

# Persuasion versus Presentation

Carl Heese and Shuo Liu\*

November 2023

## Abstract

In many economic situations, a sender communicates strategically with a receiver not only to influence the receiver's decision-making but also to shape the perception of unobserved characteristics of herself (e.g. morality, loyalty, or capability). To study such situations, we propose a model of Bayesian persuasion in which a sender communicates about a pay-off relevant state to influence decisions but also has a private type and cares about how she presents herself. Optimal communication balances persuasion and self-presentation by signaling about the sender's type. Such signaling through the design of communication has non-standard implications for the receiver: Whether the receiver fares better or worse compared to the pure persuasion setting may depend on the selected equilibrium and effects can be non-monotone with respect to the sender's type. We illustrate our findings within various classic payoff environments, for instance with quadratic losses or state-independent sender preferences. Finally, we use the model to shed new light on a series of applications.

*Keywords:* image concerns, persuasion, self-presentation, signaling

*JEL Classification:* C72, D72, D82, M50

---

\*Heese: Business School, Hong Kong University. Liu: Guanghua School of Management, Peking University. Emails: heese@hku.hk and shuo.liu@gsm.pku.edu.cn. This paper was previously circulated under the title "Image-Building Persuasion". We are grateful for the very helpful comments to James Best, Si Chen, Chen Cheng, Navin Kartik, Daniel Krähmer, Alexey Kushnir, Stephan Laueremann, Ming Li, Xiao Lin, Qianjun Lyu and seminar participants at Arizona State University, Carnegie Mellon University, Columbia University, Johns Hopkins University, Nanjing University, Peking University, University of International Business and Economics, University of Pennsylvania, University of Vienna, the SAET Annual Conference 2022 and the Econometric Society European Summer Meeting 2023. Shuo Liu acknowledges financial support from the National Natural Science Foundation of China (Grant No. 72322006).

# 1 Introduction

Many economic situations involve a sender with unobserved characteristics supplying information about a payoff-relevant state to a receiver. Examples are ubiquitous: Voters receive information about the potential impacts of a proposed reform from a politician, whose private gains from the reform’s passage are uncertain. Workers learn about the difficulty of a task from a manager, without fully knowing the extent to which their preferences align. A researcher trying to solve a problem collects notes about it from her past self, and wonders about her own problem-solving abilities, and so on. In such situations, the receiver may develop beliefs about both the state and the characteristics of the sender, that is, her *type*, based on the varying communication strategies employed by the latter.

In virtually all applications, the sender (she) is motivated to influence the induced belief about the state, typically because that is critical for convincing the receiver (he) to take a favorable action. This “persuasion motive” and its impacts on information revelation have been extensively studied in the strategic communication literature (e.g. Crawford and Sobel, 1982; Grossman, 1981; Kamenica and Gentzkow, 2011; Milgrom, 1981). However, the sender may also have an intrinsic interest in the induced belief about her own type. For instance, concerns about how one’s private traits are perceived come into play when politicians aspire to be seen as non-corrupt in pursuit of electoral support, when managers seek to appear loyal to company values for career advancements, or when individuals fancy themselves as instinctive problem solvers. Such image concerns or “presentational motives” are a prominent topic in the social psychology of communication since Goffman (1959).<sup>1</sup> They have received scant attention in the economic theory literature on communication despite their prevalence and potential tension with persuasion incentives.

This paper proposes a model of strategic communication that incorporates both persuasion and presentation motives. In line with the Bayesian persuasion paradigm pioneered by Rayo and Segal (2010) and Kamenica and Gentzkow (2011), we consider a sender who is able to commit to how she will release information about the state before communicating with a receiver. Crucially, our model also endows the sender with a private preference type that differs from the state. This novel feature gives rise to two potentially competing objectives for the sender when designing the information structure: persuading the receiver toward

---

<sup>1</sup>See, for example, Baumeister (1982), who defines self-presentation as “the use of behavior to communicate some information about oneself to others. . . self-presentational motives are to please the audience and to construct (create, maintain, and modify) one’s public self congruent to one’s ideal.”

favorable actions while presenting herself as a “good” type. We find that the interplay between persuasion and presentation motives can substantially reshape information revelation compared to pure persuasion settings. Optimal communication balances both motives, by typically revealing some information about the state but also signaling about the sender’s type. Such signaling through the design of the communication turns out to have non-standard implications for the receiver. For example, the welfare of the receiver can be non-monotone with respect to the sender’s type or the relative strength of the two motives.

As a glimpse into the new insights enabled by our model, consider the motivating binary-state example from Kamenica and Gentzkow (2011), in which a sender attempts to convince a receiver to choose her preferred action out of two possible ones. The receiver’s prior leans towards the other action. That is, absent persuasive new information, he would choose the other action. It is known that the sender can achieve the highest likelihood of her preferred action through partial disclosure: always revealing the truth in the state where the receiver ranks the preferred action highest but only sometimes so when this is not the case.

Suppose now that—unlike in the standard setting—the sender has a preference type capturing the sender’s private gains from persuading the receiver of her preferred action. Such gains may relate to undesirable traits, such as the corruptness of a politician (or the bias of a prosecutor, as in the original example by Kamenica and Gentzkow). Then, the sender has incentives to signal that she is a “high” type – one with less private gains – by disclosing information in a way that is costly for lower types to imitate. One intuitive approach to accomplish this is to more often reveal the state when the receiver prefers the other action. This is less attractive for lower types since they have more private gains from influencing the receiver towards their preferred action. When the sender engages in this type of signaling behavior, it facilitates information revelation and benefits the receiver. However, this approach of separation can work for *all* sender types only if the gain from reputation is not too large. Otherwise, some intermediate type may already be incentivized to furnish the receiver with full information. The highest types can then distinguish themselves only by *withholding* relevant information and revealing the state less often when the sender ranks the preferred action highest. In other words, an overly strong presentation motive backfires – it leads to less informative communication and welfare losses for the receiver. Ultimately, strong enough image concerns will drive all types to refrain from providing any information to the receiver, as it turns out.

Our general model allows for the same flexibility in terms of the action and the state space, as the original work by Kamenica and Gentzkow (2011). We formalize the two motives by considering a receiver who does not only update about the state from his received signal, in order to improve his choice of action but who also updates about the sender’s type based on how she generates that signal.<sup>2</sup> Further, the sender’s utility function comprises two components: a *material payoff* determined by the state and the receiver’s action, and an *image payoff* based on the sender’s true type and what type the receiver expects her to be. For the sender’s image payoff, we assume that it increases with the sender’s reputation and satisfies the canonical Spence–Mirrlees condition. In our setting, this condition ensures that higher sender types place a greater premium on enhancing their reputation compared to lower types.

As is typical in signaling games, the lack of discipline imposed by Bayes’ rule on off-equilibrium beliefs can lead to a large multiplicity of equilibria. Another complicating issue is that, even in pure persuasion settings absent image considerations, pinpointing the optimal information structures for the sender often proves infeasible. To tackle these challenges, we advance in three steps. First, to narrow down the set of possible equilibria, we establish a revelation principle showing that for any equilibrium there is an equilibrium that is pay-off equivalent in terms of both the material and (the endogenously derived) image payoffs and in which the sender’s signals are action recommendations that the receiver follows. Second, to rule out equilibria that hinge on implausible or unreasonable off-path beliefs of the receiver, we invoke a well-established equilibrium refinement, namely the D1 criterion due to Cho and Kreps (1987) and Banks and Sobel (1987). Third, rather than directly examining the sender’s choice of information structures, we investigate an auxiliary game, at the interim stage. In this game, each sender type makes choices regarding bundles of an expected material payoff and an image payoff. Here, the choice of the expected material payoff is constrained to the set of material payoffs attainable in the pure persuasion scenario. This shift in perspective highlights that the extent to which the sender can mold her image through information design is bounded by the scope of persuasion. Crucially, operating in the payoff space provides tractability – it allows us to identify the essential features of D1 equilibria, even when the sender’s optimal information structures cannot be expressed in closed form.

Our first result establishes that all equilibria satisfying the D1 criterion manifest as *semi-*

---

<sup>2</sup>As it will become clear through applications, our framework can accommodate settings where the receiver’s payoff also depends on the sender’s type in a continuation game following the initial interaction.

*separating*. Specifically, we find that there is a unique cutoff point, below which all sender types adopt information structures that reveal their types. In contrast, types exceeding this cutoff pool on the same strategy. As the sender’s image concern heightens, the cutoff decreases monotonically, resulting in an equilibrium transition from full separation to complete pooling. Furthermore, higher types necessarily obtain lower material payoffs in equilibrium. The precise level of material payoff that each below-cutoff type relinquishes for its reputation is pinned down by an envelope formula derived from the local incentive constraints related to separation being optimal. In tandem, all types above the cutoff obtain the minimum material payoff. Our characterization thus implies that the sender’s interim payoffs, both material and image ones, are the same across all equilibria.

Despite the payoff-equivalence for the sender, we find that there can be multiple equilibria. Our next results characterize these. Clearly, all of them can be Pareto ranked based on the receiver’s welfare, which, as it turns out, can vary substantially across equilibria. This is rarely the case in more conventional settings where the sender signals her private information through channels like education, advertising expenditure, or pricing, as the D1 criterion often identifies a unique equilibrium outcome (Cho and Sobel, 1990; Riley, 2001). In contrast, signaling via information design opens the door to a new potential for indeterminacy: Due to the abundance of information structures, there can be numerous alternatives yielding identical sender payoffs yet significantly different receiver payoffs. Therefore, whether the receiver fares better or worse compared to the pure persuasion benchmark may depend on the specific equilibrium chosen. Standard refinements, such as D1, offer no guidance in resolving this issue because they focus on discerning unreasonable payoff incentives of the sender. To advance our understanding nonetheless, we provide sufficient conditions under which the welfare consequences of sender’s image concerns will be robust or sensitive to equilibrium selection. In doing so, we also identify some general properties of the Pareto frontier of the equilibrium set. Most notably, we find that the receiver’s expected payoff is necessarily quasi-concave (quasi-convex) – but not always monotonic – with respect to the sender’s type in any Pareto-optimal (Pareto-worst) equilibrium. Finally, we verify a similar non-monotonicity of the receiver’s welfare when varying the relative strength of the sender’s persuasion and presentation motive.

We complement the general analysis of equilibria by specializing the main results in several classic payoff environments, including those with quadratic losses (e.g. Crawford and

Sobel, 1982; Melumad and Shibano, 1991) and state-independent sender preferences (e.g. Gentzkow and Kamenica, 2016; Lipnowski and Ravid, 2020). These supplementary exercises offer valuable insights into the nature of information structures emerging in equilibrium, an aspect that our “reduced-form” approach falls short in. In particular, we show that in these commonly studied cases, the Pareto frontier of the equilibrium set can often be supported by simple families of information structures, such as censorship, interval disclosure, or a mix between full revelation and total secrecy.

In the last part of the paper, we apply our theory to three different contexts. In each instance, we provide a tangible interpretation of the sender’s type and elucidate the origin of image concerns. Our first application considers a self-signaling environment (Bodner and Prelec, 2003) where the sender and receiver represent two selves of the same agent at different points in time. The application is cast in the context of a mental task; for example, an IQ test or an economic theory problem. We present findings that rationalize how self-image concerns, such as pride, can lead to the avoidance of supportive information that would otherwise help in solving the mental task, and thereby they lead to self-handicapping (Bénabou and Tirole, 2002). Such information avoidance is well-documented empirically (see, e.g., the survey by Golman, Hagmann and Loewenstein, 2017), and previous theoretical work has provided image-concern explanations for it in other settings; for example with moral hazard (Bénabou and Tirole, 2002) or featuring concerns about the own morality (Grossman and Van der Weele, 2017). The innovation here is to consider a setting in which the sender’s scope of information acquisition is unrestricted, thereby strengthening the previous arguments.

Next, we apply our theory to the realm of organizational economics, focusing on a moral-hazard situation where a manager controls the flow of information accessible to a worker. The manager privately knows the extent to which her preferences concerning the worker’s efforts align with the company’s leadership, as opposed to the worker. Further, the manager aspires to project an image of compliance with the company’s leadership, recognizing its positive impact on her career prospects. We find that the manager may choose to hide information from the worker in an effort to impress superiors, even if it is potentially detrimental to the company’s interest. Our result complements Jehiel (2015)’s insights on the drivers of intransparency within organizations. It sheds light on why many companies nowadays move away from performance reviews done solely by (direct) superiors to employing committees that involve third persons.

Finally, we present an application to political economy, inspired by the seminal work of Fernandez and Rodrik (1991) on policy stagnation. We contemplate a scenario where a politician conveys information about a reform to a group of voters. The politician, whose personal interest in the reform remains concealed from the voters, faces a trade-off: providing information skewed in favor of the reform increases its chance of being accepted but may be seen as self-serving. Such a perception will adversely affect the politician’s electability in future campaigns, as voters prefer leaders who act in the public’s best interest rather than their own. Our analysis suggests that in situations where future electoral outcomes hinge heavily on the perceived morality of the politician, she may opt to endorse studies that consistently align with voters’ prior skepticism toward the reform, thereby perpetuating their ignorance and policy stagnation.<sup>3</sup>

**Related literature.** Our paper primarily contributes to the burgeoning literature of Bayesian persuasion and information design. For an excellent overview of this literature, see Bergemann and Morris (2019) and Kamenica (2019). What sets our paper apart is the introduction of a novel dimension alongside the conventional persuasion motive, namely the presentation motive of the sender. Our general model remains agnostic about the origins of this motive. It could capture the instrumental benefits that individuals gain from their reputation in future interactions (Morris, 2001; Sobel, 1985), or stem from a wide array of psychological preferences (Geanakoplos, Pearce and Stacchetti, 1989), such as conformity (Bernheim, 1994), social esteem (Bénabou and Tirole, 2006), or self-image concerns (Baumeister, 1998; Bodner and Prelec, 2003; Köszegi, 2006). Regardless of its source, the critical implication of incorporating this motive into our model is that the sender’s communication strategy will inherently reflect what she privately knows: her own preference type. Several papers have investigated how private sender information may influence persuasion, e.g. Chen and Zhang (2020); Degan and Li (2021); Hedlund (2017); Koessler and Skreta (2022); Perez-Richet (2014). However, these studies typically link the sender’s private information to her interim belief about the payoff-relevant state  $XXX$  (lets verify what these papers do). In contrast, in our setting, the sender’s type is orthogonal to the state and captures relevant aspects of her preferences. More broadly, our work is complementary to a set of recent studies that explore Bayesian

---

<sup>3</sup>Similar in flavor is Kartik and McAfee (2007)’s work on electoral competition in which a candidate’s policy choice signals about his character. In their setting, signaling produces, for example, a failure of the median voter theorem.

persuasion with similar features as in our setting but on the receiver’s side, including reputational concerns (Li, 2022; Salas, 2019), psychological preferences (Lipnowski and Mathevet, 2018; Schweizer and Szech, 2018), and private information (Guo and Shmaya, 2019; Hu and Weng, 2021; Kolotilin, Mylovannov, Zapechelnyuk and Li, 2017) of the receiver.

We also make a contribution to the classic literature on signaling games following Spence (1973)’s seminal work. The semi-separating equilibrium structure, a key feature of our model, has been observed in other signaling contexts in the past (see, e.g., Bernheim, 1994; Cho and Sobel, 1990; Kartik, 2009). In these prior studies, the incomplete separation of types is mainly driven by exogenous constraints on the signal space available to the sender. In contrast, this phenomenon emerges in our model due to a boundary on the interim payoff that is endogenously determined by the scope of persuasion. More substantially, our paper takes on scenarios in which a strategic tension arises between signaling and persuasion, whereas existing research has primarily treated signaling as a means to “persuasion”, that is to influence the receiver’s action. This novel approach not only opens up new applications but also yields non-standard theoretical implications, especially with regard to the receiver’s welfare.<sup>4</sup>

Finally, our paper complements the literature examining reputation-building behavior in repeated interactions (Mailath and Samuelson, 2006, 2015). Within this literature, studies have identified various scenarios where reputational concerns can lead to advantageous outcomes (e.g. Fudenberg and Levine, 1989) or adverse effects (e.g. Ely, Fudenberg and Levine, 2008; Ely and Välimäki, 2003) in terms of welfare consequences. Results from these studies typically involve a rational player (the “normal” type) being motivated to emulate the behavior of a non-strategic player who adheres to an exogenous decision rule (the “behavioral” or “commitment” type). In our setting, whether reputation has good, bad, or ambiguous effects on the receiver’s welfare depends on various details of the game.

The remainder of the paper is organized as follows: Section 2 introduces the formal model. Section 3 presents the main theoretical results characterizing the equilibria and welfare outcomes, accompanied by specific examples in classic payoff environments. Section 4 details real-world applications of our theory. Finally, Section 5 concludes. Technical proofs and

---

<sup>4</sup>A strand of the signaling literature has considered the role of image concerns in cheap-talk communication. However, many of these studies tend to tackle more specialized questions compared to ours, such as whether a desire to influence the receiver’s future decisions (Morris, 2001; Sobel, 1985) or to appear well-informed (Ottaviani and Sørensen, 2006a,b) would encourage or discourage honest communication. Others focus on separate themes, e.g. the potential of third-party bribes to enhance information transmission when the sender likes to be perceived as non-corruptible (Durbin and Iyer, 2009).



analytical details that support the main text are provided in the Appendix.

## 2 Model

We study a communication game between a sender (she) and a receiver (he). There is a state space  $\Omega$ , with a typical state denoted by  $\omega$ , and an action space  $A$ , with a typical action denoted by  $a$ . Both  $A$  and  $\Omega$  are compact metric spaces. The players are uncertain about the state at the outset of the game and share a common prior  $\mu_0 \in \Delta(\Omega)$  about it with full support. The sender moves first by choosing an information structure  $\pi \in \Delta(\Omega \times A)$ , which is a joint distribution of the state and a signal with marginal  $\mu_0$ . The set of all possible information structures is  $\Pi$ . The receiver observes the sender's choice of information structure and the signal realization, and finally chooses an action  $a \in A$ .

**Preferences.** The receiver has a continuous utility function  $u(a, \omega)$  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 with full support. The sender's utility is the sum of two continuous functions,

$$v(a, \omega) + \phi \cdot w(p(\eta), \theta),$$

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

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

So,  $w$  satisfies the Spence-Mirrlees (or increasing difference) condition.

**Strategies and equilibrium.** A pure strategy of the sender is a mapping  $\sigma : \Theta \rightarrow \Pi$  that specifies for each type an information structure. A pure strategy of the receiver is a mapping that specifies a (possibly mixed) action for every possible information structure and signal realization. Under the best response, 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 mixes over actions that maximize his expected payoff. At the same time, the receiver also updates 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.<sup>5</sup> We analyze the perfect Bayesian equilibria in pure sender strategies (Fudenberg and Tirole, 1991, p. 333; henceforth equilibrium).

In our setting, a revelation principle holds. For any equilibrium, there is an equilibrium that is pay-off equivalent *both* in terms of the material and the image payoffs, and in which the receiver follows the sender's recommendation, that is, he chooses  $a$  after observing a signal realization  $s \in A$ . This is not an instance of the classical revelation principle (e.g., Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979) since the payoff of each sender type depends on the agent's posterior belief about her type, which in turn depends on the other types' choices. The proof is relegated to the Appendix A.1, and there we also show that it is without loss to restrict to information structures for which the signal space is contained in the receiver's action space. Given this revelation principle, 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  $\sigma$ . Here, incentive compatibility requires that for every  $\theta \in \Theta$ , the associated information structure  $\pi_\theta$  is a solution to

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

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

**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

---

<sup>5</sup>If the sender could design an information structure over both  $\omega$  and  $\theta$ , and commit to it before learning  $\theta$ , our model would become a special case of Kamenica and Gentzkow (2011). Alternatively, if the design over  $\omega$  and  $\theta$  occurred after the sender (partially) gets to know  $\theta$ , the setting would be similar to Koessler and Skreta (2022). We refrain from an extensive investigation of these alternative specifications in our context, because information design about the strength of image concerns seems difficult to justify in practice.

core idea is to restrict the receiver's out-of-equilibrium beliefs to the sender types that are "most likely" to benefit from deviations to off-path choices. The D1 criterion requires that if, for a type  $\theta$ , there is another type  $\theta'$  that has a strict incentive to deviate to the off-path choice  $\pi \in \Pi^*$  whenever  $\theta$  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  $\theta$ . Formally, given a sender strategy  $\sigma$  and the associated belief system of the receiver, we define, for any  $(\pi, \theta) \in \Pi^* \times \Theta$ , two sets of beliefs as follows:

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

and

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

An equilibrium  $(\sigma, H)$  is selected by the D1 criterion if for any  $\pi \in \Pi^*$  that is not used by any sender type under  $\sigma$ , and for any sender types  $\theta$  and  $\theta'$ ,<sup>6</sup>

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

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

---

<sup>6</sup>Considering also the off-path choices  $\pi \in \Pi \setminus \Pi^*$  will not change the set of equilibria selected by the D1 criterion. This is because, given an equilibrium, for any  $\pi \in \Pi \setminus \Pi^*$ , there is  $\pi' \in \Pi^*$  that yields the same interim material expected payoff to the sender. Specifically,  $\pi' \in \Delta(\Omega \times A)$  is given by the joint distribution of the state and the receiver action induced by  $\pi$  and decision rule of the receiver given  $\pi$ , specified by the equilibrium. 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  $\pi$  if and only if it passes the test at  $\pi'$ .

the sender's persuasion motive is entangled with her presentation motive. To make progress, we simplify the infinite-dimensional maximization problem (2) of the sender by moving the analysis to the interim stage. In particular, instead of analyzing 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.<sup>7</sup> Second, the set of feasible material interim payoffs,  $\{V \subset \mathbb{R} : \exists \pi \in \Pi^* \text{ such that } V = \mathbb{E}_\pi[v(s, \omega)]\}$  is a compact interval. To see this, consider the payoffs

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

and let  $\bar{\pi}$  and  $\underline{\pi}$  be two information structures that give rise to  $\bar{V}$  and  $\underline{V}$ , respectively.<sup>8</sup> Any feasible payoff must be weakly larger than the minimum and weakly smaller than the maximum feasible payoff. Conversely, any payoff in between can be achieved by appropriately mixing the information structures  $\bar{\pi}$  and  $\underline{\pi}$ .<sup>9</sup> Hence, the set of feasible material payoffs is  $[\underline{V}, \bar{V}]$ . Third, as we formally show in the Appendix (Lemma A1), the sender types' preferences have the following single-crossing property in the interim space: for any two bundles  $(V, \eta)$  and  $(V', \eta')$  where  $V < V'$  are feasible material payoffs and  $\eta, \eta' \in \Delta(\Theta)$ , if type  $\theta$  weakly prefers  $(V, \eta)$  over  $(V', \eta')$ , then  $(V, \eta)$  will be strictly preferred over  $(V', \eta')$  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  $\sigma$ , we define  $V(\theta; \sigma) \equiv \mathbb{E}_{\pi_\theta}[v(s, \omega)]$  and  $p(\theta; \sigma) \equiv \mathbb{E}[\tilde{\theta} | \tilde{\theta} :$

---

<sup>7</sup>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).

<sup>8</sup>Note that  $\bar{V}$  and  $\underline{V}$  are well-defined since  $\Pi^*$  is compact with respect to the weak-\* topology and since  $v$  is continuous.

<sup>9</sup>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  $\lambda$ , 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  $V$ .

$\sigma(\tilde{\theta}) = \sigma(\theta)]$ , i.e., the expected material payoff and the perceived image that the strategy induces for each type  $\theta$ , respectively. We say that a type  $\theta$  is *separating* under the strategy  $\sigma$  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  $\theta$  is *pooling*. Our first 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.<sup>10</sup> 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*

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

(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 the existence of a D1 equilibrium — one can always construct a strategy that satisfies (i) and (ii) (recall that the set of feasible material payoffs is  $[V, \bar{V}]$ ). Exactly which D1 equilibrium is chosen 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)

---

<sup>10</sup>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.

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.

**Monotone strategies and incomplete separation.** To begin, we establish some qualitative features of the sender’s strategy based on her equilibrium incentives. Recall that the sender’s central trade-off is between the material benefits derived from persuasion and the reputational gains achieved through presentation. In particular, the sender is willing to sacrifice her material payoff only if doing so results in a more favorable image. Further, since the image payoff function  $w(\cdot)$  satisfies the Spence-Mirrlees (or increasing difference) condition in (1), this kind of “money-burning” incentive is 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 image is higher.

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

Next, we show that a type cannot be pooling unless he receives the minimal feasible material payoff.

**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 weakly 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, 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  $\hat{\theta}$  cluster at the lower boundary of the material-payoff range. Similar incomplete separation at the top has been established in other contexts (Bernheim, 1994; Kartik, 2009). Indeed, Cho and Sobel (1990) demonstrated that this semi-separating structure is inherent to the equilibria selected by D1 in a broad class of costly signaling games with a compact interval of signals available to the sender. The key distinction is that, in our framework, the relevant costs or boundaries are not exogenously imposed on the signals; rather, they arise endogenously from how the receiver reacts under varying information structures.

**The cost of reputation.** We now proceed to characterize the intensity of signaling for types in the separating interval  $[0, \hat{\theta})$ , that is, we characterize how much material payoff will be sacrificed by such types. Here, the central idea is to leverage that the sender's utility function is quasi-linear with respect to her image payoff. This payoff structure reminds of the standard mechanism design setting with transfers. Thus, naturally, we advance the analysis by applying the classical envelope theorem argument (see e.g. Proposition 23.D.2 in Mas-Colell, Whinston and Green, 1995) to the local incentive compatibility constraints of the sender.

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- $\theta$  sender implies the following:

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

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

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

Combining (6) and (7) and dividing them by  $\epsilon$ , 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

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

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.,  $V(0; \sigma) = \bar{V}$ . By combining this boundary condition with 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  $\hat{\theta} \in (0, 1)$ :

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

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

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

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  $I(\cdot)$  is strictly increasing.<sup>11</sup> Also,  $I(\cdot)$  is continuous because  $w(\cdot)$  is continuously differentiable and the type distribution is absolutely continuous.

We distinguish three cases. First, if

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

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

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 over pooling with higher types, contradicting the assumption of  $\hat{\theta} < 1$ . 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 over separation (except for type 0, who may be indifferent), which contradicts the assumption of  $\hat{\theta} > 0$ . This implies that all types must be pooling in any equilibrium, and consequently, we have  $\hat{\theta} = 0$  as the unique cutoff.  $\square$

We close this section with a visual representation of the main findings of Theorem 1. Figure 1 presents all three types of sender's strategy that could emerge in an equilibrium satisfying the D1 criterion. That is, a fully separating equilibrium (Panel a), a semi-separating equilibrium (Panel b), and a fully pooling equilibrium (Panel c).

---

<sup>11</sup>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  $\mathbb{E}[\tilde{\theta}|\tilde{\theta} > \theta] \geq \mathbb{E}[\tilde{\theta}|\tilde{\theta} > \theta']$ .

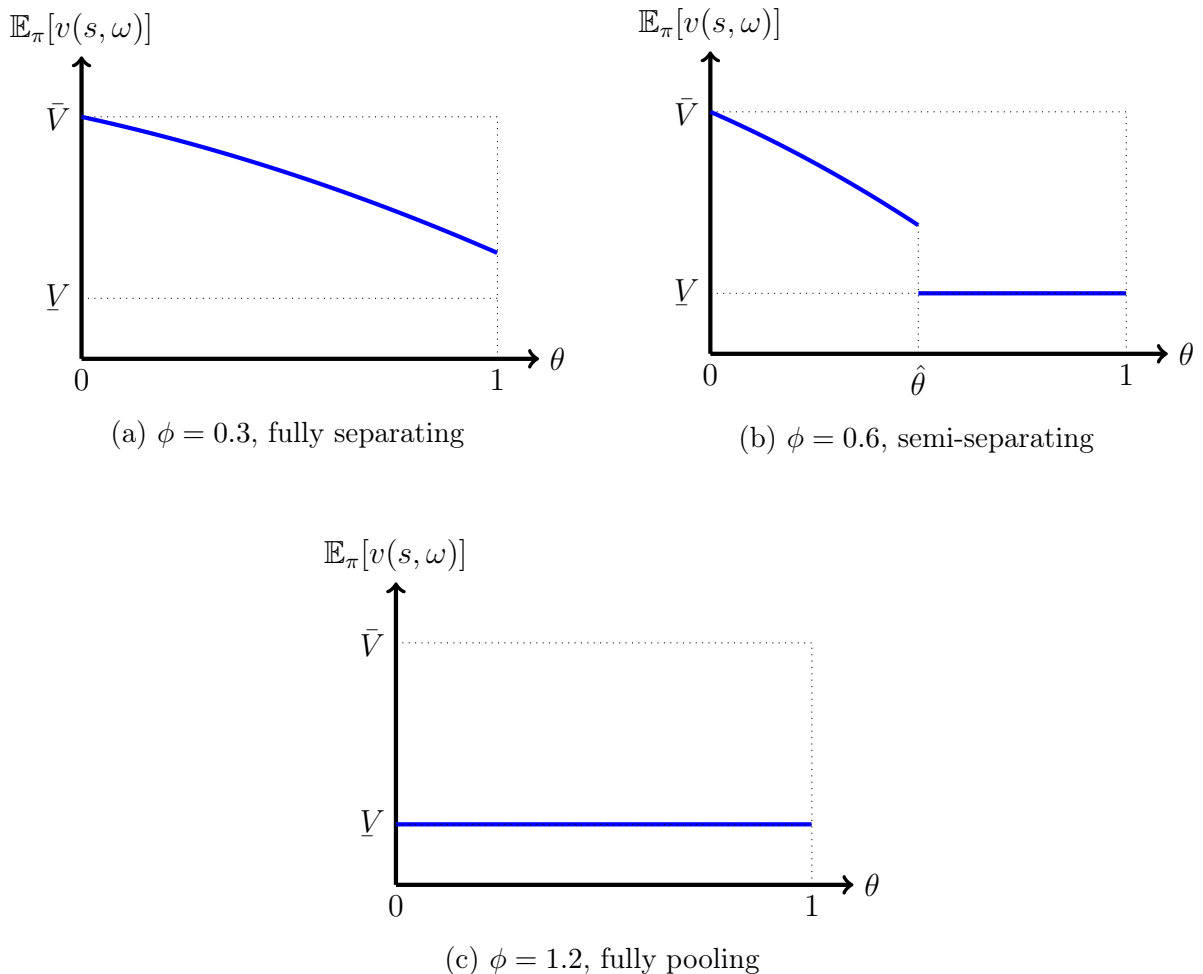


Figure 1: Sender's expected material payoff as a function of her type in a D1 equilibrium, with  $\theta \sim \mathcal{U}[0, 1]$ ,  $w(p, \theta) = p \cdot (\theta + 1)$ ,  $\bar{V} - \underline{V} = 0.6$ , and different  $\phi$ .

### 3.2 Pareto (In)efficiency and Equilibrium Multiplicity

As previously discussed, Theorem 1 reveals that the sender's interim payoffs are equivalent across all D1 equilibria. In particular, the theorem specifies 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 regarding the specific information structure that the sender will adopt, which also means that the receiver's payoff may not be definitely pinned down. This characteristic underscores the distinction of signaling through information design compared to more conventional settings, such as those where senders employ education, advertising, or pricing as signals, as the D1 criterion typically identifies a unique equilibrium outcome in those instances (see e.g. Cho

and Sobel, 1990; Riley, 2001).

Since there is no a priori reason to restrict attention to a specific class of information structures, we pursue three avenues in the following: (i) establishing simple sufficient conditions under which the implications of the sender's image concerns for receiver welfare will be robust or sensitive to equilibrium selection, (ii) analyzing the Pareto-frontier of the equilibrium set, and (iii) applying these overarching findings to various specialized yet classic payoff environments (e.g., quadratic losses or state-independent sender preferences). To simplify the discussion, we make two additional mild assumptions. First, information is valuable to the receiver:  $\bar{U} \equiv \mathbb{E}_{\mu_0} [\max_{a \in A} u(a, \omega)] > \underline{U} \equiv \max_{a \in A} \mathbb{E}_{\mu_0} [u(a, \omega)]$ . Second, in the benchmark scenario in which the sender has *no* image concerns, the receiver's equilibrium payoff – which we denote by  $U^*$  – is uniquely defined.<sup>12</sup>

### 3.2.1 When will sender's image concerns be unequivocally harmful?

When does the presence of sender's image concerns harm the receiver's welfare, irrespective of which D1 equilibrium is selected? A straightforward 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 describing the properties of the best- and worst-case scenario for the receiver. The best case is a *Pareto-optimal* equilibrium, which we define to be a D1 equilibrium for which the sender strategy  $\sigma = \{\pi_\theta\}_{\theta \in \Theta}$  maximizes the receiver's payoffs type-wise; that is, for each sender type<sup>13</sup>

$$\pi(\theta) \in \arg \max_{\pi \in \Pi^*: \mathbb{E}_\pi(v(s, \omega)) = V(\theta; \sigma)} U(\pi; \theta; \sigma). \quad (12)$$

The worst-case is a *Pareto-worst* equilibrium, which minimizes the receiver's payoffs for each sender type.

**Theorem 2.** *If  $U^* = \bar{U}$ , the receiver can never benefit from the presence of the sender's image concerns. Moreover,*

---

<sup>12</sup>Formally, the second assumption requires that  $\mathbb{E}_\pi[u(a, \omega)] = \mathbb{E}_{\pi'}[u(a, \omega)]$  for all  $\pi, \pi' \in \Pi^*$  that maximize the sender's material payoff. Without this assumption, the analysis in this section still applies if we adjust the benchmark referred to in each result accordingly to the maximum or minimum utility attainable by the receiver under any information structure that maximizes the sender's material payoff.

<sup>13</sup>An alternative definition would Pareto-rank equilibria by the *ex-ante* expected welfare of the receiver. Under this definition, multiple equilibria that only differ on a null set of types can all be Pareto-efficient. Otherwise, this alternative definition implies the same characterization of the Pareto-frontier.

- (i) provided that  $\phi < \bar{\phi}$  (so that we have a cutoff type  $\hat{\theta} > 0$ ), there exists a D1 equilibrium in which the receiver is strictly worse off compared to the case without image concerns;<sup>14</sup>
- (ii) in any Pareto-optimal D1 equilibrium, the receiver's expected payoff is decreasing with respect to the sender's type;
- (iii) in any Pareto-worst D1 equilibrium, the receiver's expected payoff is quasi-convex with respect to the sender's type.

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 receiver’s payoff implied by the strategy of a type  $\theta$  is lower than that of a higher type  $\theta' > \theta$ . This equilibrium cannot be Pareto-optimal, because of the following argument: Replacing the information structure that type  $\theta$  initially chooses with an appropriate mix of those used by types 0 and  $\theta'$  will not change the sender’s payoff, but will strictly improve the payoff of the receiver. A similar but slightly more intricate 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

---

<sup>14</sup>The statement remains valid when  $\phi \geq \bar{\phi}$  under the additional assumption that an information structure  $\pi \in \Pi^*$  satisfying  $\mathbb{E}_\pi[v(s, \omega)] = V$  and  $\mathbb{E}_\pi[u(s, \omega)] < \bar{U}$  exists.

every state  $\omega \in \Omega$ :

$$a^* \in \arg \max_{a \in A} u(a, \omega) \iff a^* \in \arg \max_{a \in A} v(a, \omega). \quad (13)$$

When (13) 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  $\Psi : \mathbb{R} \rightarrow \mathbb{R}$ , such that  $u(a, \omega) = \Psi(v(a, \omega))$  for all  $(a, \omega) \in A \times \Omega$ . Panel (a) in Figure 2 depicts the set of implementable material payoff profiles for the case where  $\Psi(\cdot)$  is 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  $\theta$  or  $\phi$ ) will transmit less information to the receiver, therefore leading to a lower receiver welfare.

Perfect alignment of material interests is by far not the only setting that implicates congruent preferences. In Appendix A.7.1, we provide an example where the players disagree on the second-best actions, even though they concur on the first-best actions in every state. Hence, although the material interests of the players are not perfectly aligned, the congruency condition (13) is satisfied. Panel (b) in Figure 2 visualizes the set of implementable material payoff profiles in this example. Especially, the upper curve in red (the lower curve in blue) delineates, for any given level of the sender's payoff  $V \in [V, \bar{V}]$ , the maximum (minimum) payoff that the receiver can attain. Consequently, 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  $A = \Omega = [0, 1]$ ,  $u(a, \omega) = -(a - \omega)^2$ , and  $v(a, \omega) = -(a - a^*(\omega, \theta))^2$ . Specifically, the sender's bliss point is given by  $a^*(\omega, \theta) = f(\theta) \cdot \omega + g(\theta)$ . Communication games in which the players' preferences take the form of such a quadratic loss

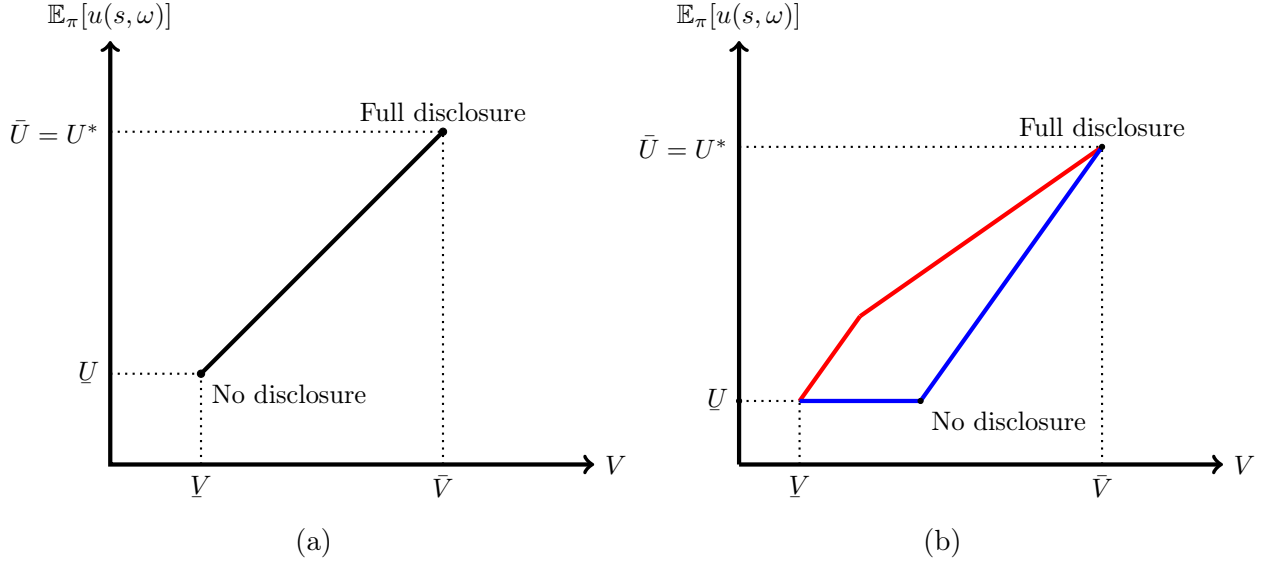


Figure 2: The equilibrium set of implementable payoffs in settings with  $U^* = \bar{U}$ . Panel (a) represents a game where the preferences of the players are perfectly aligned, with  $u(a, \omega) = v(a, \omega)$ ,  $\bar{V} = 0.5$  and  $\underline{V} = 0.1$ . Panel (b) represents a game the players' preferences are not perfectly aligned, but the congruency condition (13) holds (see Appendix A.7.1 for details). 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.

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.7.2, 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  $v(a, \omega) = u(a, \omega)$  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 Theorems 1 and 2 apply. They jointly imply that all types (except possibly type 0) will withhold information from the receiver for signaling purposes. Moreover, given that  $v(a, \omega) = u(a, \omega)$ , 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 and Samuelson, 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. In general, if reputation is “bad”, “good”, or if effects depend on equilibrium selection, depends non-trivially on details of the game, as the next two sections will further demonstrate.

### 3.2.2 When will sender’s image concerns be unequivocally beneficial?

When does the presence of image concerns benefit the receiver, irrespective of which D1 equilibrium is selected? 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,*

- (i) *provided that  $\phi < \bar{\phi}$  (so that the cutoff  $\hat{\theta}$  is smaller than 0)), there exists a D1 equilibrium in which the receiver is strictly better off compared to the case without image concerns;<sup>15</sup>*

---

<sup>15</sup>This statement remains valid when  $\phi \geq \bar{\phi}$  under the additional assumption that an information structure  $\pi \in \Pi^*$  satisfying  $\mathbb{E}_\pi[v(s, \omega)] = \underline{V}$  and  $\mathbb{E}_\pi[u(s, \omega)] > \underline{U}$  exists.

- (ii) in any Pareto-optimal D1 equilibrium, the receiver's expected payoff is quasi-concave with respect to the sender's type;
- (iii) in any Pareto-worst D1 equilibrium, the receiver's expected payoff is increasing with respect to the sender's type.

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 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 engage in a zero-sum (or constant-sum) game. Namely, when there exists a constant  $K \in \mathbb{R}$ , such that  $v(a, \omega) + u(a, \omega) = K$  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. As with perfectly aligned interests (see Example 1), the linear mapping between the players' expected payoffs implies that all D1 equilibria are payoff-equivalent to both players. 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 higher  $\theta$  or  $\phi$ ) will transmit more information to the receiver, therefore boosting his welfare.

**Example 4: Quadratic loss (continued).** Consider again the quadratic-loss games that we introduced in the previous subsection. In Appendix A.7.2, we show that under the condition  $f(\theta) < 0.5$   $\theta \in [0, 1]$ , the original game will be strategically equivalent to one where the sender has the material function  $v(a, \omega) = (a - \omega)^2$  and the image payoff function  $\hat{w}(p(\eta), \theta) = \mathbb{E}_\eta[\tilde{\theta}]/(1 - 2f(\theta))$ . Since  $v(a, \omega) + u(a, \omega) = 0$  for all  $(a, \omega) \in A \times \Omega$ , the transformed game is a zero-sum one regarding the players' material payoffs. Provided that



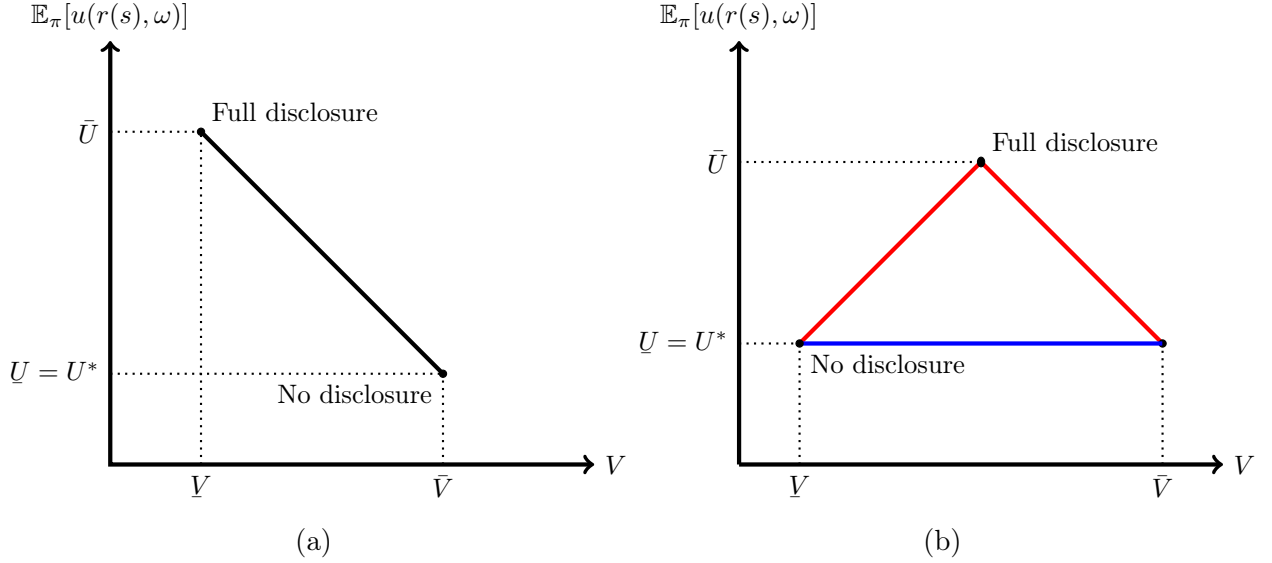


Figure 3: The equilibrium set of implementable payoffs in settings with  $U^* = \underline{U}$ . Panel (a) represents a game where the players have exactly opposite interests over the material outcomes, with  $u(a, \omega) = -v(a, \omega)$ ,  $\bar{V} = 0.5$  and  $\underline{V} = 0.1$ . In panel (b), 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.

condition (1) holds for  $\hat{w}(\cdot)$  – which can occur, for instance, if the interests of higher types are more aligned with the receiver in the sense that  $f'(\cdot) > 0$  – both Theorems 1 and 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 in a pure persuasion setting.

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 below 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 other sufficiency conditions in a setting that allows utilities of the sender and receiver to be non-linear in the state.<sup>16</sup>

**Example 5: State-independent sender preferences, I.** Suppose that  $A = \Omega = \{0, 1\}$ ,  $v(a, \omega) = a$ , and  $u(a, \omega) = \mathbb{1}_{a=\omega}$ . Thus, while the receiver wants to match the state, the sender's preference over material outcomes is state-independent: she always prefers the 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  $\mu_0$  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

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

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 that 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.7.3 for a formal proof). For instance, consider the information structure  $\underline{\pi}^q$

---

<sup>16</sup>See, e.g., Theorem 2 in Kolotilin and Wolitzky (2020). Optimal partial disclosure has been shown to take the form of censorship (Kolotilin, Mylovannov 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 Panscs, 2017).

defined as follows: Conditional on the true state, the signal  $s = 1$  is drawn with probability

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

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  $\underline{\pi}^q$  for any  $q \in [0, 1]$ , 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  $\theta$  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: types at the lower end of the separating interval 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 in the separating interval 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  $u(a, \omega) = a \cdot \omega + (1 - a) \cdot \underline{u}$ , where  $\underline{u} \in (0, 1)$  can be interpreted as the value of the receiver’s outside option  $a = 0$ . 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 Theorems 1 and 3 can be applied to study this example.<sup>17</sup>

### 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 can be uncertain in the sense that the direction of the effect depends on the selected equilibrium.<sup>18</sup>

**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  $\phi$  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 relative to the setting without image concerns and (ii) a D1 equilibrium in which the receiver is and strictly worse-off and Blackwell-less information is transmitted.*<sup>19</sup>

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 rule out equilibria by 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.

---

<sup>17</sup>Under the additional assumption that  $\omega$  is uniformly distributed, we can construct two simple classes of information structures to describe the Pareto-extremal equilibria in closed form. The Pareto-optimal information structures "censor" the state in a lower interval that varies with  $\theta$ . The Pareto-worst information structures "censor" the state in some intermediate interval. To the best of our knowledge such middle censorship has not been observed before. The details of the construction are provided in Appendix A.7.4.

<sup>18</sup>The theorem establishes that the direction of the effect is uncertain *whenever* the sender's image concerns are sufficiently small. Note that, in the other extreme, when the image concern parameter  $\phi$  is sufficiently large, all equilibria are fully pooling. Then, typically, the effect on the receiver's welfare is not uncertain.

<sup>19</sup>Note that the proof of part (iii) of Theorems 2 does not depend on the condition  $U^* = \bar{U}$ . Hence, the quasi-convexity property of the Pareto-worst equilibrium continues to hold even when  $U^* < \bar{U}$ . Likewise, the quasi-concavity property of the Pareto-optimal equilibrium, as identified by Theorem 3, remains valid here despite the condition  $U^* > \underline{U}$ .

**Remark on optimal information structures.** We view the multiplicity of equilibrium information structures in our model rather as a qualification of the information design approach 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 exemplary point within the application in Section 4.2, 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). In our setting, the previous results show that a class of simple information structures (e.g. the censoring of available information) is consistent with equilibrium requirements 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 (Kőszegi, 2006; Schwardmann and Van der Weele, 2019), moralities (Bénabou and Tirole, 2006; Chen and Heese, 2023; 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.

In the context of self-signaling, 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 public to both selves arrives according to a drift-diffusion process (e.g. Chen and Heese, 2023; Henry and Ottaviani, 2019; Morris and Strack, 2019).<sup>20</sup>

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 common state-dependent material payoff  $v(a, \omega)$  from an action choice. 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 private information in the form of the type  $\theta$  that captures the agent’s ability to figure out the solution to the task without any informational assistance. Specifically, the sender-self

---

<sup>20</sup>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.

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  $\theta$ , reflecting the idea that higher types are associated with higher abilities. The agent further derives a “diagnostic utility”  $\psi \cdot \mathbb{E}_\eta[\tilde{\theta}]$  from being perceived as a high type by her receiver-self, where  $\psi > 0$ . It is straightforward to verify that this dual-self game maps into the following specification of our general model where the sender has the utility function  $u^S(a, \omega, \theta, \eta) = v(a, \omega) + (\delta f(\theta) + \psi \cdot \mathbb{E}_\eta[\tilde{\theta}]) / (1 - f(\theta))$ , where  $\delta \equiv \mathbb{E}_{\mu_0}[\max_{a \in A} v(a, \omega)]$  is a constant, while the receiver has the utility function  $u(a, \omega) = v(a, \omega)$ .<sup>21</sup>

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 *et al.* (2017).<sup>22</sup>

## 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 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 reputational concerns of the manager.<sup>23</sup>

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

---

<sup>21</sup>A sender of type  $\theta$  chooses  $\pi \in \Pi^*$  to maximize  $(1 - f(\theta)) \cdot \mathbb{E}_\pi[v(s, \omega)] + \mathbb{E}_{\mu_0}[\max_{a \in A} v(a, \omega)] \cdot f(\theta) + \psi \cdot \mathbb{E}_\eta[\tilde{\theta}]$  (subject to the obedience constraints from the action recommendations  $s$ ). This is equivalent to maximizing  $\mathbb{E}_\pi[v(s, \omega)] + (\delta f(\theta) + \psi \cdot \mathbb{E}_\eta[\tilde{\theta}]) / (1 - f(\theta))$ . Note that the function  $w(p, \theta) = (\delta f(\theta) + \psi p) / (1 - f(\theta))$  is continuously differentiable and satisfies our condition (1).

<sup>22</sup>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).

<sup>23</sup>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.

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  $u(a, \omega) = -(a - \omega)^2$ , 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  $\eta$  is interpreted as the senior management’s belief about the manager’s type. The image payoff specification posits that higher types care more about the impression that they leave to the boss. This seems reasonable because, 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.<sup>24</sup> 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 always 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. 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,

---

<sup>24</sup>See Kamenica and Gentzkow (2011). For details specific to our setting, see also Appendix A.7.2 and the analysis of Example 2 in Section 3.2.1.



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 organizations are surveyed and discussed in, e.g. Hermalin (2001) and 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.<sup>25</sup> 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.

### 4.3 Populist Sentiments and Policy Stagnation

In a seminal study, Fernandez and Rodrik (1991) address why governments often fail to adopt reforms considered efficiency-enhancing by experts, which they describe as “one of the fundamental questions of political economy”. Indeed, this question is particularly puzzling in the current era, where political leaders emphasize the importance of science and evidence-based policy making for progress and growth.<sup>26</sup> Fernandez and Rodrik (1991) demonstrate that such policy stagnation may occur when voters are uncertain about the idiosyncratic

---

<sup>25</sup>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.

<sup>26</sup>See, e.g., Mallapaty (2022), Prillaman (2022), and the article “Politics will be poorer without Angela Merkel’s scientific approach” by the editorial board of *Nature* (2021) .

impacts of the reform *ex ante*, therefore rejecting it even though the reform for sure will benefit a *majority* of the democratic public *ex post* and is welfare-enhancing overall. However, it remains unclear why such uncertainty persists, as the government could in principle seek to educate voters about the potential consequences of the reform, especially given the increasing availability of data and development of information technology. In what follows, we use our framework to show that this phenomenon can be rationalized by the (over-)disciplining effect of reputational concerns of policy makers.

To this end, we describe a stylized model of politics in which information first flows from experts to a politician, and further to a democratic public (represented as a group of voters) who then accept or reject a reform accordingly. The public's prior opinion is marked by reform skepticism. That is, absent persuasive new information about the potential consequences, the public would reject the reform outright, leading to policy stagnation as in Fernandez and Rodrik (1991). The politician could ask experts to provide such information to the public in principle. However, the politician has reputational concerns: the public may perceive her advocacy for the reform as driven by personal interests, hurting her chances in future elections.<sup>27</sup> One may expect these reputation concerns to encourage information sharing, as the politician tries to appear neutral. Yet paradoxically, we demonstrate that these very concerns can perpetuate an equilibrium in which the public remains uninformed. In particular, the public may remain anchored in their initial skepticism towards the reform even when implementing it would actually benefit *all* voters and when the politician could have committed to share that information.

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 proposed reform being publicly debated. To obtain this information, the politician can commission a study  $\pi$  that specifies a distribution of results for each possible state.<sup>28</sup> For instance, the politician may 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 could select 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 (relative to a result opposing

---

<sup>27</sup>The effect of election incentives on politicians' conduct is a prominent focus in the literature on electoral accountability, as surveyed by Ashworth (2012) and Duggan and Martinelli (2017).

<sup>28</sup>For the purpose of the exposition of the applied setting, we allow for any signal space  $S$  not necessarily contained in the receiver's action space  $A$ . The details in the appendix show that this is without loss, given a version of the revelation principle.

it).

In the second stage, the politician observes the result of the study and then chooses whether to keep it private or disclose it to the public in the form of a verifiable report. Implementing the reform ( $a = 1$ ) enhances every voter's welfare by 1 if  $\omega = 1$  but reduces welfare by 1 if  $\omega = 0$ , relative to maintaining the status quo ( $a = 0$ ). The public is initially skeptical of the reform, with a common prior belief  $\mu_0 \equiv \Pr(\omega = 1) \in (0, 0.5)$ . Thus, only when the disclosed result is sufficiently compelling to overcome these predispositions, the politician's communication is effective in generating support for 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  $\theta$  is linked to the "corruptness" of the politician, or more broadly speaking to the private interests that the politician has in the reform or her willingness to push actions against the interest of the public (e.g. advocating  $a = 1$  when  $\omega = 0$ ). Higher values of  $\theta$  reflect greater alignment with public interest, consistent with lower private benefits from the adoption of the reform.

The third and last stage introduces reputational concerns of the politician. In this stage, the politician runs against another candidate in an election. Each voter receives a payoff  $\alpha\theta + \epsilon$  if a politician of type  $\theta$  is elected to office, where  $\alpha > 0$  is a parameter and  $\epsilon \in \mathbb{R}$  is a common preference shock (e.g. changes in the economic environment) drawn according to a continuous cumulative distribution function  $G$  with full support. This specification could be interpreted as capturing the intuition that more corrupt politicians are more likely to act against the public's interest once elected. Alternatively, it may simply reflect that voters intrinsically desire more "ethical" politicians in office. We fix the voters' expected payoff from electing the opposing candidate as  $\underline{u}$ , which satisfies  $G(\underline{u}) \in (0, 1)$ . Voters observe the preference shock  $\epsilon$ , but not the politician's type  $\theta$ . However, they make a Bayesian inference about  $\theta$  based on the politician's "past record", which in our model amounts to whether and how she previously attempted to influence public opinion about the reform through her strategy which involves the commissioning of a study and subsequent disclosure choices. Thus, voters will support the politician if and only if  $\alpha p + \epsilon \geq \underline{u}$ , where  $p$  is the voters' posterior expectation regarding the politician's type. Accordingly, the likelihood of the politician winning the election is  $1 - G(\underline{u} - \alpha p)$ . Finally, the politician's payoffs upon winning and losing the election are given by  $w_2(\theta)$  and  $w_3(\theta)$ , respectively. We assume that the ratio  $(w_2(\theta) - w_3(\theta))/w_1(\theta)$  is

continuously differentiable and strictly increasing in  $\theta$ , so that less corrupt types care more about the election relative to the reform.<sup>29</sup> This monotonicity assumption ensures that the relevant single-crossing property holds, creating the potential for the politician to shape her image through strategic communication.<sup>30</sup>

In Appendix A.9.2, we demonstrate in detail how this political economy model can be solved in reduced form with the previous analysis. Specifically, we show that the equilibrium incentives of the politician can be mapped into a specialized setting of Example 5 studied in Section 3.2. The previous findings apply: higher types necessarily commission studies less favorable to the reform, expressing fewer and/or weaker endorsements of the reform to the public.

This equilibrium outcome and its key driving force can be intuitively understood in terms of “populist sentiments”: the politician seeks to position herself in the debate surrounding the reform – through strategically commissioning and revealing study results – in a way that appeals to the public she is “on their side” (i.e., not corrupt). Indeed, this interpretation aligns with the standard political science definition of populism as “a political philosophy supporting [...] the people in their struggle against the privileged elite” (see the corresponding item in the [American heritage dictionary](#)).<sup>31</sup>

A critical insight from our analysis is that the effect of populism on the politician’s communication strategy is ambivalent, and does not necessarily translate monotonically into the welfare of the public. Specifically, consider an increase in  $\alpha$ , which could represent heightened populist sentiment among the public, whereby voters become more concerned about the politician’s corruption when deciding their electoral support. As we formally show in Appendix A.9.3, this parallels the effect of increasing  $\phi$  – the relative weight that the sender places on image versus material payoffs – in the general model. Intuitively, a larger  $\alpha$  strengthens the signaling motives of the politician because her public image becomes more decisive for the election outcome. Thus, similar to the results from Section 3.2, the public’s welfare is non-monotone in  $\alpha$ . In particular, when  $\alpha$  is sufficiently small, all equilibria are fully separating –

---

<sup>29</sup>Given that  $w_1(\cdot)$  is strictly decreasing, the desired monotonicity holds if all types are purely office-motivated (i.e.,  $w_2(\cdot) - w_3(\cdot)$  is a constant function), a setting commonly studied in the literature on electoral competition (Persson and Tabellini, 2002). Moreover, as we formally show in Appendix A.9.1, this monotonicity assumption can also be derived from a setting where the politician is both office- and policy-motivated.

<sup>30</sup>In comparison, if  $((w_2(\cdot) - w_3(\cdot))/w_1(\cdot))' < 0$  holds, then more corrupt types have higher incentives to be elected. In this case, the disciplining effect of the election concerns disappears – the unique equilibrium is such that all types promote the reform in the same way as in the benchmark game with  $\phi = 0$ .

<sup>31</sup>See also Mudde (2004), Acemoglu, Egorov and Sonin (2013), and the recent VoxEU debate on populism (available at <https://cepr.org/debates/populism>).

meaning voters can fully distinguish between corrupt and non-corrupt politicians – and the payoff from reform choices increase in  $\alpha$  in the Pareto optimal equilibrium (see Panel (b) of Figure 3 and the discussion surrounding Example 5). However, when  $\alpha$  is large, reputational incentives may “over-discipline” the politician.

In the extreme when the future electoral outcome hinges heavily on the perceived morality of the politician (sufficiently high  $\alpha$ ), all politician types conform to the public’s prior skepticism by recommending the status quo with probability one, regardless of the state. As such, the public learns nothing about either the reform or the politician’s integrity. Critically, even when the reform would enhance welfare for all parties involved, the public’s equilibrium belief remains equal to their skeptical prior, so the status quo persists and policies stagnate. This result demonstrates how Fernandez and Rodrik (1991)’ observation that asymmetric information of the public can result in policy stagnation is even valid in a “general equilibrium” model in which the public’s information is endogenous and all uncertainty could principally be resolved.

## 5 Conclusion

We have developed a novel framework that enables studying the strategic disclosure of information as jointly determined by two countervailing forces: the standard motive to persuade an audience towards actions preferred by the sender, and the relatively underexplored motive to manage impressions regarding unobserved characteristics of the sender. Our main results delineate the Pareto frontier of the equilibrium set. For example, they demonstrate that the sender’s information choices and the receiver’s welfare can exhibit intriguing non-monotonicity with respect to the relative strength of the two motives. Since concerns about one’s image admit diverse interpretations ranging from psychological preferences to reputational considerations in dynamic interactions, our model offers broad applicability. We demonstrate this versatility across multiple contexts, generating new insights into a number of issues that have received considerable attention from researchers and practitioners. These insights include observations related to information avoidance in self-signaling situations, harmful intransparencies in organisations due to managers’ career concerns, and how heightened populism can engender policy stagnation.

We close by suggesting two directions for future research, each of which relaxes some re-

restrictions made in our current framework. First, our model assumes that the sender’s payoff is separable with respect to the material allocations and her type-specific gains from reputation. This quasi-linear structure – which is commonly employed in applied works – greatly simplifies our analysis, as it ensures that the relevant single-crossing property holds at the interim stage. Plainly, the analysis continues to hold if one directly imposes this single-crossing property on the sender’s interim payoffs. However, this assumption remains restrictive, because it requires the payoff difference between any pair of *probability distributions* over ex-post allocations to be single-crossing in the sender’s type. As discussed by Kartik, Lee and Rappoport (2023) (and see also Kushnir and Liu, 2019), only a limited set of ex post payoff specifications could generate this property. Nevertheless, Chen, Ishida and Suen (2022) recently provide a general analysis for costly signaling under double-crossing preferences, which nest single-crossing as a special case. By identifying environments where the double-crossing property naturally arises at the interim stage, one could combine our techniques and those by Chen *et al.* (2022) to obtain additional insights.

Second, our model precludes any correlation between the state and the sender’s type, which may limit its suitability for some settings and merits further study. For instance, the alignment of preferences between managers and their supervisors may vary across states, and the necessity of reform could correlate with the corruptness of the incumbent politician. Such correlations naturally lead to an informed principal problem, which is known to be difficult in the literature (Myerson, 1983). Although a comprehensive analysis tackling this challenging issue is beyond the scope of our paper, we note that there have been several exciting works recently aiming to develop a toolkit for studying informed principal problems in the context of information design (e.g. Koessler and Skreta, 2022; Zapechelnyuk, 2023). Integrating these cutting-edge approaches with our framework provides fertile ground for new applications.

# Appendix

## A.1 A Revelation Principle in Our Setting

We are going to prove a revelation principle in our setting. This revelation principle asserts that it is without loss to consider information structures for which the signal space  $S$  is contained in the receiver's action space  $A$  and for which it is weakly optimal for the receiver to choose action  $a = s$  after a signal realization  $s \in A$ . That is, all sender types choose information structures from the set

$$\Pi^* \equiv \left\{ \pi \in \Pi : s \in \arg \max_{a \in A} \mathbb{E}[u(a, \omega) | s; \pi] \quad \forall s \in \text{supp}(\pi) \right\}. \quad (16)$$

This revelation principle is not an instance of the standard revelation principle (e.g., Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979) since the sender's payoffs depend on the agents' posterior belief about the sender type (and these beliefs in turn depend on the sender strategy as a whole function of the sender type.)

So, for a second—unlike in the specification of the main model—we also consider information structures  $\pi$  for which the signal space is not contained in the receiver's action space.

Take any perfect Bayesian equilibrium of the sender-receiver game, given by a sender strategy  $\sigma = \{\pi_\theta\}_{\theta \in \Theta}$ , decision rules  $\hat{a}(\pi) : \text{supp}(\pi) \rightarrow \Delta(A)$  for each information structure  $\pi$ , and a belief system  $H$  that attaches a belief distribution  $\eta(\pi)$  to any possible information structure. We will construct now a payoff-equivalent equilibrium given by a sender strategy  $\sigma'$  for which  $\sigma(\theta) \in \Pi^*$  for all  $\theta \in \Theta$  and a belief system  $H'$  that attaches a belief distribution  $\eta'(\pi)$  to any information structure  $\pi \in \Pi$ . The receiver's strategy is given by the collection of functions  $\hat{a}'(\pi)$  for any  $\pi \in \Pi^*$  that map any signal  $s \in A$  to itself and by some arbitrary sequentially rational decision rule  $\hat{a}'(\pi)$  if  $\pi \in \Pi \setminus \Pi^*$ .<sup>32</sup>

First, we construct  $\sigma'$ . For any  $\theta \in \Theta$ , consider  $\sigma(\theta)$ . Together  $\sigma(\theta)$  and  $\hat{a}(\sigma(\theta))$  induce a joint distribution of the state  $\omega$  and the action  $a$  taken by the receiver. Set  $\sigma'(\theta)$  to be equal to this induced distribution. One checks that  $\sigma'(\theta)$  satisfies the obedience constraints;

---

<sup>32</sup>Note that this pay-off equivalent equilibrium, strictly speaking, will be an equilibrium in a (slightly) different game than the equilibrium that we start with in this proof. In particular, we do not specify off-path beliefs and receiver mappings for information structures for which the signal space is not contained in  $A$ . This is purely for notational convenience since otherwise the sender's strategy space is not a set (but a class) but we would like to use set notation.

thus,  $\sigma'(\theta) \in \Pi^*$ . Before specifying the belief system  $H'$ , we make a preliminary observation. Take any  $\theta, \theta'$  for which  $\sigma'(\theta) = \sigma'(\theta')$ . This implies that the expected material payoff of the sender is the same under  $\sigma'(\theta)$  and  $\sigma'(\theta')$ . Further, the expected material payoff of the sender is the same under  $\sigma(\theta)$  and  $\sigma(\theta')$ . This implies that

$$\mathbb{E}_{\eta(\pi(\theta))}(\tilde{\theta}) = \mathbb{E}_{\eta(\pi(\theta'))}(\tilde{\theta}) \quad (17)$$

since otherwise, in the equilibrium that we started with, either  $\theta$  has a strict incentive to mimick  $\theta'$  or the other way around. This leads to a contradiction, since then the equilibrium that we started with would not be an equilibrium. The observation means that, for all types  $\theta$  that are assigned the same information structure under  $\sigma'$ , they must have gotten the same image payoff under  $\sigma$ . Moreover, by construction, if  $\theta$  and  $\theta'$  pool given  $\sigma$ , necessarily they pool under  $\sigma'$ . Together, this implies that

$$\mathbb{E}(\tilde{\theta} | \sigma'(\tilde{\theta}) = \sigma'(\theta)) = \mathbb{E}_{\eta(\theta)}(\tilde{\theta}) \quad (18)$$

for all  $\theta \in \Theta$ . Since Bayes' rule implies that  $\mathbb{E}_{\eta'(\sigma'(\theta))}(\tilde{\theta}) = \mathbb{E}(\tilde{\theta} | \sigma'(\tilde{\theta}) = \sigma'(\theta))$ , we obtain

$$\mathbb{E}_{\eta'(\sigma'(\theta))}(\tilde{\theta}) = \mathbb{E}_{\eta(\sigma(\theta))}(\tilde{\theta}) \quad (19)$$

for all  $\theta \in \Theta$ . We complete the specification of  $H'$  by setting

$$\eta'(\pi) = \eta(\pi) \quad (20)$$

for any  $\pi \in \Pi$  for which there is no  $\theta \in \Theta$  so that  $\sigma'(\theta) = \pi$ .

Finally, we argue that  $(\sigma', H')$  constitutes a perfect Bayesian equilibrium. This follows since the material payoff of each type  $\theta$  is the the same under  $\sigma'$  and  $\sigma$  by construction, and the image payoff is also the same given (19). Therefore, since  $\sigma$  was part of an equilibrium, no type has an incentive to mimick any other type given  $(\sigma', H')$ . Further, since any out-of-equilibrium information structure  $\pi \in \Pi$  leads to the same image payoff under  $H$  and  $H'$ , and since  $(\sigma, H)$  were part of a PBE, no type has an incentive to deviate to an out-of-equilibrium information given  $(\sigma', H')$ .



## A.2 The Single-Crossing Property

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

- (a) *all types  $\theta' < \theta$  strictly prefer  $(V, \eta)$  over  $(V', \eta)$ ,*
- (b) *all types  $\theta' > \theta$  strictly prefer  $(V', \eta')$  over  $(V, \eta)$ .*

PROOF. Indifference of type  $\theta$  means

$$V - V' = \phi \cdot [w(p(\eta'), \theta) - w(p(\eta), \theta)]. \quad (21)$$

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 (21) 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.3 Proof of Lemma 1

Let  $\sigma$  be an equilibrium strategy. Incentive compatibility implies that, for all sender types  $\theta, \theta' \in \Theta$  with  $\theta' < \theta$ ,

$$V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta) \geq V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta) \quad (22)$$

and

$$V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta') \geq V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta'). \quad (23)$$

Summing up (22) and (23), we obtain (with some rearrangement)

$$w(p(\theta; \sigma), \theta) - w(p(\theta'; \sigma), \theta) \geq w(p(\theta; \sigma), \theta') - w(p(\theta'; \sigma), \theta'). \quad (24)$$

Since  $\theta > \theta'$  and  $w(\cdot)$  has strictly increasing differences, (24) 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 (23) to hold it is necessary that  $V(\theta; \sigma) \leq V(\theta'; \sigma)$ .  $\square$

## A.4 Proof of Lemma 2

Take an equilibrium with sender strategy  $\sigma$  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 9).

Recall the notation  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}]$  for the receiver's posterior expectation about the sender's type upon observing the latter player chooses  $\pi^\varepsilon$ . We argue that in equilibrium,  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] \geq \sup J$  must hold. To prove this argument, we distinguish two cases. First, suppose that  $\pi^\varepsilon$  is a choice on the equilibrium path under the strategy  $\sigma$ , 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  $\theta$  was arbitrary and the receiver's on-path beliefs must satisfy Bayes' rule, the claim  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] = \mathbb{E}[\tilde{\theta}|\pi^\varepsilon; \sigma] \geq \sup J$  immediately follows.

Second, suppose that no types will choose  $\pi^\varepsilon$  under the strategy  $\sigma$ . In this case, take any  $\theta \in J$  with  $\theta < \sup J$ . Since the type distribution has full support, it holds  $\sup J > \mathbb{E}(\tilde{\theta}|\tilde{\theta} \in J) > \inf J$ . Therefore, there is  $\epsilon > 0$  small enough so that, for the off-path belief  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] = \sup J$ ,  $\theta$  strictly prefers  $\pi^\varepsilon$  over  $\pi$ , and for  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] = \inf J$ ,  $\theta$  strictly prefers  $\pi$  over  $\pi^\varepsilon$ . An application of the intermediate value theorem shows that 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  $\pi^\varepsilon$ , then the type- $\theta$  sender would be indifferent between choosing  $\pi$  and  $\pi^\varepsilon$ . Moreover, given that  $V - \varepsilon < V$ , any sender with  $\theta' < \theta$  would strictly prefer  $\pi$  to  $\pi^\varepsilon$  whenever type  $\theta$  is being indifferent between these two pairs. Hence, due to the single-crossing property (Lemma A1), the D1 criterion requires that the receiver assigns zero weight to types  $\theta' < \theta$  upon observing that  $\pi^\varepsilon$  was chosen by the sender. As a result, we have  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] \geq \theta$ . Since the choice of  $\theta < \sup J$  was arbitrary, it again follows that the claim  $\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] \geq \sup J$  must hold.

Next, since the distribution of the type  $\theta$  has full support, we further have

$$\mathbb{E}_{\eta(\pi^\varepsilon)}[\tilde{\theta}] \geq \sup J > \mathbb{E}[\tilde{\theta}|\tilde{\theta} \in J].$$

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

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

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

## A.5 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  $(\sigma, H)$  constitute a D1-equilibrium. The belief system is such that, for any  $\pi \in \Pi^*$ :

- If  $\hat{\theta} = +\infty$  and  $\mathbb{E}_\pi[v(s, \omega)] \geq \bar{V} - \phi \int_0^1 \frac{\partial w(x, x)}{\partial p} dx$ , the receiver assigns probability one to the unique type  $\theta \in [0, 1]$  for which  $\mathbb{E}_{\pi_\theta}[v(s, \omega)] = \mathbb{E}_\pi[v(s, \omega)]$ ;
- If  $\hat{\theta} < 1$  and  $\mathbb{E}_\pi[v(s, \omega)] > \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx$ , the receiver assigns probability one to the unique type  $\theta \in [0, 1]$  for which  $\mathbb{E}_{\pi_\theta}[v(s, \omega)] = \mathbb{E}_\pi[v(s, \omega)]$ ;
- If  $\hat{\theta} < 1$  and  $\mathbb{E}_\pi[v(s, \omega)] = V$ , the receiver's posterior belief is  $\mathbb{E}_{\eta(\pi)}(\tilde{\theta}) = \mathbb{E}_{\mu_0}(\tilde{\theta} | \tilde{\theta} \in [\hat{\theta}, 1])$ .

Finally, the remaining out-of-equilibrium beliefs  $\eta(\pi)$  for  $\pi \in \Pi \setminus \Pi^*$  are given as follows:

- Both, if  $\hat{\theta} = +\infty$  and  $\bar{V} - \phi \int_0^1 \frac{\partial w(x, x)}{\partial p} dx > \mathbb{E}_\pi[v(s, \omega)] > V$ ,
- or if  $\hat{\theta} < 1$  and  $\bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx \geq \mathbb{E}_\pi[v(s, \omega)] > V$ ,

the receiver assigns probability one to type  $\min\{\hat{\theta}, 1\}$ .

**Sequential Rationality.** Note that, given  $H$ , any  $\pi$  and  $\pi'$  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  $\bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx \geq V > \underline{V}$  will be sub-optimal for the sender, as she could always obtain a higher material payoff by deviating to  $\pi_{\min(1, \hat{\theta}) - \epsilon}$  while undermining her image only by arbitrarily small  $\epsilon > 0$ . By the same argument, when  $\hat{\theta} = +\infty$ , any information structure

that induces a material payoff  $V$  with  $\bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx > V > \underline{V}$  will be sub-optimal for the sender. 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  $\sigma$ , 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 from 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  $\sigma$ , 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  $\pi$  that yields the minimal material payoff  $\underline{V}$ ) and separating herself (by choosing some  $\pi$  that gives rise to  $\mathbb{E}_{\pi}[v(s, \omega)] = \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx$ ). 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 information structure  $\pi' \in \Pi^*$ . For any type  $\theta$ , provided that  $D^0(\pi', \theta)$  – the set of beliefs for which  $\theta$  weakly prefers to deviate from her choice  $\pi_{\theta}$  to  $\pi'$  – 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  $\theta'$  with  $\pi_{\theta'} \neq \pi_\theta$ . We have already shown that this type has *strict* incentives *not* to mimic  $\theta$ . This implies  $\underline{p}(\pi', \theta') > \underline{p}(\pi', \theta)$ , and therefore  $D^0(\pi', \theta') \subsetneq D(\pi', \theta)$ . Conversely, for any type  $\theta''$  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  $\theta''$  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  $\mathbb{E}_{\pi'}[v(s, \omega)] = V(\theta; \sigma)$ . If  $\hat{\theta} = +\infty$  (i.e., the strategy  $\sigma$  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  $\theta \in [0, 1]$ . This implies that the D1 criterion imposes no constraint on the out-of-equilibrium belief after observing  $\pi'$ , and our choice is consistent. If  $\hat{\theta} \in [0, 1]$  (i.e., the strategy  $\sigma$  is semi-separating), then it is necessary that

$$V < \mathbb{E}_{\pi'}[v(s, \omega)] \leq \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx. \quad (25)$$

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

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

where the first inequality follows from condition (i) and since  $w(x, \cdot)$  is strictly increasing in the second argument. The second inequality follows by two observations: First, the cut-off type  $\hat{\theta}$  is indifferent between joining the pool—which is the type's choice given condition (ii)—and separating as well as choosing  $\pi'$  assuming that it leads to a belief  $\underline{p}(\pi', \hat{\theta})$ , that is,

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

Second, the single-crossing property of Lemma A1 then implies the second inequality. It is then clear that  $\underline{p}(\pi', \theta) > \underline{p}(\pi', \hat{\theta})$  for all  $\theta < \hat{\theta}$ . Further, since

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

Lemma A1 and (25) jointly imply that

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

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  $\pi'$ . However, our belief system was just chosen this way.  $\square$

## A.6 Proof of Theorem 2

Part (i): Take any information structure  $\pi^N \in \Pi^*$  that minimizes the sender's payoff among the information structures that provide no information about the state, and let  $V^N = \mathbb{E}_{\pi^N}[v(s, \omega)]$ . By assumption,  $U^*$  is unique and given by  $\bar{U}$ . Further, we assumed that  $\bar{U} > \underline{U}$ , so it is necessary that  $\bar{V} > V^N$ . Otherwise, an information structure that provides no information to the receiver would be sender-optimal in the pure persuasion benchmark, which necessitates  $\underline{U} = \bar{U}$ . 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). Therefore, there must exist a non-empty interval  $(0, \check{\theta}) \subseteq (0, \hat{\theta})$  such that  $V(\theta; \sigma) \geq V^N$  holds for all  $\theta \in (0, \check{\theta})$ .

Now, 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  $\tilde{\pi}_\theta$ : conditional on each state, 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  $\pi^N$ . It is straightforward to verify that  $\tilde{\pi}_\theta \in \Pi^*$ ,  $\mathbb{E}_{\tilde{\pi}_\theta}[v(s, \omega)] = V(\theta; \sigma)$ , and

$$\mathbb{E}_{\tilde{\pi}_\theta}[u(s, \omega)] = \lambda(\theta) \cdot \bar{U} + (1 - \lambda(\theta)) \cdot \underline{U} < \bar{U}. \quad (26)$$

Next, define a strategy  $\check{\sigma}$  for the sender as follows: for all  $\theta \in (0, \check{\theta})$ , let  $\check{\sigma}(\theta) = \tilde{\pi}_\theta$ ; for all other  $\theta$ , let  $\check{\sigma}(\theta) = \sigma(\theta)$ . By Theorem 1,  $\check{\sigma}$  is part of a D1 equilibrium. Moreover, since the type distribution is continuous and has full support, (26) 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.

Assume, for the sake of contradiction, that the receiver's expected payoff is not decreasing everywhere. Then, there must exist  $\theta, \theta' \in [0, 1]$  such that  $\theta < \theta'$  and

$$\mathbb{E}_{\pi_\theta}[u(s, \omega)] < \mathbb{E}_{\pi_{\theta'}}[u(s, \omega)] \leq \bar{U}. \quad (27)$$

If  $V(\theta; \sigma) = V(\theta'; \sigma)$ , we could simply ask type  $\theta$  to adopt the same information structure as type  $\theta'$ , thereby increasing the welfare of the receiver without affecting the sender's. Hence, without loss of generality, we may focus on the scenario  $\bar{V} \geq V(\theta; \sigma) > V(\theta'; \sigma)$ . Now consider the following information structure  $\tilde{\pi}_\theta$ : conditional on each state, with probability  $\lambda = (V(\theta; \sigma) - V(\theta'; \sigma)) / (\bar{V} - V(\theta'; \sigma)) \in [0, 1]$ , 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  $\tilde{\pi}_\theta \in \Pi^*$ ,  $\mathbb{E}_{\tilde{\pi}_\theta}[v(s, \omega)] = V(\theta; \sigma)$ , and

$$\mathbb{E}_{\tilde{\pi}_\theta}[u(s, \omega)] = \lambda \cdot \bar{U} + (1 - \lambda) \cdot \mathbb{E}_{\pi_{\theta'}}[u(s, \omega)] > \mathbb{E}_{\pi_\theta}[u(s, \omega)]. \quad (28)$$

Therefore, it is possible to construct a D1 equilibrium strategy  $\tilde{\sigma}$  that always yields a weakly higher payoff to the receiver than  $\sigma$ , and this payoff difference will even be strict when the sender's type is  $\theta$ . Hence, the strategy  $\sigma$  cannot be Pareto-optimal if the associated payoff for the receiver is not decreasing everywhere within the interval  $[0, 1]$ .

Part (iii): To prove quasi-convexity, it suffices to demonstrate that the receiver's payoff is 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. Define  $V_{min}^N$  and  $V_{max}^N$  as the minimum and maximum material payoffs that the sender may obtain when the receiver acts under no information, respectively. Note that these two values may differ since if the receiver is indifferent given no information, the information structure has some freedom in choosing the recommended action of the receiver. Let  $\pi_{min}^N$  and  $\pi_{max}^N$  be the information structures that lead to these two material payoffs for the sender, respectively. Also, let  $\theta_{min}^N = \sup\{\theta \in [0, 1] : V(\theta; \sigma) \geq V_{max}^N\}$  and  $\theta_{max}^N = \inf\{\theta \in [0, 1] : V(\theta; \sigma) \leq V_{min}^N\}$ .

First, we argue that the receiver's expected payoff must be decreasing everywhere on  $[0, \theta_{min}^N]$ . To prove this, suppose by contradiction that there exist  $\theta, \theta' \in [0, \theta_{min}^N]$  such that  $\theta < \theta'$  and

$$\underline{U} \leq \mathbb{E}_{\pi_\theta}[u(s, \omega)] < \mathbb{E}_{\pi_{\theta'}}[u(s, \omega)]. \quad (29)$$

If  $V(\theta; \sigma) = V(\theta'; \sigma)$ , we could simply ask type  $\theta'$  to adopt the same information structure as type  $\theta$ , which clearly decreases the welfare of the receiver without affecting the sender's. Hence, without loss of generality, we may focus on the scenario  $V(\theta; \sigma) > V(\theta'; \sigma) \geq V_{max}^N$ . Next, consider the following information structure  $\tilde{\pi}_{\theta'}$ : conditional on each state, with probability  $\lambda' = (V(\theta'; \sigma) - V_{max}^N)/(V(\theta; \sigma) - V_{max}^N) \in [0, 1]$ , the information structure generates a signal  $s$  according to  $\pi_{\theta}$ ; with the remaining probability  $1 - \lambda'$ , the signal is generated according to  $\pi_{max}^N$ . It is straightforward to check that  $\tilde{\pi}_{\theta'} \in \Pi^*$ ,  $\mathbb{E}_{\tilde{\pi}_{\theta'}}[v(s, \omega)] = V(\theta'; \sigma)$ , and

$$\mathbb{E}_{\tilde{\pi}_{\theta'}}[u(s, \omega)] = \lambda' \cdot \mathbb{E}_{\pi_{\theta}}[u(s, \omega)] + (1 - \lambda') \cdot \underline{U} < \mathbb{E}_{\pi_{\theta'}}[u(s, \omega)]. \quad (30)$$

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

Second, for types  $\theta \in [\theta_{min}^N, \theta_{max}^N]$ , we claim that the receiver's payoff is constant in  $\theta$  in the Pareto-worst equilibrium. To see this, note that we can replace the info structure of each sender type  $\theta \in [\theta_{min}^N, \theta_{max}^N]$  with one mixing over  $\pi_{min}$  and  $\pi_{max}$  so that they give the sender payoff  $V(\theta)$ . Then, the receiver's payoff is  $\underline{U}$  since he receives no information from  $\theta$ . So, indeed, in the Pareto-worst equilibrium, the receiver's payoff is constant on this interval.

. Thus, the receiver's expected payoff will stay constant at  $\underline{U}$  within this interval.

Lastly, we argue that the receiver's expected payoff must be increasing everywhere on  $[\theta_{max}^N, 1]$ . Suppose not. Then, there must exist  $\theta, \theta' \in [\theta_{max}^N, 1]$  such that  $\theta < \theta'$  and

$$\underline{U} \leq \mathbb{E}_{\pi_{\theta'}}[u(s, \omega)] < \mathbb{E}_{\pi_{\theta}}[u(s, \omega)]. \quad (31)$$

If  $V(\theta; \sigma) = V(\theta'; \sigma)$ , it would be feasible to have type  $\theta$  adopt the same information structure as type  $\theta'$ , which clearly decreases the welfare of the receiver without altering the sender's. Hence, without loss, we may focus on the scenario  $V(\theta'; \sigma) < V(\theta; \sigma) \leq V_{min}^N$ . Using a similar construction of "grand" information structures involving  $\pi_{min}^N$  as before, it can be shown that there exists a D1 equilibrium strategy that always gives a weakly lower payoff to the receiver than  $\sigma$ , and this payoff difference will even be strict when the sender's type is  $\theta$ . Hence, the receiver's expected payoff must be increasing on  $[\theta_{max}^N, 1]$ . This concludes our proof of the receiver's expected payoff being quasi-convex in the whole interval  $[0, 1]$ .  $\square$



## A.7 Results and Proofs Related to the Examples

### A.7.1 The Utility-Frontier with Almost-Perfectly-Aligned Preferences

In the following, we provide formal details regarding the second instance of congruent preferences (Example 1) that we discussed in the main text. Formally, we specify  $A = \Omega = \{-1, 0, 1\}$  and define the material payoff functions for the players as follows:

$$u(a, \omega) = \begin{cases} 1 & \text{if } a = \omega, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad v(a, \omega) = \begin{cases} 1 & \text{if } a = \omega, \\ 0 & \text{if } a \neq \omega \text{ and } a \neq -1, \\ -1 & \text{if } a \neq \omega \text{ and } a = -1. \end{cases} \quad (32)$$

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 with 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 selects it by mistake. By contrast, the receiver is indifferent between different types of errors. Despite the discrepancy in the payoff functions, the congruency condition (13) is met since the players agree on the first-best action in each state.

Let the prior distribution  $\mu_0$  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, denoted as  $\Pr(a = \omega)$ ; (ii) the total probability that the receiver will wrongly take the action  $a = -1$ , denoted as  $\Pr(a = -1 | \omega \neq -1)$ . 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, 0.5 is also an upper bound for  $\Pr(a = -1 | \omega \neq -1)$ : If  $\Pr(a = -1 | \omega \neq -1) > 0.5$ , 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 the recommended action  $-1$ .

Now consider an information structure  $\underline{\pi} \in \Pi$  which recommends the action  $a = -1$  with probability one 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  $\underline{\pi} \in \Pi^*$  and that  $\underline{\pi}$  achieves the two aforementioned 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,  $\pi$  must give the lowest possible payoff to the sender among all information structures, thus  $\bar{V} = \mathbb{E}_{\pi}(v(s, \omega))$ . A calculation shows  $\mathbb{E}_{\pi}(v(s, \omega)) = 0$ .

Given the analysis above, we deduce that the set of implementable payoff profiles, formally defined as  $\mathcal{W} = \{(V, U) : \exists \pi \in \Pi^* \text{ such that } V = \mathbb{E}_{\pi}[v(s, w)] \text{ and } U = \mathbb{E}_{\pi}[u(s, w)]\}$ , must fall within the rectangle  $[V, \bar{V}] \times [\underline{U}, \bar{U}] = [0, 1] \times [0.4, 1]$ . 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 maximum and minimum 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 boils down to identifying the set of  $\Pr(a = -1 | \omega \neq -1)$  values that the sender can 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  $\pi^U \in \Pi^*$  characterized by the following conditional probabilities (of recommending different actions in different states):  $\pi^U(-1 | -1) = 1$ ,  $\pi^U(1 | 1) = \pi^U(-1 | 1) = \pi^U(-1 | 0) = 0.5$ ,  $\pi^U(0 | 0) = (U - 0.4)/0.4$ , and  $\pi^U(1 | 0) = (0.8 - U)/0.4$ . The resulting payoff to the sender,  $U - 0.4$ , is the minimal one across all information structures that induce the receiver to choose  $a = \omega$  with probability  $U$ . Consequently, for all  $U \in [0.4, 0.6]$ ,  $(U - 0.4, U)$  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  $\pi^U \in \Pi^*$  with  $\pi^U(-1 | -1) = 1$ ,  $\pi^U(1 | 1) = \pi^U(0 | 0) = (U - 0.2)/0.8$ , and  $\pi^U(-1 | 1) = \pi^U(-1 | 0) = (1 - U)/0.8$ . Thus, for each  $U \in (0.6, 1]$ ,  $(2U - 1, U)$  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 maximum material payoff of the sender is achieved when the receiver *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  $(0, 0.4)$  and  $(0.4, 0.4)$  are implementable and  $\underline{U} = 0.4$ , any payoff profile  $(V, 0.4)$  with  $V \in (0, 0.4)$  is also a boundary point of  $\mathcal{W}$  given the convexity of  $\mathcal{W}$ . Taken together, we obtain the blue curve depicted in the figure.

### A.7.2 Transforming the Quadratic-Loss Games

Consider the quadratic-loss game that we discussed in Examples 2 and 4. For any  $\pi \in \Pi^*$ , the obedience constraints of the receiver imply  $s = \mathbb{E}_\pi(\omega'|s)$ . As a result, the expected material loss of a type- $\theta$  sender is

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

where the outer expectation is taken with respect to  $\omega$  and  $s$  (but not  $\omega'$ ). The second equality follows from several arguments. First, the law of iterated expectation yields  $\mathbb{E}_\pi [\mathbb{E}[\omega'|s]] = \mathbb{E}_{\mu_0}[\omega]$ , where we denote by  $\mu_0$  the prior belief about the state. Second, it holds  $\mathbb{E}_\pi [\mathbb{E}[\omega'|s]\omega] = \mathbb{E}_s(\mathbb{E}_\omega(\mathbb{E}(\omega'|s) \cdot \omega|s)) = \mathbb{E}_s(\mathbb{E}(\omega'|s)^2) = \mathbb{E}_\pi(\mathbb{E}(\omega'|s)^2)$ , where the first step is due to Fubini's theorem, which says that we can integrate sequentially, first over  $\omega$  and then over  $s$ . The second step follows from linearity of the expected value, and the third step is valid since the function to be integrated does not depend on  $\omega$ . For the third equality above, we use  $K(\theta) \equiv \mathbb{E}_{\mu_0} [(a^*(\omega, \theta))^2] - 2g(\theta) \cdot \mathbb{E}_{\mu_0}[\omega]$  to collect all the ( $\theta$ -specific) constant terms. A similar calculation shows that  $\mathbb{E}_\pi [(s - \omega)^2] = -\mathbb{E}_\pi[\mathbb{E}[\omega'|s]^2] + \mathbb{E}_{\mu_0}[\omega^2]$ .

Now, suppose that  $f(\theta) > 0.5$  for all  $\theta \in [0, 1]$  and compare the following two utility functions of the sender, given by a material and an image payoff function each,  $v(a, \omega) - (a - a^*(\omega, \theta))^2 + \phi \cdot$  and  $w(p(\eta), \theta)$  compared with  $\hat{v}(\omega, \theta) = -(a - \omega)^2$  and  $\hat{w}(p(\eta), \theta) = w(p(\eta), \theta)/(2f(\theta) - 1)$ . We claim that, these two utility functions represent the same preference over the pairs  $(\pi, \eta)$  with  $\pi \in \Pi^*$  and  $\eta$  any belief over the types, for each type  $\theta \in [0, 1]$ . This is because, for all  $\theta \in [0, 1]$  and all such  $(\pi, \eta)$  and  $(\pi', \eta')$ , we have

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

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  $\hat{v}(\cdot)$  and  $\hat{w}(\cdot)$ .

Similarly, if  $f(\theta) < 0.5$  for all  $\theta \in [0, 1]$ , 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.7.3 Receiver-Optimality in Example 5

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.

Take any information structure  $\pi \in \Pi^*$  that recommends the high action with the unconditional probability  $q$ , and let  $\pi(a|\omega)$  be the conditional probability that it recommends action  $a$  when the state is  $\omega$ . Then, it is necessary that

$$\mu_0 \cdot \pi(1|1) + (1 - \mu_0) \cdot \pi(1|0) = q. \quad (33)$$

Therefore, the receiver's expected utility under  $\pi$  is given by

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

REcall that the information structure  $\bar{\pi}^q$  has  $\bar{\pi}^q(1|0) = 0$  when  $q \in [0, \mu_0]$ , so we see that it is receiver-optimal in this case. At the same time, note that, using (33), the receiver's expected utility can also be written as

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

Recall that when  $q \in (\mu_0, 2\mu_0]$ , then  $\bar{\pi}^q(1|1) = 1$ . Thus,  $\bar{\pi}$  is also receiver-optimal in this case.

#### A.7.4 Pareto-Extremal Equilibria in Example 6

Consider the setting specified in Example 6, and suppose the state  $\omega$  is uniformly distributed on  $[0, 1]$ . Since  $v(a, \omega) = a$ , the sender's expected material payoff equals the probability of the receiver taking action  $a = 1$ . Clearly, the minimum of this probability is 0, which the sender can achieve by providing no information. Below we prove that the maximum of this probability is  $2 - 2\underline{u}$ .

Take any information structure  $\pi \in \Pi^*$ , and let  $\pi(1|\omega)$  be the probability that it generates the signal  $s = 1$ . We argue that, for the purpose of maximizing the sender's material payoff, it is without loss of generality to focus on  $\pi$  with increasing  $\pi(1|\cdot)$ . To see this, consider any  $\pi \in \Pi^*$  and suppose  $\pi(1|\omega) > \pi(1|\omega')$  for some  $\omega, \omega'$  with  $\omega < \omega'$ . Then, one can construct a new information structure  $\hat{\pi}$  by swapping the distributions of the recommended action for  $\omega$  and  $\omega'$ . Given that  $\omega < \omega'$ , this relaxes the obedience constraints (which require  $\mathbb{E}[\omega|s = 1] \geq \underline{u}$  and  $\mathbb{E}[\omega|s = 0] \leq \underline{u}$ ) and preserves the overall probability of  $a = 1$  since we are only changing the information structure on a measure zero of states; so it gives the same sender payoff.

Next, take any  $\pi \in \Pi^*$  such that  $\pi(1|\cdot)$  is increasing. Let  $\hat{\omega} = \inf\{\omega : \pi(1|\omega) = 1\}$ . We argue that if  $\hat{\omega} > 1 - 2\underline{u}$ , then  $\pi$  cannot be maximizing the sender's material payoff. Without loss of generality, suppose that  $\mathbb{E}[\omega|s = 1] = \underline{u}$  under  $\pi$ . Together with  $\hat{\omega} > 1 - 2\underline{u}$ , this implies that the probability of getting the signal  $s = 1$  is strictly positive when the state lies in  $[0, 1 - 2\underline{u})$ . But then, one can reassign this probability mass to the interval  $[1 - 2\underline{u}, \hat{\omega})$  without violating the incentive compatibility constraints. In fact, doing so will relax the previously binding constraint  $\mathbb{E}[\omega|s = 1] = \underline{u}$ , meaning that it would become feasible for the sender to induce the receiver to take action  $a = 1$  with even higher probability.

In sum, the above arguments demonstrate that the preceding arguments show that the sender can attain her maximum material payoff using the following information structure, which generates a deterministic signal  $s$  contingent on the true state:  $s = 1$  if  $\omega \geq 1 - 2\underline{u}$ , and  $s = 0$  otherwise. Under this information structure, the total probability of the receiver choosing action  $a = 1$  is  $2\underline{u}$ .

Lastly, we show that how one can use two simple classes of “interval disclosure” information structures to describe the entire Pareto-frontier of the equilibrium set. For this purpose, for every  $q \in [0, 2\underline{u}]$ , we define two different information structures. The first, denoted as  $\bar{\pi}^q$ , generates a deterministic signal  $s$  contingent on the true state:  $s = 1$  if  $\omega \geq 1 - q$ , and  $s = 0$

otherwise. The second information structure, denoted as  $\underline{\pi}^q$ , also follows a deterministic signal-generating rule:  $s = 1$  if  $\omega \in [\underline{u} - q/2, \underline{u} + q/2]$ , and  $s = 0$  otherwise. It can be verified that both  $\bar{\pi}^q$  and  $\underline{\pi}^q$  can induce the receiver to choose the non-default action  $a = 1$  with a probability of  $q$ . Furthermore,  $\bar{\pi}^q$  ( $\underline{\pi}^q$ ) yields the highest (lowest) expected payoff to the receiver among all information structures that implement the same marginal distribution of actions: For instance, if an information structure  $\pi \in \Pi^*$ , which induces  $\Pr(a = 1) = q$  like  $\bar{\pi}^q$ , advises the receiver to choose  $a = 0$  with some positive probability when  $\omega \geq 1 - q$ , it must also suggest  $a = 1$  in certain situations when  $\omega < 1 - q$ . By exchanging these two recommendations, we can create an information structure that increases the receiver's payoff while keeping his total probability of choosing  $a = 1$  unchanged. Hence, similar to Example 5, there exists a Pareto-optimal (Pareto-worst) D1 equilibrium in which each type  $\theta$  chooses the information structure  $\bar{\pi}^{q(\theta)}$  ( $\underline{\pi}^{q(\theta)}$ ), where  $q(\theta)$  represents the probability that type  $\theta$  would induce the receiver to take action  $a = 1$ .

## A.8 Proof of Theorem 4

Take the information structure  $\pi^N \in \Pi^*$  that minimizes the expected material payoff of the sender across all information structures in  $\Pi^*$  that convey no information about the state to the receiver. Take the information structure  $\pi^F \in \Pi^*$  that minimizes the expected material payoff of the sender across all information structures in  $\Pi^*$  that convey full information about the state to the receiver. Let  $V^N$  and  $V^F$  be the expected material payoffs under  $\pi^N$  and  $\pi^F$ , respectively. Since the receiver's payoff  $U^*$  in the pure persuasion benchmark satisfies  $U^* \in (\underline{U}, \bar{U})$ , it is necessary that  $\bar{V} > \max\{V^N, V^F\}$ . Otherwise, providing full or no information is sender-optimal in the benchmark, which would imply  $U^* = \bar{U}$ , or  $U^* = \underline{U}$ .

If  $\phi$  is sufficiently small, all D1 equilibria will be fully separating and the interim payoff of the highest type  $V(1; \sigma)$  will be sufficiently close to  $\bar{V}$  so that  $V(1; \sigma) > \max\{V^N, V^F\}$  irrespective of the which D1 equilibrium strategy  $\sigma$  is selected. In particular, there exists a D1 equilibrium in which each type  $\theta$  uses a “grand” information structure that mixes appropriately between the sender-optimal information structure  $\bar{\pi}$  absent image concerns and the information structure  $\pi^N$ . Clearly, 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  $\pi^F$ . Clearly, the receiver must be strictly better off in this equilibrium relative to the equilibrium without image concerns.  $\square$

## A.9 Additional Details of the Application in Section 4.3

### A.9.1 Micro-Founding the Monotonicity Assumption in Section 4.3

In this subsection, we provide a setting of electoral competition which endogenizes the key assumption in Section 4.3, namely, that the ratio  $(w_2(\cdot) - w_3(\cdot))/w_1(\cdot)$  is strictly increasing. We suppose that, in the third stage, the incumbent politician described in Section 4.3 – who we now call candidate  $A$  – competes with another candidate  $B$  (challenger) for an election. Each candidate  $j = A, B$  has a private type  $\theta_j \in [0, 1]$ , which is independently distributed with a mean denoted by  $\bar{\theta}_j$ .

The candidate who wins the election will get to choose a policy  $y \in \mathbb{R}$ . The voters have a common preference over policies, represented by  $-|y - y^*|$ . For each candidate  $j = A, B$ , with probability  $\theta_j$ , she will have the same policy preference as the voters. With the remaining probability  $1 - \theta_j$ , the candidate's preference will be  $-|y - (y^* + 1)|$ . Thus, the higher  $\theta_j$  (i.e., the less corrupt the candidate), the more likely that the candidate will act perfectly according to the voters' interest once elected.

Recall that the voters may update their belief about candidate  $A$ 's type upon observing the latter's choices in previous stages (whereas the belief about candidate  $B$  is given by the prior). Let  $\epsilon$  be the common preference shock that directly adds to each voter's utility whenever  $A$  is elected, and  $p$  is the public posterior about  $A$ 's type. It is straightforward to show that voters would support candidate  $A$  if and only if  $\epsilon \geq \bar{\theta}_B - p$ . The winning probability of candidate  $A$  is then given by  $1 - G(\bar{\theta}_B - p)$ .

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

$$G(\bar{\theta}_B - p) \cdot [-\theta_A(1 - \bar{\theta}_B) - (1 - \theta_A)\bar{\theta}_B] = -G(\bar{\theta}_B - p) \cdot (w_2(\theta_A) - w_3(\theta_A)),$$

where  $w_2(\theta_A) = 0$  and  $w_3(\theta_A) = -\theta_A - \bar{\theta}_B + 2\bar{\theta}_B\theta_A$  captures the expected payoffs upon winning and losing the election, respectively. Given that  $w_1(\cdot)$  is strictly decreasing, it is easy to check that  $(w_2(\cdot) - w_3(\cdot))/w_1(\cdot)$  is strictly increasing in  $\theta$  whenever  $\bar{\theta}_B < 0.5$  is additionally satisfied.

### A.9.2 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  $\Theta_\pi$  be the set of types that choose the study  $\pi$  in an equilibrium. Note that disclosing a result  $s$  with  $\pi(s|1)/\pi(s|0) \geq \ell_0 \equiv (1 - \mu_0)/\mu_0$  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 image payoff than disclosing  $s$ . In this game, the analogue of the single-crossing property holds for bundles of expected material payoffs attached to strategies of the sender (which are given by the choice of study and subsequent disclosure choices) and expected beliefs about the sender type. This way, the set  $\Theta_\pi$  partitions into two (potentially empty) and disjoint sets  $\Theta_\pi^-$  and  $\Theta_\pi^+$  so that all types in  $\Theta_\pi^-$  are strictly smaller than all types in  $\Theta_\pi^+$  and do not disclose  $s$  while all types in  $\Theta_\pi^+$  disclose  $s$ . If  $\Theta_\pi^-$  is non-empty, then the ordering of the sets  $\Theta_\pi^-$  and  $\Theta_\pi^+$  implies that non-disclosure leads to a strictly lower image payoff as well as a weakly lower material payoff than disclosure. This cannot be in equilibrium, so that  $\Theta_\pi^-$  must be empty. The argument just made is a "classic unraveling argument", and shows that, 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 voters' posterior belief about the politician's type would only depend on the chosen study. Consequently, a type- $\theta$  politician obtains the following payoff from choosing a study  $\pi$ :

$$\Pr(\pi(s|1)/\pi(s|0) \geq \ell_0) \cdot w_1(\theta) + (1 - G(\underline{u} - \alpha p)) \cdot w_2(\theta) + G(\underline{u} - \alpha p) \cdot w_3(\theta), \quad (34)$$

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 ( $s = 0$ ) about the reform. 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 (34) is equivalent to

$$\max_{\pi} \Pr(s = 1 | \pi) + \frac{w_2(\theta)}{w_1(\theta)} - G(\underline{u} - \alpha p) \cdot \frac{w_2(\theta) - w_3(\theta)}{w_1(\theta)},$$

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  $\pi(1|1)/\pi(1|0) \geq \ell_0$ . Thus, the equilibrium problem of the current application maps into



our setting by specializing Example 5 of Section 3.2 with

$$u(a, \omega) = \mathbb{1}_{a=\omega}, \quad v(a, \omega) = \mathbb{1}_{a=1}, \quad \phi = 1, \quad \text{and} \quad w(p, \theta) = \frac{w_2(\theta)}{w_1(\theta)} - G(\underline{u} - \alpha p) \cdot \frac{w_2(\theta) - w_3(\theta)}{w_1(\theta)}.$$

(Note here that in the main text we described the receiver's utilities as  $u(a, \omega) = 0$  if  $a = 0$  and  $u(a, \omega) = 1$  if  $a = 1 = \omega$  and  $u(a, \omega) = -1$  if  $a = 1$  and  $\omega = 0$ . However, this utility specification leads to the same best response of the receiver as  $u(a, \omega) = \mathbb{1}_{a=\omega}$  as a function of the belief over the state  $\omega$ .)

### A.9.3 Non-Monotone Welfare Effects of Populism

We argue that a change in  $\alpha$  can have a non-monotone effect on the welfare of the public. For this purpose, we parameterize the reduced-form version of the dynamic game with  $w_2(\theta) = 0$ ,  $w_3(\theta)/w_1(\theta) = -\theta - 1$ , and  $\epsilon$  is uniformly distributed on  $[0, 1]$ . Assume also that  $1 > \underline{u} > \alpha$  always holds. Taken together, we have

$$w(p, \theta) = -(\underline{u} - \alpha p) \cdot (\theta + 1) \quad \text{and} \quad \frac{\partial w(p, \theta)}{\partial p} = \alpha(\theta + 1).$$

It then follows from Theorem 1 and our analysis of Example 5 that in any D1 equilibrium, all types  $\theta$  below a unique cutoff  $\hat{\theta}$  are separating, and each  $\theta < \hat{\theta}$  will commission a study that implements the reform with the following probability:

$$q(\theta) = 2\mu_0 - \int_0^\theta \alpha(x + 1)dx = 2\mu_0 - \alpha \left( \frac{\theta^2}{2} + \theta \right).$$

In contrast, all types  $\theta \geq \hat{\theta}$  will opt for a completely uninformative study, in which case the reform will for sure be rejected by the public. Additionally, it can be verified that

- $\hat{\theta} = +\infty$  when  $\alpha < 4\mu_0/3$ ;
- $\hat{\theta} \in (0, 1)$  and is strictly decreasing in  $\alpha$  when  $4\mu_0/3 \leq \alpha < 2\mu_0/\mathbb{E}_\Gamma[\tilde{\theta}]$ ;
- $\hat{\theta} = 0$  when  $\alpha \geq 2\mu_0/\mathbb{E}_\Gamma[\tilde{\theta}]$ .

Now consider the Pareto-optimal equilibrium which, as shown in Example 5, can be sustained by using the family of information structures (14). When  $\alpha$  is sufficiently small, the equilibrium is fully separating, and higher types provide more information about the reform to the public. In this case, a local increase in  $\alpha$  will benefit the public by incentivizing the

politician to be even more honest. However, a large  $\alpha$  could hurt the public by incentivizing the politician to provide less information (when some types are already recommending against reforms that are beneficial) and/or hindering the learning of the politician's true type (when pooling occurs in equilibrium).

## 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. (1982). A self-presentational view of social phenomena. *Psychological bulletin*, **91** (1), 3.
- (1998). The self. In D. Gilbert, S. Fiske and G. Lindzey (eds.), *The Handbook of Social Psychology*, vol. 1, McGraw-Hill, pp. 680–740.
- 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.
- BERGEMANN, D. and MORRIS, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, **57** (1), 44–95.
- 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, C.-H., ISHIDA, J. and SUEN, W. (2022). Signaling under double-crossing preferences. *Econometrica*, **90** (3), 1225–1260.
- CHEN, S. and HEESE, C. (2023). Fishing for good news: Motivated information acquisition, CRC TR 224 Discussion Paper No. 223.
- CHEN, Y. and ZHANG, J. (2020). Signalling by Bayesian persuasion and pricing strategy. *The Economic Journal*, **130** (628), 976–1007.
- 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.
- DEGAN, A. and LI, M. (2021). Persuasion with costly precision. *Economic Theory*, **72** (3), 869–908.
- 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.
- DURBIN, E. and IYER, G. (2009). Corruptible advice. *American Economic Journal: Microeconomics*, **1** (2), 220–242.
- EDITORIALS (2021). Politics will be poorer without Angela Merkel’s scientific approach. *Nature*, **597** (7922), 304–304.
- 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. and LEVINE, D. K. (1989). Reputation and equilibrium selection in games with a patient player. *Econometrica*, **57** (4), 759–778.
- , 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.
- GEANAKOPOLOS, 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. (2016). A Rothschild-Stiglitz approach to Bayesian persuasion. *American Economic Review, Papers & Proceedings*, **106** (5), 597–601.
- GIBBARD, A. (1973). Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pp. 587–601.
- GOFFMAN, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor.
- GOLMAN, R., HAGMANN, D. and LOEWENSTEIN, G. (2017). Information avoidance. *Journal of Economic Literature*, **55** (1), 96–135.
- GREEN, J. and LAFFONT, J.-J. (1977). Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica: Journal of the Econometric Society*, pp. 427–438.
- 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.
- HU, J. and WENG, X. (2021). Robust persuasion of a privately informed receiver. *Economic Theory*, **72**, 909–953.
- 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. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, **11**, 249–272.
- 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.
- , LEE, S. and RAPPOPORT, D. (2023). Single-crossing differences in convex environments. *Review of Economic Studies*, forthcoming.
- and MCAFEE, R. P. (2007). Signaling character in electoral competition. *American Economic Review*, **97** (3), 852–870.
- KOESSLER, F. and SKRETA, V. (2022). Informed information design. *Journal of Political Economy*, forthcoming.
- 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.
- KUSHNIR, A. and LIU, S. (2019). On the equivalence of Bayesian and dominant strategy implementation for environments with nonlinear utilities. *Economic Theory*, **67** (3), 617–644.
- LI, H. (2022). Transparency and policymaking with endogenous information provision. ArXiv preprint arXiv:2204.08876.
- LIPNOWSKI, E. and MATHEVET, L. (2018). Disclosure to a psychological audience. *American Economic Journal: Microeconomics*, **10** (4), 67–93.
- and RAVID, D. (2020). Cheap talk with transparent motives. *Econometrica*, **88** (4), 1631–1660.
- MAILATH, G. J. (1987). Incentive compatibility in signaling games with a continuum of types. *Econometrica*, **55** (6), 1349–1365.

- and SAMUELSON, L. (2006). *Repeated Games and Reputations: Long-Run Relationships*. Oxford University Press.
- and — (2015). Reputations in repeated games. In H. P. Young and S. Zamir (eds.), *Handbook of Game Theory with Economic Applications*, vol. 4, Elsevier, pp. 165–238.
- 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. (1979). Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pp. 61–73.
- (1983). Mechanism design by an informed principal. *Econometrica*, **51** (6), 1767–1797.
- (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.
- OTTAVIANI, M. and SØRENSEN, P. N. (2006a). Professional advice. *Journal of Economic Theory*, **126** (1), 120–142.
- and SØRENSEN, P. N. (2006b). Reputational cheap talk. *The RAND Journal of Economics*, **37** (1), 155–175.
- 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.
- RAYO, L. and SEGAL, I. (2010). Optimal information disclosure. *Journal of Political Economy*, **118** (5), 949–987.
- RILEY, J. G. (2001). Silver signals: Twenty-five years of screening and signaling. *Journal of Economic Literature*, **39** (2), 432–478.
- SALAS, C. (2019). Persuading policy-makers. *Journal of Theoretical Politics*, **31** (4), 507–542.
- 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*, 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>.
- SOBEL, J. (1985). A theory of credibility. *The Review of Economic Studies*, **52** (4), 557–573.
- SPENCE, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, **87** (3), 355–374.
- 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.
- ZAPECHELNYUK, A. (2023). On the equivalence of information design by uninformed and informed principals. *Economic Theory*, forthcoming.
- ZHONG, W. (2018). Information design possibility set, working paper.