigma) \equiv \mathbb{E}[\tilde{\theta} | \tilde{\theta} \in \Theta^*(\theta; \sigma)]$. Incentive compatibility implies that, for all sender types $\theta, \theta' \in \Theta$ with $\theta' < \theta$,

$$V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta) \geq V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta) \tag{17}$$

and

$$V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta') \geq V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta'). \tag{18}$$

Summing up (17) and (18), we obtain (with some rearrangement)

$$w(p(\theta;\sigma),\theta) - w(p(\theta';\sigma),\theta) \geq w(p(\theta;\sigma),\theta') - w(p(\theta';\sigma),\theta'). \tag{19}$$

Since $\theta > \theta'$ and $w(\cdot)$ has strictly increasing differences, (19) implies that $p(\theta;\sigma) \geq p(\theta';\sigma)$. Given that the sender always prefers higher images, we also have $w(p(\theta;\sigma),\theta') \geq w(p(\theta';\sigma),\theta')$. Hence, for (18) to hold it is necessary that $V(\theta;\sigma) \leq V(\theta';\sigma)$. □

A.3 Proof of Lemma 2

Take an equilibrium with σ and suppose that it satisfies D1. Suppose that there exists a non-singleton $J \subseteq [0,1]$ such that all types $\theta \in J$ choose the same $\pi \in \Pi^*$ with $\mathbb{E}_\pi[v(s,\omega)] = V > \underline{V}$. Take an information structure $\pi^\varepsilon \in \Pi^*$ that satisfies $\mathbb{E}_{\pi^\varepsilon}[v(s,\omega)] = V - \varepsilon$, which must exist for sufficiently small $\varepsilon > 0$ (see footnote 10).

Let $\mathbb{E}[\tilde{\theta}|\pi^\varepsilon;\sigma]$ be the receiver's posterior expectation about the sender's type upon observing the latter player chooses π^ε. We argue that in equilibrium, $\mathbb{E}[\tilde{\theta}|\pi^\varepsilon;\sigma] \geq \sup J$ must hold. To prove this argument, we distinguish two cases. First, suppose that π^ε is a choice on the equilibrium path under the strategy σ, i.e., there exists $\theta \notin J$ such that $\sigma(\theta) = \pi^\varepsilon$. Then, by Lemma 1, we have $\theta \geq \sup J$. Since the choice of θ was arbitrary and the receiver's on-path beliefs must satisfy Bayes' rule, the claim $\mathbb{E}[\tilde{\theta}|\pi^\varepsilon;\sigma] \geq \sup J$ immediately follows.

Second, suppose that no types will choose π^ε under the strategy σ. In this case, take any $\theta \in J$ with $\theta < \sup J$. By continuity of $w(\cdot)$, for sufficiently small $\varepsilon > 0$ there must exist a posterior expectation $\hat{p} \in [0,1]$ such that if the receiver would hold a belief with this expectation and be obedient to the realization of the signal upon observing π^ε, then the type-θ sender would be indifferent between choosing π and π^ε. Moreover, given that $V - \varepsilon < V$, any sender with $\theta' > \theta$ would strictly prefer π^ε to π whenever type θ is being indifferent between these two pairs, while a sender with $\theta' < \theta$ would hold the exact opposite preference. Hence, due to this single-crossing property (Lemma A1), the D1 criterion requires that the receiver assigns zero weight to types $\theta' \leq \theta$ upon observing that π^ε was chosen by the sender. As a result, we have $\mathbb{E}[\tilde{\theta}|\pi^\varepsilon;\sigma] > \theta$. Since the choice of $\theta < \sup J$ was arbitrary, it again follows that the claim $\mathbb{E}[\tilde{\theta}|\pi^\varepsilon;\sigma] \geq \sup J$ must hold.

Next, since the type distribution $\Gamma(\cdot)$ has full support, we further have

$$\mathbb{E}[\tilde{\theta}|\pi^{\varepsilon};\sigma] \geq \sup J > \mathbb{E}[\tilde{\theta}|\tilde{\theta} \in J].$$

Then, for sufficiently small $\varepsilon > 0$, the expected payoff from π^{ε} will be strictly higher than that from π for all types $\theta \in J$:

$$(V - \varepsilon) + \phi \cdot w\left(\mathbb{E}[\tilde{\theta}|\pi^{\varepsilon};\sigma], \theta\right) > V \cdot f(\theta) + \phi \cdot w\left(\mathbb{E}[\tilde{\theta}|\tilde{\theta} \in J], \theta\right),$$

using that $w(\cdot)$ is strictly increasing in its first argument. This contradicts with σ being an equilibrium strategy. □

A.4 Proof of Theorem 1: The If-Part

To prove the if-statement of Theorem 1, we verify that for any strategy $\sigma = \{\pi_\theta\}_{\theta\in\Theta}$ that satisfies $\pi_\theta \in \Pi^*$ for all $\theta \in [0,1]$ and both conditions (i) and (ii), there is a system of beliefs $H = \{\eta(\pi)\}_{\pi\in\Pi}$ of the receiver so that (σ, H) constitute a D1-equilibrium. The belief system is such that, for any $\pi \in \Pi^*$ publicly chosen by the sender:

- if $\mathbb{E}_\pi[v(s,\omega)] \geq \bar{V} - \phi \int_0^{\min\{\hat{\theta},1\}} \frac{\partial w(x,x)}{\partial p} dx$, the receiver assigns probability one to the unique type $\theta \in [0, \min\{\hat{\theta}, 1\}]$ for which $\mathbb{E}_{\pi_\theta}[v(s,\omega)] = \mathbb{E}_\pi[v(s,\omega)]$;
- if $\bar{V} - \phi \int_0^{\min\{\hat{\theta},1\}} \frac{\partial w(x,x)}{\partial p} dx > \mathbb{E}_\pi[v(s,\omega)] > \underline{V}$, the receiver assigns probability one to type $\min\{\hat{\theta}, 1\}$;
- if $\mathbb{E}_\pi[v(s,\omega)] = \underline{V}$, the receiver updates his belief by restricting the type space to the subset $[\min\{\hat{\theta}, 1\}, 1]$ and invoking Bayes' rule.

Finally, the out-of-equilibrium beliefs $\eta(\pi)$ for $\pi \in \Pi \setminus \Pi^*$ can be completed by following the procedure that we described in footnote 6.

Sequential Rationality. Note that, given H, any π and π' that give rise to the same material payoff will be equally preferred by the sender, as they will also lead to the same posterior beliefs about the sender's type. In addition, when $\hat{\theta} \in (0,1)$ (i.e., the cut-off type is in the interior), any information structure that induces a material payoff V with $\bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x,x)}{\partial p} d\, x < V < \underline{V}$ will be sub-optimal for the sender, as she could always obtain

a higher material payoff without undermining her image. Hence, to verify the sequential rationality of the sender's strategy, it suffices to show that no type $\theta \in [0,1]$ of the sender can strictly benefit from mimicking another type $\theta' \in [0,1]$. Since all types $\theta, \theta' < \hat{\theta}$ are separating according to σ, we have

$$\begin{aligned}
&V(\theta;\sigma) + \phi \cdot w(\theta,\theta) \\
=&V(\theta';\sigma) + \phi \cdot w(\theta',\theta) + [V(\theta;\sigma) - V(\theta';\sigma)] + \phi \cdot [w(\theta,\theta) - w(\theta',\theta)] \\
=&V(\theta';\sigma) + \phi \cdot w(\theta',\theta) - \int_{\theta}^{\theta'} V'(x;\sigma)dx - \phi \cdot \int_{\theta}^{\theta'} \frac{\partial w(x,\theta)}{\partial p} dx \\
=&V(\theta';\sigma) + \phi \cdot w(\theta',\theta) + \int_{\theta}^{\theta'} \left[\frac{\partial w(x,x)}{\partial p} - \frac{\partial w(x,\theta)}{\partial p}\right] dx \\
>&V(\theta';\sigma) + \phi \cdot w(\theta',\theta),
\end{aligned}$$

where the second equality follows condition (i), and the strict inequality follows since $w(\cdot)$ has strictly increasing differences. Thus, no type in $[0,\hat{\theta})$ would want to mimic another type in the same interval. In addition, since all types in $[\hat{\theta},1]$ will get the same material payoff and image payoff according to σ, none of them can benefit from mimicking others in the same interval. Lastly, if $\hat{\theta} \in (0,1)$ (so that both separating and pooling types exist), then by construction the cut-off type $\hat{\theta}$ is indifferent between pooling with higher types (by choosing some $\bar{\pi}$ that yields the minimal material payoff $\underline{V}$) and separating herself (by choosing some π that gives rise to $\mathbb{E}_{\pi}[v(s,\omega)] = \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x,x)}{\partial p} d\,x$). Hence, Lemma A1 implies that the types in the separating interval $[0,\hat{\theta})$ cannot benefit from mimicking those in the pooling interval $[\hat{\theta},1]$, and vice versa.

D1 criterion. Take an off-path communication protocol $\pi' \in \Pi^*$. For any type θ, provided that $D^0(\pi',\theta)$ (i.e., the set of beliefs for which θ weakly prefers to deviate from her choice π_θ to π') is not empty, we define

$$\underline{p}(\pi',\theta) = \inf_{\eta \in D^0(\pi',\theta)} \mathbb{E}_{\eta}[\tilde{\theta}].$$

Note that since $\partial w(p,\theta)/\partial p > 0$, we have $\eta \in D^0(\pi',\theta) \iff \mathbb{E}_{\eta}[\tilde{\theta}] \geq \underline{p}(\pi',\theta)$ and $\eta \in D(\pi',\theta) \iff \mathbb{E}_{\eta}[\tilde{\theta}] > \underline{p}(\pi',\theta)$.

We distinguish two cases. First, suppose that there is $\theta \in [0,1]$ such that $\mathbb{E}_{\pi'}[v(s,\omega)] = V(\theta;\sigma)$, which implies that $\underline{p}(\pi',\theta) = \mathbb{E}[\tilde{\theta}|\pi_\theta;\sigma]$. Consider any type θ' with $\pi_{\theta'} \neq \pi_\theta$. We have

already shown that this type has *strict* incentives *not* to mimic θ. This implies $\underline{p}(\pi', \theta') > \underline{p}(\pi', \theta)$, and therefore $D^0(\pi', \theta') \subsetneq D(\pi', \theta)$. Conversely, for any type θ'' with $\pi_{\theta''} = \pi_\theta$, clearly $\underline{p}(\pi', \theta'') = \underline{p}(\pi', \theta)$, and therefore $D^0(\pi', \theta'') = D^0(\pi', \theta) \supsetneq D(\pi', \theta)$. Thus, the D1 criterion requires that the receiver restricts his out-of-equilibrium belief to those types θ'' with $\pi_{\theta''} = \pi_\theta$. However, our belief system was just chosen this way.

Second, suppose that there is no $\theta \in [0, 1]$ such that that $\mathbb{E}_{\pi'}[v(s, \omega)] = V(\theta; \sigma)$. If $\hat{\theta} = +\infty$ (i.e., the strategy σ is fully separating), then it is necessary that $\mathbb{E}_{\pi'}[v(s, \omega)] < V(1; \sigma)$. In this scenario, on-path incentive compatibility guarantees that $D^0(\pi', \theta) = \emptyset$ for all $\theta \in [0, 1]$, so we can freely choose the out-of-equilibrium beliefs of the receiver for such communication protocols. If $\hat{\theta} \in [0, 1)$ (i.e., the strategy σ is semi-separating), then it is necessary that

$$\underline{V} < \mathbb{E}_{\pi'}[v(s, \omega)] \leq \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx. \tag{20}$$

Hence, for all $\theta < \hat{\theta}$, we have

$$\begin{aligned} V(\theta; \sigma) + \phi \cdot w(\theta, \theta) &> \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx + \phi \cdot w(\hat{\theta}, \theta) \\ &\geq \mathbb{E}_{\pi'}[v(s, \omega)] + \phi \cdot w\left(\underline{p}(\pi', \hat{\theta}), \theta\right), \end{aligned}$$

where the strict inequality follows condition (i), and the weak inequality is jointly implied by (20), the indifference condition of the cut-off type $\hat{\theta}$, and Lemma A1. It is then clear that $\underline{p}(\pi', \theta) > \underline{p}(\pi', \hat{\theta})$ for all $\theta < \hat{\theta}$. Further, since

$$\underline{V} + \phi \cdot w\left(\mathbb{E}[\tilde{\theta} | \tilde{\theta} \geq \hat{\theta}], \hat{\theta}\right) = \mathbb{E}_{\pi'}[v(s, \omega)] + \phi \cdot w\left(\underline{p}(\pi', \hat{\theta}), \hat{\theta}\right),$$

Lemma A1 and (20) jointly imply that

$$\underline{V} + \phi \cdot w\left(\mathbb{E}[\tilde{\theta} | \tilde{\theta} \geq \hat{\theta}], \theta\right) > \mathbb{E}_{\pi'}[v(s, \omega)] + \phi \cdot w\left(\underline{p}(\pi', \hat{\theta}), \theta\right)$$

for all $\theta > \hat{\theta}$. As a result, we also have $\underline{p}(\pi', \theta) > \underline{p}(\pi', \hat{\theta})$ for all $\theta > \hat{\theta}$. In sum, we can conclude that $D^0(\pi', \theta) \subsetneq D(\pi', \hat{\theta})$ for all $\theta \neq \hat{\theta}$, so the D1 criterion requires that the receiver assigns probability one to type $\hat{\theta}$ when he observes π'. However, our belief system was just chosen this way. $\square$

A.5 Proof of Theorem 2

Part (i): Take any information structure $\pi^N \in \Pi^*$ that conveys no information about the state to the receiver, and let $V^N = \mathbb{E}_{\pi^N}[v(s,\omega)]$. By assumption, $U^* = \bar{U}$ is uniquely defined and $\bar{U} > \underline{U}$, so it is necessary that $\bar{V} > V^N$. Take an arbitrary D1 equilibrium strategy $\sigma = \{\pi_\theta\}_{\theta\in[0,1]}$. For every type $\theta \in (0,\hat{\theta})$, recall that her expected payoff $V(\theta;\sigma)$ will be uniquely pinned down by the envelope formula (5). Then, there must exist a non-empty interval $(0,\check{\theta}) \subseteq (0,\hat{\theta})$, such that $V(\theta;\sigma) \geq V^N$ for all $\theta \in (0,\check{\theta})$.

Let $\bar{\pi}$ be an information structure that yields the expected payoff $\bar{V}$ to the sender. For each $\theta \in (0,\check{\theta})$, consider the following information structure $\check{\pi}_\theta$: with probability $\lambda(\theta) = (V(\theta;\sigma)-V^N)/(\bar{V}-V^N) < 1$, the receiver observes a signal s drawn according to $\bar{\pi}$; with the remaining probability $1-\lambda(\theta)$, the signal is generated according to π^N. It is straightforward to check that $\check{\pi}_\theta \in \Pi^*$, $\mathbb{E}_{\check{\pi}_\theta}[v(s,\omega)] = V(\theta;\sigma)$, and

$$\mathbb{E}_{\check{\pi}_\theta}[u^R(s,\omega)] = \lambda(\theta)\cdot\bar{U} + (1-\lambda(\theta))\cdot\underline{U} < \bar{U}. \tag{21}$$

Next, define a strategy $\check{\sigma}$ of the sender as follows: for all $\theta \in (0,\hat{\theta})$, $\check{\sigma}(\theta) = \check{\pi}_\theta$; for all other θ, let $\check{\sigma}(\theta) = \sigma(\theta)$. By Theorem 1, $\check{\sigma}$ is part of a D1 equilibrium. Moreover, since the type distribution Γ is continuous and has full support, (21) implies that the ex-ante expected payoff of the receiver must be strictly lower than U^*, meaning that he is harmed by the presence of the sender's image concerns.

Part (ii): Let $\sigma = \{\pi_\theta\}_{\theta\in[0,1]}$ be the sender's strategy in a Pareto-optimal D1 equilibrium. Suppose by contradiction that the receiver's expected payoff is not decreasing everywhere within $[0,\hat{\theta})$. Then, there must exist $\theta,\theta' \in [0,\hat{\theta})$ such that $\theta < \theta'$ and

$$\mathbb{E}_{\pi_\theta}[u^R(s,\omega)] < \mathbb{E}_{\pi_{\theta'}}[u^R(s,\omega)]. \tag{22}$$

Since U^* is uniquely defined and both θ and θ' are separating, we have $\bar{V} \geq V(\theta;\sigma) > V(\theta';\sigma)$. We consider the following information structure $\check{\pi}_\theta$: with probability $\lambda = (V(\theta;\sigma) - V(\theta';\sigma))/(\bar{V} - V(\theta';\sigma)) > 0$, the information structure generates a signal s according to $\bar{\pi}$; with the remaining probability $1-\lambda$, the signal is generated according to $\pi_{\theta'}$. It is

straightforward to check that $\check{\pi}_\theta \in \Pi^*$, $\mathbb{E}_{\check{\pi}_\theta}[v(s,\omega)] = V(\theta;\sigma)$, and

$$\mathbb{E}_{\check{\pi}_\theta}[u^R(s,\omega)] = \lambda \cdot \bar{U} + (1-\lambda) \cdot \mathbb{E}_{\pi_{\theta'}}[u^R(s,\omega)] > \mathbb{E}_{\pi_\theta}[u^R(s,\omega)]. \tag{23}$$

Therefore, it is possible to construct a D1 equilibrium strategy $\check{\sigma}$ that always gives the receiver a weakly higher payoff than σ, and this payoff difference will even be strict when the sender's type is θ. Hence, the strategy σ cannot be Pareto-optimal if the associated payoff for the receiver is not decreasing within the separating interval $[0, \hat{\theta})$.

Part (iii): Since $U^* = \bar{U}$, quasi-convexity is equivalent to the receiver's payoff being either monotonically decreasing or U-shaped with respect to the sender's type. Let $\sigma = \{\pi_\theta\}_{\theta\in[0,1]}$ be the sender's strategy in a Pareto-worst D1 equilibrium. Take any information $\pi^N \in \Pi^*$ that conveys no information about the state to the receiver, and let V^N be the associated payoff of the sender. We distinguish two cases. First, suppose that

$$\bar{V} - \phi \cdot \int_0^{\min\{\hat{\theta},1\}} \frac{\partial w(x,x)}{dp} dx \geq V^N. \tag{24}$$

We claim that in this case, the receiver's expected payoff must be decreasing everywhere on $[0, \hat{\theta})$. To prove this, suppose by contradiction that there exist $\theta, \theta' \in [0, \hat{\theta})$ such that $\theta < \theta'$ and (22) holds. Then, consider the following information structure $\check{\pi}_{\theta'}$: with probability $\lambda' = (V(\theta';\sigma) - V^N)/(V(\theta;\sigma) - V^N) < 1$, the information structure generates a signal s according to π_θ; with the remaining probability $1-\lambda'$, the signal is generated according to π^N. It is straightforward to check that $\check{\pi}_{\theta'} \in \Pi^*$, $\mathbb{E}_{\check{\pi}_{\theta'}}[v(s,\omega)] = V(\theta';\sigma)$, and

$$\mathbb{E}_{\check{\pi}_{\theta'}}[u^R(s,\omega)] = \lambda' \cdot \mathbb{E}_{\pi_\theta}[u^R(s,\omega)] + (1-\lambda') \cdot \underline{U} < \mathbb{E}_{\pi_{\theta'}}[u^R(s,\omega)]. \tag{25}$$

Therefore, it is possible to construct a D1 equilibrium strategy $\check{\sigma}$ that always gives the receiver a weakly higher payoff than σ, and this payoff difference will even be strict when the sender's type is θ'. Hence, if (24) holds, the receiver's expected payoff $\mathbb{E}_{\check{\pi}_\theta}[u^R(s,\omega)]$ must be monotonically decreasing in θ within the interval $[0, \hat{\theta})$.

Second, suppose that (24) does not hold. Then, there must exist $\theta^N \in [0, \hat{\theta})$, such that $V(\theta^N;\sigma) = V^N$. Using similar construction of "grand" information structures involving π^N, it can be shown that the receiver's expected payoff must be first decreasing in θ on $[0, \theta^N]$, and then increasing on $[\theta^N, \hat{\theta})$. □

A.6 Results and Proofs Related to the Examples

A.6.1 The Utility-Frontier with Almost-Perfectly-Aligned Preferences

Consider the game with almost perfectly aligned preferences, which we introduced in Example 1. Let the prior distribution μ_0 be such that $\Pr(\omega = 1) = \Pr(\omega = 0) = 0.4$ and $\Pr(\omega = -1) = 0.2$. It is clear that $\bar{V} = \bar{U} = 1$ and $\underline{U} = 0.4$. To solve $\underline{V}$, first note that the sender's expected material payoff depends mainly on two things: (i) the total probability that the receiver will take the right action, $\Pr(a = \omega)$; (ii) the total probability that the receiver will wrongly take the action $a = -1$, $\Pr(a = -1|\omega \neq -1)$. Regardless of which information structure $\pi \in \Pi^*$ is used by the sender, it is necessary that $\Pr(a = \omega) \geq 0.4$, because the receiver cannot do strictly worse than sticking to his prior-optimal action. At the same time, 0.4 is also an upper bound for $\Pr(a = -1|\omega \neq -1)$: If $\Pr(a = -1|\omega \neq -1) > 0.4$, the receiver would necessarily hold a posterior with $\Pr(\omega = -1|s = -1) < 1/3$, which means that it cannot be rational for him to take recommended action -1.

Now consider an information structure $\underline{\pi} \in \Pi^*$ which recommends the action $a = -1$ with probability 1 in state $\omega = -1$, and it recommends $a = 1$ or $a = -1$ with equal probabilities in the other two states. It can be checked that $\underline{\pi}$ achieves the above two bounds on the receiver's decision-making probabilities simultaneously. Since the sender is worse off when the receiver less often takes the right action and more often chooses the action $a = -1$ in the wrong states, $\underline{\pi}$ must give the lowest possible payoff to the sender among all information structures, which is $\underline{V} = 0$.

We, therefore, know that the set of implementable payoff profiles, formally defined as $\mathcal{W} = \{(V, U) : \exists\, \pi \in \Pi^* \text{ such that } V = \mathbb{E}_\pi[v(s, w)] \text{ and } U = \mathbb{E}_\pi[u^R(s, w)]\}$, must lie in the rectangle $[\underline{V}, \bar{V}] \times [\underline{U}, \bar{U}] = [0, 1] \times [0.4, 1]$. In addition, $\mathcal{W}$ is closed and convex (Zhong, 2018). Hence, to characterize $\mathcal{W}$, it suffices to answer the following question: for a given level of the receiver's payoff $U \in [\underline{U}, \bar{U}]$, what are the maximal and the minimal material payoffs that the sender can achieve by using some information structure $\pi \in \Pi^*$, respectively? Note that the receiver's expected payoff equals exactly the ex-ante probability that he takes the right action. Hence, the question reduces to identifying the set of $\Pr(a = -1|\omega \neq -1)$ that the sender may induce without violating the requirement $\Pr(a = \omega) = U$.

For $U \in [0.4, 0.6]$, the previous upper bound on $\Pr(a = -1|\omega \neq -1)$ can still be achieved. This is made possible by the information structure $\underline{\pi}^U \in \Pi^*$ characterized by the following

conditional probabilities (of recommending different actions in different states): $\underline{\pi}^U(-1|-1) = 1$, $\underline{\pi}^U(1|1) = \underline{\pi}^U(-1|1) = \underline{\pi}^U(-1|0) = 0.5$, $\underline{\pi}^U(0|0) = (U - 0.4)/0.4$, and $\underline{\pi}^U(1|0) = (0.8 - U)/0.4$. The resulting payoff to the sender, $U - 0.4$, is the minimal one across all information structures that induce the receiver to choose $a = -1$ with probability U. Consequently, for all $U \in [0.4, 0.6]$, $(U, U - 0.4)$ is on the boundary of $\mathcal{W}$, which corresponds to a point on the red curve (below the kink) in Panel (b) of Figure 2. As for $U \in (0.6, 1]$, the highest probability that the receiver will wrongly choose $a = -1$ becomes $1 - U$. This (revised) upper bound can be achieved by an information structure $\underline{\pi}^U \in \Pi^*$ with $\underline{\pi}^U(-1| - 1) = 1$, $\underline{\pi}^U(1|1) = \underline{\pi}^U(0|0) = (U - 0.2)/0.8$, and $\underline{\pi}^U(-1|1) = \underline{\pi}^U(-1|0) = (1 - U)/0.8$. Thus, for each $U \in (0.6, 1]$, $(U, 2U - 1)$ is on the boundary of $\mathcal{W}$, and it corresponds to a point on the red curve (this time above the kink) in the figure.

Finally, for all $U \in [0.4, 1]$, the maximal payoff of the sender is necessarily achieved when she *never* chooses $a = -1$ in states $\omega \in \{0, 1\}$. Hence, every payoff profile (U, U) with $U \in [0.4, 1]$ is in the boundary of $\mathcal{W}$. In addition, since both $(0.4, 0)$ and $(0.4, 0.4)$ are implementable and $\underline{U} = 0.4$, any payoff profile $(0.4, V)$ with $V \in (0, 0.4)$ is also an boundary point of $\mathcal{W}$. Taken together, we obtain the blue curve depicted in the figure.

A.6.2 Transforming the Quadratic-Loss Games

Consider the quadratic-loss game that we discussed in Examples 2 and 4. Given the sender's choice of information structure π, the receiver has a unique best response for every signal realization $s \in \text{supp}(\pi)$: $\hat{a}(s) = \mathbb{E}[\omega|s]$. As a result, the expected material loss of a type-θ sender is

$$\begin{aligned}
&\mathbb{E}_\pi\left[(\hat{a}(s) - a^*(\omega,\theta))^2|s\right] \\
=&\,\mathbb{E}_\pi\left[(\mathbb{E}[\omega|s])^2\,|s\right] + \mathbb{E}_\pi\left[(a^*(\omega,\theta))^2\,|s\right] - 2\mathbb{E}_\pi\left[\mathbb{E}[\omega|s]\cdot(f(\theta)\cdot\omega + g(\theta))|s\right] \\
=&\,\mathbb{E}_\pi\left[(\mathbb{E}[\omega|s])^2\right] + \mathbb{E}_{\mu_0}\left[(a^*(\omega,\theta))^2\right] - 2f(\theta)\cdot\mathbb{E}_\pi\left[(\mathbb{E}[\omega|s])^2\right] - 2g(\theta)\cdot\mathbb{E}_{\mu_0}[\omega] \\
=&\,(1 - 2f(\theta))\cdot\mathbb{E}_\pi[\mathbb{E}[\omega|s]^2] + K(\theta),
\end{aligned}$$

where the second equality follows the law of iterated expectation, and we use $K(\theta) \equiv \mathbb{E}_{\mu_0}\left[(a^*(\omega,\theta))^2\right] - 2g(\theta)\cdot\mathbb{E}_{\mu_0}[\omega]$ to collect all the (θ-specific) constant terms. In addition, we have $\mathbb{E}_\pi\left[(\hat{a}(s) - \omega)^2|s\right] = -\mathbb{E}_\pi[\mathbb{E}[\omega|s]^2] + \mathbb{E}_{\mu_0}[\omega^2]$.

Now, suppose that $f(\theta) > 0.5\ \forall\theta \in [0, 1]$ and compare the following two utility functions

of the sender: $u^S(a,\omega,\theta,\eta) = -(a-a^*(\omega,\theta))^2+\phi\cdot w(p(\eta),\theta)$, and $\hat{u}^S(a,\omega,\theta,\eta) = -(a-\omega)^2+\phi\cdot\hat{w}(p(\eta),\theta)$, where $\hat{w}(p(\eta),\theta) = w(p(\eta),\theta)/(2f(\theta)-1)$. We claim that, taking the receiver's best response $\hat{a}(\cdot)$ as given, these two utility functions represent the same preference over the pairs (π,η) for all types $\theta\in[0,1]$. This is because, for all $\theta\in[0,1]$ and all (π,η) and (π',η'), we have

$$\begin{aligned}
&\mathbb{E}_\pi[u^S(\hat{a}(s),\omega,\theta,\eta)|s] \geq \mathbb{E}_{\pi'}[u^S(\hat{a}(s),\omega,\theta,\eta')|s]\\
\iff &(2f(\theta)-1)\cdot\left[\mathbb{E}_\pi[\mathbb{E}[\omega|s]^2]-[\mathbb{E}_{\pi'}[\mathbb{E}[\omega|s]^2]\right]+\phi\cdot[w(p(\eta),\theta)-w(p(\eta'),\theta)]\geq 0\\
\iff &\mathbb{E}_\pi[\mathbb{E}[\omega|s]^2]-[\mathbb{E}_{\pi'}[\mathbb{E}[\omega|s]^2]+\phi\cdot[\hat{w}(p(\eta),\theta)-\hat{w}(p(\eta'),\theta)]\geq 0\\
\iff &\mathbb{E}_\pi[\hat{u}^S(\hat{a}(s),\omega,\theta,\eta)|s] \geq \mathbb{E}_{\pi'}[\hat{u}^S(\hat{a}(s),\omega,\theta,\eta')|s].
\end{aligned}$$

Hence, under the current parametric assumption, the quadratic-loss game in Example 2 has the same equilibrium set as a game where the receiver's utility function remains unchanged, but the sender's utility function is instead given by $\hat{u}^S(\cdot)$.

Similarly, if $f(\theta)<0.5\ \forall\theta\in[0,1]$, then, as described in Example 4, we effectively have a quadratic loss game where the sender's material payoff function is given by $v(a,\omega) = -(a-\omega)^2$, while her image payoff function is given by $\hat{w}(p(\eta),\theta) = w(p(\eta),\theta)/(1-2f(\theta))$.

A.6.3 Receiver-Optimality with State-Independent Sender Preferences

Consider our first example of state-independent sender preferences (Example 5), where both the state and the action spaces are binary. Recall the information structure $\bar{\pi}^q$, which is defined according to (14) for each $q\in[0,2\mu_0]$. We argue that, among all information structures that induce the receiver to choose the high action with probability q, $\bar{\pi}^q$ is the one that gives the highest payoff to the receiver.

To prove our claim, note that, under the information structure $\bar{\pi}^q$, the receiver's payoff is given by

$$\mathbb{E}_{\bar{\pi}^q}[u^R(s,\omega)] = \begin{cases} 1-\mu_0+q & \text{if } q\in[0,\mu_0],\\ 1+\mu_0-q & \text{if } q\in(\mu_0,2\mu_0].\end{cases}$$

Now take any other information $\pi\in\Pi^*$ such that may induce the receiver to choose the high action with probability q, and let $\pi(a|\omega)$ be the conditional probability that it recommends

action a when the state is ω. Then, it is necessary that

$$\mu_0 \cdot \pi(1|1) + (1 - \mu_0) \cdot \pi(1|0) = q. \tag{26}$$

Therefore, the receiver's expected utility under π is given by

$$\begin{aligned}\mathbb{E}_\pi[u^R(s,\omega)] &= \mu_0 \cdot \pi(1|1) + (1-\mu_0) \cdot \pi(0|0) \\ &= q - (1-\mu_0) \cdot \pi(1|0) + (1-\mu_0) \cdot (1 - \pi(1|0)) \\ &= 1 - \mu_0 + q - 2(1-\mu_0) \cdot \pi(1|0).\end{aligned}$$

Since $\pi(1|0) \geq 0$, we have $\mathbb{E}_\pi[u^R(s,\omega)] \leq 1 - \mu_0 + q$, so $\bar{\pi}^q$ is clearly receiver-optimal when $q \in [0, \mu_0]$. At the same time, note that, using (26), the receiver's expected utility can also be written as

$$\mathbb{E}_\pi[u^R(s,\omega)] = 1 + (2\pi(1|1) - 1) \cdot \mu_0 - q.$$

Then, since $\pi(1|1) \leq 1$, we also have $\mathbb{E}_\pi[u^R(s,\omega)] \leq 1 + \mu_0 - q$. Hence, $\bar{\pi}^q$ is also receiver-optimal when $q \in (\mu_0, 2\mu_0]$.

A.7 Proof of Theorem 4

Since we can always concentrate the mass of the type distribution to sufficiently small types, the first statement of the theorem is implied by the second.[27] To prove the second statement, take any information structure $\pi^N \in \Pi^*$ ($\pi^F \in \Pi^*$) that conveys no (full) information about the state to the receiver. Let V^N and V^F be the expected material payoffs of the sender under π^N and π^F, respectively. Since $U^* \in (\underline{U}, \bar{U})$ is uniquely defined, it is necessary that $\bar{V} > \max\{V^N, V^F\}$.

If ϕ is sufficiently small, all D1 equilibria will be fully separating, and the interim payoff of the highest type will satisfy $V(1;\sigma) > \max\{V^N, V^F\}$ irrespective of the which D1 equilibrium strategy σ is selected. In particular, there exists a D1 equilibrium in which each type θ uses a "grand" information structure that mixes appropriately between the sender-optimal information structure $\bar{\pi}$ absenting image concerns and the information structure π^N. Clearly,

[27] Note that as the type distribution converges to a degenerate distribution at $\theta = 0$, $I(0)$ will converge to zero. Hence, sufficiently small types will necessarily be separating in equilibrium when the type distribution puts sufficiently large mass on them.

the receiver is strictly worse off in this equilibrium relative to the equilibrium without image concerns. Similarly, there also exists a D1 equilibrium in which the sender's strategy is always a combination of $\bar{\pi}$ and π^F. It is straightforward to verify that the receiver must be strictly better off in this equilibrium relative to the equilibrium without image concerns. □

A.8 Populist Sentiments and Policy Stagnation

A.8.1 Reduced Form Description of the Equilibrium of the Dynamic Game

We explain that the dynamic game as in 4.3 has a reduced form description in terms of the model in Section 2. We begin by arguing that in any equilibrium, the politician will always disclose the result of the study, regardless of whether it is positive about the reform or not.

To see this, let Θ_π be the set of types that choose the study π in an equilibrium. Note that disclosing any result s with $\pi(s|1)/\pi(s|0) \geq \ell_0$ will for sure lead to the adoption of the reform. This implies that, for a type $\theta \in \Theta_\pi$ to prefer keeping this result private, non-disclosure must lead to a higher reputation than disclosing s. By virtue of the single-crossing property, all types $\theta' \in \Theta_\pi$ with $\pi' < \pi$ would strictly prefer non-disclosure to disclosure. But then, the highest type among the non-disclosing ones would have a strict incentive to deviate, and the classic unraveling argument applies. Hence, in any equilibrium, all results s with $\pi(s|1)/\pi(s|0) \geq \ell_0$ will necessarily be disclosed. An analogous argument establishes that any result s with $\pi(s|1)/\pi(s|0) < \ell_0$ will also be disclosed.

Given that the politician would always disclose what she learns from the study, (on the equilibrium path) the citizen's posterior belief about the politician's type would only depend on the chosen study. Consequently, a type-θ politician obtains the following payoff from choosing a study π:

$$\Pr(\pi(s|1)/\pi(s|0) \geq \ell_0) \cdot w_1(\theta) + \phi \cdot (1 - G(\underline{u} - p)) \cdot w_2(\theta), \tag{27}$$

where $p = \mathbb{E}[\theta|\pi]$. Without loss of generality, suppose that each politician type chooses a study that only gives rise to a binary result – either positive ($s = 1$) or negative ($s = 0$). Naturally, disclosing the positive result is equivalent to making a recommendation to pass the reform, while disclosing the negative result is the same as recommending to maintain the

status quo. Thus, for each type of politician, maximizing (27) is equivalent to

$$\max_{\pi} \Pr(s = 1 \,|\, \pi) + \phi \cdot (1 - G(\underline{u} - p)) \cdot \frac{w_2(\theta)}{w_1(\theta)},$$

subject to the constraints that $\pi(1|\omega), \pi(0|\omega) \in [0, 1]$ and $\pi(0|\omega) + \pi(1|\omega) = 1 \; \forall \, \omega \in \{0, 1\}$, and $\pi(1|1)/\pi(1/0) \geq \ell_0$.

Finally, we see that the equilibrium problem maps into our setting by specializing Example 4 of Section 3.2 with $u^R(a, \omega) = \mathbb{1}_{a=\omega}$, $v(a, \omega) = \mathbb{1}_{a=1}$, and $w(p, \theta) = (1 - G(\underline{u} - p))\, w_2(\theta)/w_1(\theta)$.

A.8.2 Micro-Founding the Assumptions in Section 4.3

In what follows, we provide a setting of electoral competition that endogenizes the key assumption in Section 4.3, namely, that the ratio $w_2(\cdot)/w_1(\cdot)$ is strictly increasing. We suppose that, in the second stage, the politician described in Section 4.3 – who we now call candidate A – competes with another candidate B for the election. Each candidate $j = A, B$ has a private type $\theta_j \in [0, 1]$, which is i.i.d. with the distribution function Γ. We use θ_0 to denote the mean of the type distribution.

The candidate who wins the election will get to choose a policy $y \in \mathbb{R}$. The citizen's policy preference is $-|y - y^*|$. For each candidate $j = A, B$, with probability θ_j, she will have the same policy preference as the citizen. With the remaining probability $1 - \theta_j$, the candidate's preference will be $-|y - (y^* + 1)|$. Thus, the higher θ_j (i.e., the less corrupt the candidate), the more likely that the preferences of the candidate and the citizen are aligned.

Recall that the citizen may update his belief about candidate A's type upon observing the latter's choice of commissioned study (whereas the belief about candidate B is given by the prior). Therefore, it is straightforward to show that the citizen would prefer to vote for candidate A if and only if $\varepsilon \geq \theta_0 - p$, where $p = \mathbb{E}[\theta_A|\pi]$. The winning probability of candidate A is then given by $1 - G(\theta_0 - p)$.

Overall, for candidate A, her expected payoff from the electoral competition is

$$G(\theta_0 - p) \cdot [-\theta_A(1 - \theta_0) - (1 - \theta_A)\theta_0] = -G(\theta_0 - p) \cdot w_2(\theta_A),$$

where $w_2(\theta_A) = \theta_A + \theta_0 - 2\theta_0\theta_A$. Given that $w_1(\cdot)$ is strictly decreasing, it is easy to check that $w_2(\cdot)/w_1(\cdot)$ is strictly increasing whenever $\theta_0 < 0.5$ is additionally satisfied. This condition captures that the public's prior about the corruptness of the candidates is relatively high.

References

Acemoglu, D., Egorov, G. and Sonin, K. (2013). A political theory of populism. *The Quarterly Journal of Economics*, **128** (2), 771–805.

Ashworth, S. (2012). Electoral accountability: Recent theoretical and empirical work. *Annual Review of Political Science*, **15** (1), 183–201.

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, **55** (3), 647–661.

Baumeister, R. F. (1998). *The self.* The Handbook of Social Psychology.

Ben-Porath, E., Dekel, E. and Lipman, B. L. (2014). Optimal allocation with costly verification. *American Economic Review*, **104** (12), 3779–3813.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, **117** (3), 871–915.

— and Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, **112** (4), 848–886.

— and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, **96** (5), 1652–1678.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, **102** (5), 841–877.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In I. Brocas and J. D. Carrillo (eds.), *The Psychology of Economic Decisions*, vol. 1, Oxford University Press, pp. 105–123.

Che, Y.-K., Kim, J. and Mierendorff, K. (2013). Generalized reduced-form auctions: A network-flow approach. *Econometrica*, **81** (6), 2487–2520.

Chen, S. and Heese, C. (2021). Fishing for good news: Motivated information acquisition, cRC TR 224 Discussion Paper No. 223.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, **102** (2), 179–221.

— and Sobel, J. (1990). Strategic stability and uniqueness in signaling games. *Journal of Economic Theory*, **50** (2), 381–413.

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

Crocker, J. and Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin*, **130** (3), 392–414.

Dillon, K. (2017). New managers should focus on helping their teams, not pleasing their bosses. *Harvard Business Review*.

Duggan, J. and Martinelli, C. (2017). The political economy of dynamic elections: Accountability, commitment, and responsiveness. *Journal of Economic Literature*, **55** (3), 916–84.

Ellingsen, T. and Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, **98** (3), 990–1008.

Ely, J., Fudenberg, D. and Levine, D. K. (2008). When is reputation bad? *Games and Economic Behavior*, **63** (2), 498–526.

Ely, J. C. and Välimäki, J. (2003). Bad reputation. *The Quarterly Journal of Economics*, **118** (3), 785–814.

Fehr, E. and Rangel, A. (2011). Neuroeconomic foundations of economic choice – Recent advances. *Journal of Economic Perspectives*, **25** (4), 3–30.

Fernandez, R. and Rodrik, D. (1991). Resistance to reform: Status quo bias in the presence of individual-specific uncertainty. *American Economic Review*, **81** (5), 1146–1155.

Fudenberg, D., Newey, W., Strack, P. and Strzalecki, T. (2020). Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences*, **117** (52), 33141–33148.

— and Tirole, J. (1991). *Game Theory*. MIT Press.

Galperti, S. (2019). Persuasion: The art of changing worldviews. *American Economic Review*, **109** (3), 996–1031.

Geanakoplos, J., Pearce, D. and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, **1** (1), 60–79.

Gentzkow, M. and Kamenica, E. (2017). Disclosure of endogenous information. *Economic Theory Bulletin*, **5** (1), 47–56.

Golman, R., Hagmann, D. and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, **55** (1), 96–135.

Green, J. R. and Stokey, N. L. (2007). A two-person game of information transmission. *Journal of Economic Theory*, **135** (1), 90–104.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, **24** (3), 461–483.

Grossman, Z. and Van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, **15** (1), 173–217.

Guo, Y. and Shmaya, E. (2019). The interval structure of optimal disclosure. *Econometrica*, **87** (2), 653–675.

Hedlund, J. (2017). Bayesian persuasion by a privately informed sender. *Journal of Economic Theory*, **167**, 229–268.

Henry, E. and Ottaviani, M. (2019). Research and the approval process: The organization of persuasion. *American Economic Review*, **109** (3), 911–55.

Hermalin, B. E. (2001). Economics and corporate culture. In S. Cartwright, P. C. Earley and C. L. Cooper (eds.), *The International Handbook of Organizational Culture and Climate*, New York: John Wiley & Sons, pp. 217–261.

Ivanov, M. (2021). Optimal monotone signals in Bayesian persuasion mechanisms. *Economic Theory*, **72** (3), 955–1000.

Jehiel, P. (2015). On transparency in organizations. *The Review of Economic Studies*, **82** (2), 736–761.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, **101** (6), 2590–2615.

—, Kim, K. and Zapechelnyuk, A. (2021). Bayesian persuasion and information design: Perspectives and open issues. *Economic Theory*, **72** (3), 701–704.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, **76** (4), 1359–1395.

Koessler, F. and Skreta, V. (2021). Information design by an informed designer, CEPR Discussion Paper No. DP15709.

Kohlberg, E. and Mertens, J.-F. (1986). On the strategic stability of equilibria. *Econometrica*, **54** (5), 1003–1037.

Kolotilin, A., Corrao, R. and Wolitzky, A. (2022a). Persuasion as matching, mimeo.

—, Mylovanov, T. and Zapechelnyuk, A. (2022b). Censorship as optimal persuasion. *Theoretical Economics*, **17** (2), 561–585.

—, —, — and Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, **85** (6), 1949–1964.

— and Wolitzky, A. (2020). Assortative information disclosure, UNSW Economics Working Paper 2020-08.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, **4** (4), 673–707.

Krajbich, I., Oud, B. and Fehr, E. (2014). Benefits of neuroeconomic modeling: New policy interventions and predictors of preference. *American Economic Review: Papers & Proceedings*, **104** (5), 501–506.

Kreps, D. M. (1990). Corporate culture and economic theory. In J. E. Alt and K. A. Shepsle (eds.), *Perspectives on positive political economy*, Cambridge: Cambridge University Press, pp. 90–143.

Lipnowski, E. and Mathevet, L. (2018). Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, **10** (4), 67–93.

Mailath, G. J. (1987). Incentive compatibility in signaling games with a continuum of types. *Econometrica*, **55** (6), 1349–1365.

—, Samuelson, L. *et al.* (2006). *Repeated Games and Reputations: Long-Run Relationships*. Oxford University Press.

Mallapaty, S. (2022). What Xi Jinping's third term means for science. *Nature*, **611** (7934), 20–21.

Mas-Colell, A., Whinston, M. D. and Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

Melumad, N. D. and Shibano, T. (1991). Communication in settings with no transfers. *The RAND Journal of Economics*, **22** (2), 173–198.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pp. 380–391.

Morris, S. (2001). Political correctness. *Journal of Political Economy*, **109** (2), 231–265.

— and Strack, P. (2019). The wald problem and the relation of sequential sampling and ex-ante information costs, available at SSRN: https://ssrn.com/abstract=2991567.

Mudde, C. (2004). The populist zeitgeist. *Government and Opposition*, **39** (4), 541–563.

Myerson, R. B. (2009). Learning from Schelling's *Strategy of Conflict*. *Journal of Economic Literature*, **47** (4), 1109–25.

Nikandrova, A. and Pancs, R. (2017). Conjugate information disclosure in an auction with learning. *Journal of Economic Theory*, **171**, 174–212.

Perez-Richet, E. (2014). Interim Bayesian persuasion: First steps. *American Economic Review: Papers & Proceedings*, **104** (5), 469–74.

Persson, T. and Tabellini, G. (2002). *Political Economics: Explaining Economic Policy*. MIT Press.

Prillaman, M. (2022). Billions more for US science: How the landmark spending plan will boost research. *Nature*, **608**, 249.

Ramey, G. (1996). D1 signaling equilibria with multiple signals and a continuum of types. *Journal of Economic Theory*, **69** (2), 508–531.

Ratcliff, R., Smith, P. L., Brown, S. D. and McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, **20** (4), 260–281.

Schelling, T. C. (1980). *The Strategy of Conflict*. Harvard University Press.

Schlenker, B. R. (2012). Self-presentation. In M. R. Leary and J. P. Tangney (eds.), *Handbook of Self and Identity*, *25*, The Guilford Press, pp. 492–518.

Schwardmann, P. and Van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, **3** (10), 1055–1061.

Schweizer, N. and Szech, N. (2018). Optimal revelation of life-changing information. *Management Science*, **64** (11), 5250–5262.

Smolin, A. and Yamashita, T. (2022). Information design in concave games, available at arXiv: https://arxiv.org/abs/2202.10883.

Tamura, W. (2018). Bayesian persuasion with quadratic preferences, working paper.

Terstiege, S. and Wasser, C. (2022). Competitive information disclosure to an auctioneer. *American Economic Journal: Microeconomics*, **14** (3), 622–64.

Zhong, W. (2018). Information design possibility set, working paper.