Information Design with Image Concerns
— preliminary and incomplete —

Carl Heese and Shuo Liu*

November 2022

Abstract

In many economic situations, a sender communicates strategically with a receiver not only to influence his decision-making but also to influence how certain unobserved characteristics of herself (e.g., loyalty, integrity, or unselfishness) are perceived. To study such strategic interactions, we introduce image concerns into the canonical framework of information design. We characterize how the sender optimally sacrifices her persuasive influence on the decision in order to boost her reputation, by manipulating the communication about a payoff-relevant state. We describe when this harms and benefits the receiver unambiguously, revealing also that effects often depend on the selected equilibrium and are non-monotone in the sender's characteristics. We illustrate our findings within various standard preference settings (e.g., with quadratic losses or state-independent sender preferences). Finally, the model provides insights into a wide range of applications, such as how performance evaluation in hierarchical organizations can backfire by engendering harmful upward signalling, and if and when competitive elections may discipline the instrumentalisation of media channels by politicians.

Keywords: information design, persuasion, image concerns, signalling

JEL Classification: C72, D72, D82, M50

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

1 Introduction

This paper develops and applies a novel model of strategic communication and information control. In many economic situations, a sender strategically chooses how to communicate with a receiver not only to influence the decision-making of the later. Instead, the sender *also* cares about how the communication reflects about certain unobserved characteristics of herself, e.g., loyalty, unselfishness, or integrity. These inferences about the sender type may matter for decisions by a third party, at a later point of time, or only because the sender is "behavioural".

Consider a politician who seeks to take influence on current policy-making through media channels in order to advance his own interests, which might not align with the interests of the common voter. The politician *also* attempts to being perceived as defending the common voters' interests.[1] This may be rational since his public image may matter for future elections.

A similar tension arises in situations of information control: consider a hierarchical organization with senior management, a mid-manager (he) and a subordinate (she). The mid-manager controls the information flow to the subordinate about the requirements of tasks, and by doing so, may take influence on her effort level. He *also* cares about the signal that his communication strategy sends to senior management, knowing that mid-managers who appear to have internalized the company guidelines to combating moral hazard will be rewarded.

To model these strategic interactions, we introduce image (or reputational) concerns into the canonical framework of information design (Kamenica and Gentzkow, 2011). There is a pay-off relevant state, a sender (he) with a one-dimensional private preference type, and a receiver. The sender chooses how to communicate with the receiver in the form of a joint distribution of a recommended action and a pay-off relevant state. The receiver chooses an action upon observing the recommendation. The sender has the standard *persuasion motive*; that is, he cares about the action of the receiver. Additionally, he cares about the belief implied by his communication strategy about her private preference type. Thus, there is an additional *signaling motive*. This motive might reflect that the implied belief matters for a

[1] In particular, *populist* politics typically attempts to being perceived as defending the common voter against the elite. See the *American Heritage Dictionary* that defines "populism" as a political philosophy supporting the rights and power of the people in their struggle against the privileged elite." See http://ahdictionary.com/word/search.html?q=populism, and also Acemoglu, Egorov and Sonin (2013) who follow this notion.

third party, future interactions, or for behavioral reasons.

As typical in signaling games, choices of off-equilibrium beliefs generate a large multiplicity of equilibria in our setting. To rule out unreasonable equilibria, we invoke a standard equilibrium refinement, the D1 criterion by Cho and Kreps (1987).[2]

Our first set of results characterizes the structure of equilibria satisfying the D1 criterion. All equilibria are *semi-separating*: There is a unique cutoff so that in any equilibrium, all types below the cutoff separate by using different strategies, and all types above pool on the same strategy. We go on by characterizing the (opportunity) cost of signaling the own preference type for the types on the separating interval. Our characterization will imply that all equilibria are payoff-equivalent for all sender types. Thus, they are Pareto ranked, only varying in the receiver's payoff from her decision.[3]

The second set of results characterizes the implications of image concerns for the receiver's welfare from her decision. We provide sufficient conditions for when the receiver is unambiguously worse (better) off relative to the setting without image concerns. For a large set of environments, however, the welfare effect is ambiguous. The direction of it depends on the selected equilibrium and on the distribution of sender types. The equilibria differ by the signals chosen by the sender types, reflecting different *cultures of communication* (compare with Schelling, 1980).[4] Since all equilibria survive standard refinements, differences are *not* driven by choices of off-path beliefs.

We provide further results about the equilibrium set: we characterize its Pareto frontier and show that information transmission about the state is often non-monotone in the sender type. In other words, "honesty" about the state depends non-trivially on the sender's type, which may reflect a politician's integrity, or similar traits, depending on the concrete application. Throughout, we illustrate our findings within standard preference settings of the literature on sender-receiver games. Theses include settings with quadratic loss as in Crawford and Sobel (1982) or Melumad and Shibano (1991),[5] and others with state-independent sender

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

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

[4]We discuss Schelling's view on equilibrium multiplicities and the lessons that can be learned in our context in detail in Section 5.

[5]Such settings have also received attention in the information design literature, see, e.g., Galperti (2019); Jehiel (2015); Kamenica and Gentzkow (2011).

preferences. In some of these concrete settings, we characterize the equilibrium information structures in closed-form.

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

We contribute to the information design literature by introducing a novel consideration next to the standard persuasion motive. The sender trades off the motive to signal about a preference type and his persuasive influence on a receiver's action. This trade-off speaks to a wide array of applications. We present some of these and further discussion of the related applied literature in Section 4. In particular, the image concerns may be given a behavioural interpretation. This way, we relate the information design literature to a broader literature in behavioral economics and psychology, which has discussed a multitude of incarnations of image concerns: e.g. conformity Bernheim (1994), social reputation (Bénabou and Tirole, 2006), and self-image concerns (Bodner and Prelec, 2003). In this context, we note that the model may be interpreted as one of signaling from a sender-self to a receiver-self. Naturally, both selves share the same preferences over material outcomes. For such common value settings, we provide sharp welfare results: image concerns unambiguously reduce material payoffs. This relates to empirical evidence in social psychology. The literature on self-presentation (Schlenker, 2012) documents that individuals exhibit a wide array of behaviour that is factually bad for them but presumably useful for self-presentation.
There is a related literature on information design with signaling concerns, which has however either maintained that all senders share a common preference type (Hedlund, 2017) or that the pay-off relevant state is identical to the sender's type (Koessler and Skreta, 2021; Perez-Richet, 2014). Consequently, the existing work has not studied the trade-off between the motive to signal about a preference type and the persuasion motive, and the analysis in it relies on different techniques.

We contribute to the literature on signaling games. Most signaling games study situations in which a sender takes an action in order to influence a receiver's decision. His action may be influential since it may signal something about his private information to the receiver for whom this information is pay-off relevant (Cho and Sobel, 1990; Mailath, 1987; Ramey, 1996, see, e.g.,). In our model, the sender's action, his communication strategy, has a direct influence on the receiver's action, by revealing information about a state. It may also signal something about his private information to the receiver for whom this information is however

not pay-off relevant.[6]

This "separation" of the influence on the receiver's decision (persuasion motive) and the inference about the sender's type (signaling motive) is interesting in terms of new applications but also since there are non-trivial conceptual and theoretical implications.

If the influence on the receiver's decision channels through the information conveyed by signaling—like in most signaling games— in all separating equilibria, the (equilibrium) meaning of the message sent by a specific type is the same and unambiguous. So are the receiver decisions and welfare. In our setting, messages convey "factual" information about a state but also "personal" information about the sender's type. We show that the factual meaning of recommendations depends on the selected equilibrium. Only the personal meaning is pinned down uniquely but this information is not pay-off relevant for the receiver. This way, non-standard results about the receiver's welfare arise: e.g., it can be non-monotone in the sender's type despite an appropriate single-crossing condition, and it varies with the selected equilibrium.[7]

The paper's structure as follows: Section 2 presents the formal model. Section 3 contains all results about the equilibria and welfare, and applies the results to numerous standard preference settings. Section 4 formalizes two applications, relating to the initial examples: one on hierarchical organizations, and another on the disciplining scope of competitive elections for the instrumentalisation of media channels by politicians. Here, we also discuss insights for the literature on the use of incentives in organizations and that on electoral accountability. Section 5 contains further discussion and Section 6 concluding remarks.

2 Model

We study a communication game between a sender (she) and a receiver (he). There is a state space Ω, with a typical state denoted by ω, and an action space A, with a typical action denoted by a. Both A and Ω are compact metric spaces. The players are uncertain about the state at the outset of the game and share a prior $\mu_0 \in \text{int}(\Delta(\Omega))$. The sender moves first by choosing a *communication protocol*, which consists of two mappings: (i) an

[6]It is not pay-off relevant for the current decision of the receiver. Further, it need not be pay-off relevant for any receiver decision, including future ones.

[7]This second implication also depends on the sender having access to a rich set of communication strategies, not just a one-dimensional strategy space—as typical in information design.

information structure $\pi : \Omega \to \Delta(S)$, where $S = supp(\pi)$ is a compact metric space; (ii) an action-recommendation plan $r : S \to A$, which specifies, for each signal, a recommended action; $\mathcal{C}$ is the set of communication protocols. The receiver observes the sender's choice of communication protocol and the signal realization, and finally chooses an action $a \in A$ (which need not coincide with what is recommended by the sender).

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

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

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

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

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

Strategies and equilibrium. A pure strategy of the sender is a mapping $\sigma : \Theta \to \mathcal{C}$ that specifies for each type a communication protocol. A pure strategy of the receiver is a mapping that specifies an action for every possible communication protocol and signal realization. We analyze the perfect Bayesian equilibria in pure sender strategies (Fudenberg and Tirole, 1991, p. 333; henceforth equilibrium). In equilibrium, given the sender's choice of communication protocol and the subsequent signal realization, the receiver forms posteriors beliefs about the state using Bayes' rule, and then takes an action to maximize his expected payoff.[8] At the same time, the receiver may also update his beliefs about the sender's type. In other words, the sender's strategy influences not only the material outcome of the game, but also

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

her image in the eyes of the receiver.

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

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

where

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

That is, given the receiver's system of beliefs and the constraint that following the sender's recommendation is indeed optimal for the receiver, no sender type can be strictly better-off by deviating from the strategy σ.[9]

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 signalling games also arises in our model. We follow the literature and invoke a standard equilibrium refinement, the D1 criterion due to Cho and Kreps (1987) and Banks and Sobel (1987). The core idea is to restrict the receiver's out-of-equilibrium beliefs to the sender types that are "most likely" to benefit from the deviation to the off-path choice. Formally, given a sender strategy σ and an associated belief system of the receiver, we define for any $(\pi, r, \theta) \in \mathcal{C}^* \times \Theta$ the sets

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

[9]Note that restricting the sender's choice to the set $\mathcal{C}^* \subsetneq \mathcal{C}$ (i.e., the set of protocols with which it is indeed rational for the receiver to follow the sender's recommendation) is without loss of generality. This is because if $(\pi, r) \in \mathcal{C} \setminus \mathcal{C}^*$ and it induces the receiver to use some (sequentially rational) decision rule $\hat{r}(\cdot)$, we can always set $\eta(\pi, r) = \eta(\pi, \hat{r})$ to make sure that (π, r) is sub-optimal for the sender whenever (2) holds.

and

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

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

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

In words, condition (3) requires that if, for a type θ, there is another type θ' that has a strict incentive to deviate to the off-path choice $(\pi, r) \in \mathcal{C}^*$ whenever θ has a weak incentive to do so, then the receiver's out-of-equilibrium beliefs upon observing this choice of the sender shall not put any weight on θ.[10] An equilibrium that passes this test a *D1 equilibrium*; henceforth, often simply called equilibrium if no misunderstanding is possible.

3 Analysis

3.1 A Reduced-Form Characterization of Equilibria

Kamenica and Gentzkow (2011) analyze the benchmark scenario in which the sender does not have image concerns ($\phi = 0$), that is, she is purely guided by the persuasion motive. It is known that, even in that special setting, the equilibrium information structure is often intractable. This problem does not get any easier, if not more difficult, in our model, because the sender's persuasion motive may be entangled with her signalling 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 communication protocols directly, we focus on the expected material payoff that the sender obtains by the choice of a communication protocol. Similar "reduced-form approaches" have proven useful in a variety of mechanism design settings.[11]

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

[11] See e.g. the theory on reduced form auctions (Che, Kim and Mierendorff, 2013) and the literature on costly state verification (Ben-Porath, Dekel and Lipman, 2014).

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.[12] Second, the set of expected material payoffs that the sender can implement, i.e., $\mathcal{V} \equiv \{V \subset \mathbb{R} : \exists(\pi, r) \in \mathcal{C}^*$ such that $V = \mathbb{E}_\pi[v(r(s), \omega)]\}$, is a compact interval. To see this, let us define

$$\bar{V} \equiv \max_{(\pi,r)\in\mathcal{C}^*} \mathbb{E}_\pi[v(r(s),\omega)],\ \underline{V} \equiv \min_{(\pi,r)\in\mathcal{C}^*} \mathbb{E}_\pi[v(r(s),\omega)], \tag{4}$$

and let $(\bar{\pi}, \bar{r})$ and $(\underline{\pi}, \underline{r})$ be communication protocols that give rise to $\bar{V}$ and $\underline{V}$, respectively.[13] It is clear that $\mathcal{V} \subseteq [\underline{V}, \bar{V}]$. Since any $V \in [\underline{V}, \bar{V}]$ can be implemented by appropriately mixing the protocols $(\bar{\pi}, \bar{r})$ and $(\underline{\pi}, \underline{r})$, the reverse inequality $\mathcal{V} \supseteq [\underline{V}, \bar{V}]$ also holds.[14] Hence, $\mathcal{V} = [\underline{V}, \bar{V}]$. Third, as we formally show in the Appendix (Lemma A1), the sender's interim payoff has the following single-crossing property: for any $(V, \eta), (V', \eta') \in \mathcal{V} \times \Delta(\Theta)$ with $V < V'$, if type θ weakly prefers (V, η) over (V', η'), then (V, η) will be strictly preferred over (V', η') by all types $\theta' > \theta$.

The above properties allow us to apply techniques from the costly signalling literature (e.g. Cho and Sobel, 1990; Mailath, 1987; Ramey, 1996) to partially characterize the set of D1 equilibria. Given a sender strategy σ, we define $V(\theta; \sigma) \equiv \mathbb{E}_{\pi_\theta}[v(r_\theta(s), \omega)]$ and $p(\theta; \sigma) \equiv \mathbb{E}[\tilde{\theta}|\tilde{\theta} : \sigma(\tilde{\theta}) = \sigma(\theta)]$, i.e., the expected material payoff and the perceived image that the strategy induces for each type θ, respectively. We say that a type θ is *separating* under the strategy σ if $\sigma(\theta') \neq \sigma(\theta)$ for all $\theta' \neq \theta$ (in which case we necessarily have $p(\theta; \sigma) = \theta$). Otherwise, we say that θ is *pooling*. Our main result shows that there exists a *unique* cutoff $\hat{\theta}$ such that all types $\theta < \hat{\theta}$ ($\theta \geq \hat{\theta}$) will be separating (pooling) in any equilibrium that satisfies

[12] This 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)

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

[14] Take any solutions $(\bar{\pi}, \bar{r})$ and $(\underline{\pi}, \underline{r})$ of the sender's max- and minimization problems in (4), respectively. Without loss of generality, assume that $\bar{\pi}$ and $\underline{\pi}$ have no overlapping support (we can always relabel signal names). 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" communication protocol $(\hat{\pi}, \hat{r})$: with probability λ, the sender draws a signal s according to $\underline{\pi}$ and make recommendation according to $\underline{r}(\cdot)$, and with probability $1 - \lambda$, according to $\bar{\pi}$ and $\bar{r}(\cdot)$. It is straightforward to check that $(\hat{\pi}, \hat{r}) \in \mathcal{C}^*$ and $\mathbb{E}_{\hat{\pi}}[v(\hat{r}(s), \omega)] = V$, i.e., $(\hat{\pi}, \hat{r})$ indeed implements V.

D1.[15] 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 $\sigma = \{(\pi_\theta, r_\theta)\}_{\theta \in \Theta}$ of the sender with $(\pi_\theta, r_\theta) \in \mathcal{C}^*$ for all $\theta \in [0,1]$ is part of a D1 equilibrium if and only if the conditions (i) and (ii) are both satisfied:*

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

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

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

Theorem 1 implies that there exist D1 equilibria — one can always construct a strategy that satisfy (i) and (ii) (recall the property that $\mathcal{V} = [\underline{V}, \bar{V}]$). Exactly which strategy is chosen among the qualified ones is immaterial for the sender, because from her perspective all of them are equivalent in terms of payoffs. As a consequence, the set of D1 equilibria can be Pareto-ranked according to the welfare of the receiver. In later analysis, we will provide examples and applications which also feature payoff equivalence for the receiver, or which permit an analytical description of the equilibria that are maximal and minimal 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 the conditions (i) and (ii) will be derived (among others) from the on-path incentive compatibility constraints. With attention to detail, one can further construct the appropriate out-of-equilibrium beliefs that prevent off-path deviations and satisfy D1.

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

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

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

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

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

The intuition behind Lemma 2 is as follows. Given the single-crossing property of the sender's interim preferences, a higher type in a pooling set will be more likely to benefit from an off-path choice that slightly reduces her material payoff than any types lower than her. To be consistent with the D1 criterion, such an unexpected move must convince the receiver that the sender's type is higher than any one 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 it is not feasible. It is not feasible if the material payoff is already "used up" by the pooling types, i.e., the material payoff is minimal and given by $\underline{V}$.

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

costly signalling games where the sender faces a compact signal space. Our setup is quite different in that there is no exogenous cost structure on the set of available signals: *a priori*, it is costless for the sender to choose any communication protocol.

The cost of reputation. We now proceed to characterize the endogenous cost of signalling (i.e., the extent to which one's material payoff need to be sacrificed) for types in the separating interval $[0, \hat{\theta})$. Here, the central idea is to leverage that the sender's utility function is quasi-linear with respect to her reputation. This payoff structure reminds of the standard mechanism design setting with transfers. Thus, naturally, we advance the analysis by applying a classical envelope theorem argument to the (local) incentive compatibility constraints of the sender.[16]

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

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

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

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

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

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

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

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

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

Further, whenever $\hat{\theta} > 0$, the type $\theta = 0$ is in the separating interval and getting the lowest possible image payoff. Thus, incentive compatibility also requires that this type must

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

be earning the highest possible material payoff, i.e., $V(0;\sigma) = \bar{V}$. Combining this boundary condition and the differentiable equation (8), we immediately obtain the payoff formula (5) and conclude that it must hold for all $\theta \in [0, \hat{\theta})$ in any D1 equilibrium.

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

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

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

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

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

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

for all $\theta \in [0, 1]$. Note that I is strictly increasing;[17] I is continuous in $\hat{\theta}$ because $w(\cdot)$ is continuously differentiable and because the type distribution Γ is absolutely continuous.

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

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

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

First, if

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

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

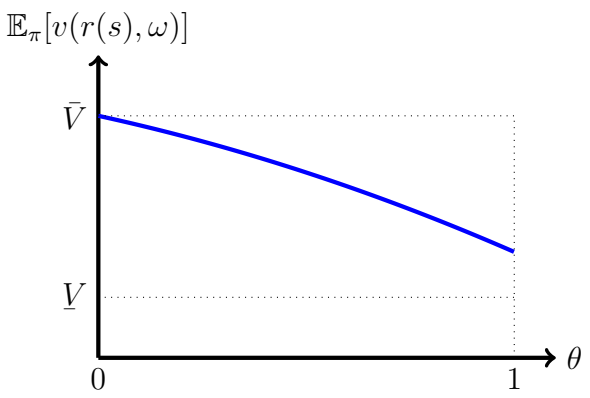
Second, consider the case $(\bar{V} - \underline{V})/\phi \geq I(1)$ or, equivalently, $\phi \leq \underline{\phi} \equiv (\bar{V} - \underline{V})/I(1)$.. Suppose that there would be an equilibrium with cutoff $\hat{\theta} < 1$. Then, all types $\theta < 1$ would strictly prefer separating to pooling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

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

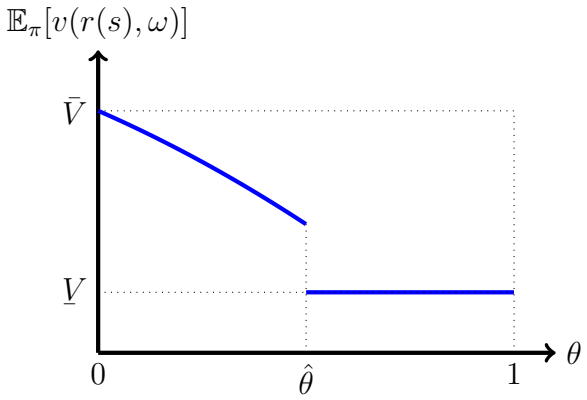
We close this section with a graphical illustration of the main findings of Theorem 1. Figure 1 presents all three types of the sender's strategy that could emerge in an equilibrium satisfying the D1 criterion. That is, the cases $\phi \leq \bar{\phi}$ (Panel a), $\underline{\phi} < \phi < \bar{\phi}$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

3.2 Equilibrium Multiplicity and Pareto (In)efficiency

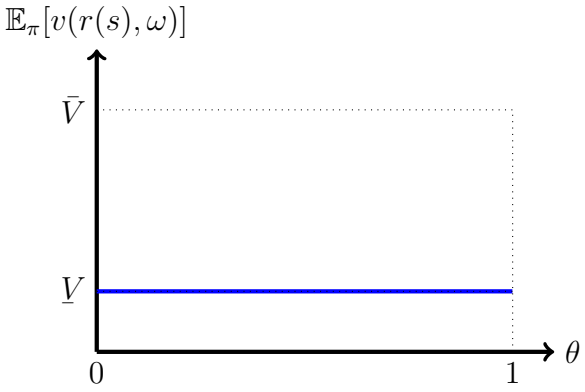
As mentioned, Theorem 1 implies that the sender's interim payoffs are equivalent across all D1 equilibria. In particular, the theorem specifies explicitly the level of material payoff that each sender type will give up in order to separate herself from lower types. Given the abundance of possible information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



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



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



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

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

receiver's welfare will (or will not) be robust to equilibrium selection, and (ii) analyzing the Pareto-frontier of the equilibrium set.

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

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

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

identifying what the best- and the worst-case scenarios for the receiver may look like within the set of D1 equilibria.

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

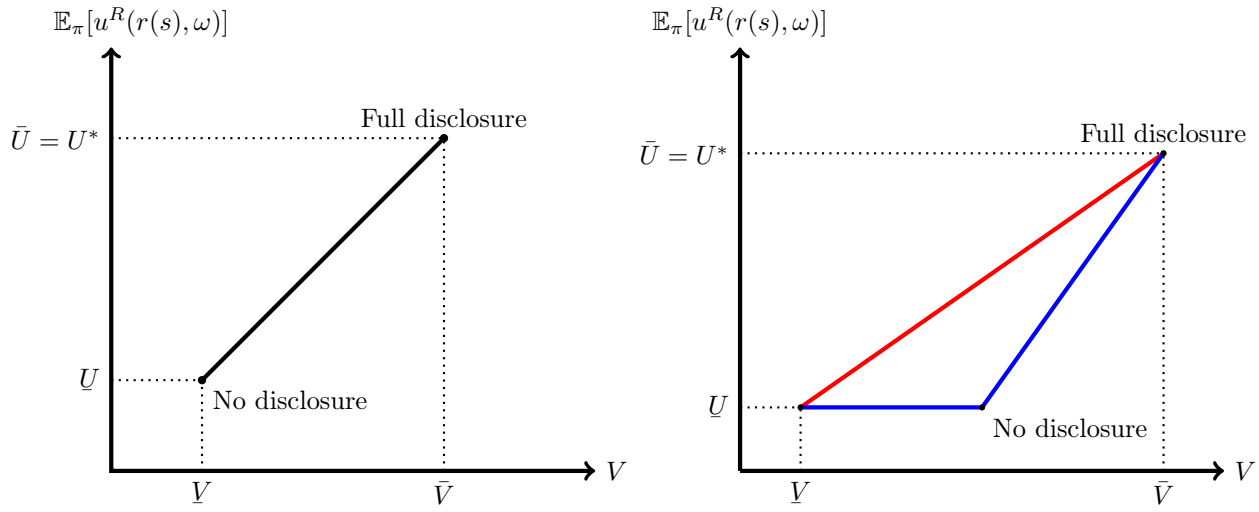
(i) there exists a D1 equilibrium in which the receiver is strictly worse-off compared to the case without image concerns;

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

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

Intuitively, the conditions of Theorem 2 imply that a no-disclosure protocol is suboptimal for the sender when she is purely guided by material interests, since otherwise the receiver would not have been able to enjoy his full-information payoff. Therefore, an image-concerned sender can always separate herself from those very low types by occasionally sending a completely uninformative signal to the receiver, which obviously engenders a negative "side-effect" on the receiver's payoff. As for the properties of the Pareto-extremal equilibria, our proof mainly exploits the convexity of the set of payoff profiles that can be implemented via information design: For instance, if, within the separating interval of a D1 equilibrium, the receiver's payoff implied by the strategy of a type θ is lower than the payoff implied by the strategy of a higher type $\theta' > \theta$, then this equilibrium cannot be Pareto-optimal, as follows. Replacing the communication protocol that type θ initially chooses with an appropriate mix of those used by types 0 and θ' will not change the sender's payoff, but will strictly improve the payoff of the receiver. A similar constructive argument (which involves the no-disclosure protocol instead of the one used by type 0) shows that any Pareto-worst D1 equilibrium must be either decreasing or U-shaped with respect to the sender's type. Otherwise, it would have been feasible to further reduce the receiver's payoff without altering the sender's.

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



(a) Common interest

(b) Almost common interest

Figure 2: The equilibrium set of implementable payoffs in settings with $U^* = \bar{U}$. In panel (a), we have a common-interest game with $u^R(a,\omega) = v(a,\omega)$, $\bar{V} = 0.5$ and $\underline{V} = 0.1$. In panel (b), we have an almost-common-interest game as described in Example 3. There, the red curve depicts the utility-frontier of the Pareto-optimal D1 equilibria, while the blue curve corresponds to the Pareto-worst D1 equilibria.

Example 1: Common-interest games. Suppose that there exists a strictly *increasing* function $\Psi : \mathbb{R} \to \mathbb{R}$, such that $u^R(a,\omega) = \Psi(v(a,\omega))$ for all $(a,\omega) \in A \times \Omega$. In such settings, the material interests of the players are perfectly aligned. Thus, plainly, the material payoff of the sender is maximized if he provides full information to the receiver (and is minimized if she only babbles). Hence, we have $U^* = \bar{U} = \Psi(\bar{V})$ and, whenever $\bar{V} > \underline{V}$ and $\phi \leq \bar{\phi}$ hold, Theorem 2 applies.

Panel (a) in Figure 2 depicts the set of implementable material payoff profiles in a common-interest game, in which $\Psi(\cdot)$ is a linear function so that the mapping between the *expected payoffs* of the two players is also a linear one. Consequently, the D1 equilibria in this game are not only payoff-equivalent to the sender (as already asserted by Theorem 1), but also to the receiver. A particularly simple D1 equilibrium is one in which the sender always commits to a protocol that reveals either everything or nothing about the true state, with the frequency of the former action decreasing in the sender's type. In this case, a more image-concerned sender (captured by either a higher θ or ϕ) will transmit less information to the receiver, therefore intensifying the negative impact on the latter's welfare.

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

In Appendix A.6, 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 $v(a,\omega) = -(a-\omega)^2$ and the image payoff function $\hat{w}(p(\eta),\theta) = w(p(\eta),\theta)/(2f(\theta)-1)$. This transformation manifests that the players' interests are sufficiently aligned under the current specification, insomuch that a sender purely guided by material interests would be willing to share all information with the receiver. However, if function $\hat{w}(\cdot)$ satisfies the key condition (1) – which can be the case, for instance, if $f'(\cdot) < 0$, meaning that higher types put less weight on the state-dependent term relative to the state-independent target $g(\cdot)$— then both Theorem 1 and Theorem 2 apply. They jointly imply that all types (except possibly type 0) will withhold information from the receiver for signalling purposes. Moreover, given that $u^R(a,\omega) = v(a,\omega)$, the equilibrium payoffs of both the sender and the receiver are uniquely pinned down by the D1 criterion.

Example 3: An almost-common-interest game. Next, we provide an example in which the receiver's equilibrium payoff is indeterminate. Yet, we are able to fully characterize the Pareto-frontier of the equilibrium set. Let $A = \Omega = \{-1,0,1\}$. The material payoff functions

of the players are:

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

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

Since both players' material payoffs are uniquely maximized under full disclosure ($\bar{V} = \bar{U} = U^* = 1$), Theorem 2 is applicable here. Now, assume further that the prior distribution of the state is given by $\Pr(\omega = 1) = \Pr(\omega = 0) = 0.4$ and $\Pr(\omega = -1) = 0.2$. As we show in Appendix A.7, while no disclosure minimizes the receiver's payoff ($\underline{U} = 0.4$), it does not minimize the sender's. Instead, the sender achieves her minimal material payoff by recommending $a = -1$ with probability 1 in state $\omega = -1$, while recommending $a = 1$ or $a = -1$ with equal probabilities in the other two states. Note that this communication protocol results in the payoffs $\underline{V} = 0$ and $\underline{U} = 0.4$ for the sender and the receiver, respectively. Based on the above observations, we identify the entire set of implementable material payoff profiles. Panel (b) in Figure 2 provides a graphical representation of this set, where the red (blue) curve pins down, for any given level of sender's payoff $V \in [\underline{V}, \bar{V}]$, the maximal (minimal) payoff that the receiver can obtain. Thus, in any Pareto-extremal D1 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.

Theorem 2 and the examples following it relate to the literature on "bad reputation" in dynamic games (see, e.g., Ely, Fudenberg and Levine, 2009; Ely and Välimäki, 2003). An overarching finding of this literature is that reputational concerns are harmful if they are based on a desire to separate from a bad type rather than to mimick a good commitment type (see the discussion in Mailath, Samuelson *et al.*, 2006). The forces behind our results are quite different and more subtle: the sender tries to separate herself from the type that is *least image-concerned*, which requires her to avoid taking the strategy that would be *endogenously*

chosen by the latter. In the current set-up, that strategy happens to be the one that maximizes the material payoffs of both players.

3.2.2 When will sender's image concerns be beneficial?

We now study when the presence of image concerns can only be beneficial. Analogous to the previous subsection, we focus on settings where 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 sender's image concerns. Moreover, if $\phi < \bar{\phi}$ (so that we have a cutoff type $\hat{\theta} > 0$), then*

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

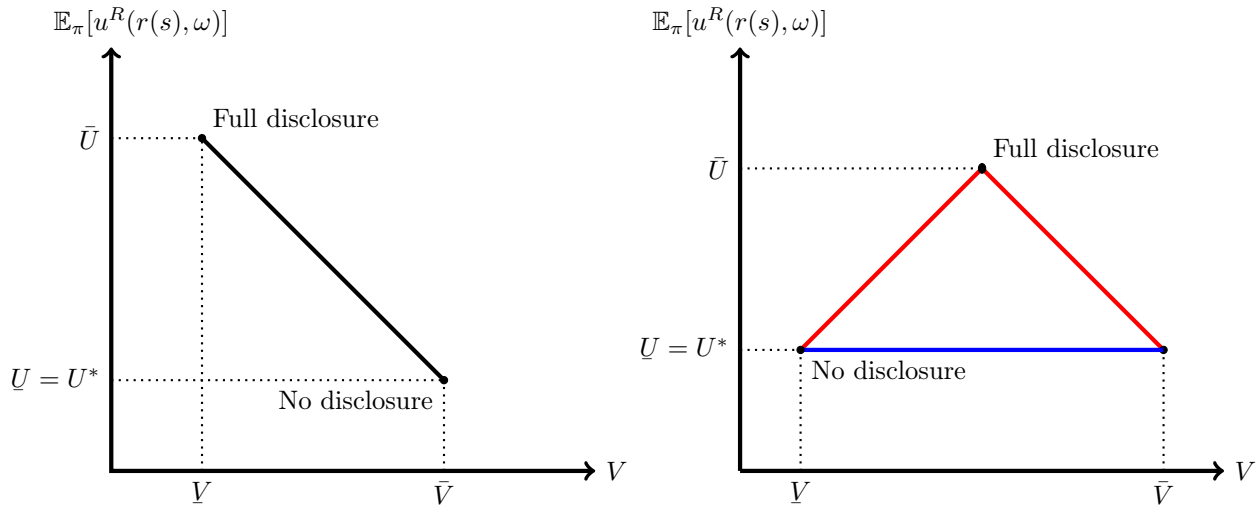
(ii) in any Pareto-optimal D1 equilibrium, the receiver's expected payoff is quasi-concave with respect to the sender's type θ on the separating interval $[0, \hat{\theta})$.

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

Both the proof and the intuition of Theorem 3 are analogous to Theorem 2.

Example 4: Conflicting-interest games. Suppose that there exists a strictly *decreasing* function $\Psi : \mathbb{R} \to \mathbb{R}$, such that $u^R(a, \omega) = \Psi(v(a, \omega))$ for all $(a, \omega) \in A \times \Omega$. In such settings, the players have exactly opposite material interests. It is clear that the material payoff of the sender is maximized if she provide no information to the receiver (and is minimized if she provides full information). Hence, we have $U^* = \underline{U} = \Psi(\bar{V})$ and Theorem 3 applies provided that $\bar{V} > \underline{V}$ and $\phi \leq \bar{\phi}$ additionally hold.

Panel (a) in Figure 3 depicts the set of implementable material payoff profiles in a game with conflicting interests. Similar to Example 1, the function $\Phi(\cdot)$ is chosen to be linear, which gives rise to the linear mapping between the players' expected payoffs as we see from the figure. This property ensures that all D1 equilibria are payoff-equivalent to both the



(a) Conflicting interest

(b) Partially conflicting interest

Figure 3: The equilibrium set of implementable payoffs in settings with $U^* = \underline{U}$. In panel (a), we have a completely conflicted-interest game with $u^R(a,\omega) = -v(a,\omega)$, $\bar{V} = 0.5$ and $\underline{V} = 0.1$. In panel (b), we have partially conflicted-interest game as described in Example 5. There, the red curve depicts the utility-frontier of the Pareto-optimal D1 equilibria, while the blue curve corresponds to the Pareto-worst D1 equilibria.

sender and the receiver. A particularly simple D1 equilibrium is one in which the sender always commits to a protocol that reveals either everything or nothing about the true state, with the frequency of the former action increasing in the sender's type. In this case, a more image-concerned sender (captured by either a higher θ or ϕ) will transmit more information to the receiver, therefore intensifying the positive impact on the latter's welfare.

Example 5: Quadratic-loss games (continued). Consider again the quadratic-loss games that we introduced in the previous subsection. In Appendix A.6, we show that under the condition $f(\theta) < 0.5$ $\theta \in [0,1]$, the initial game will be strategically equivalent to one in which the sender has the material function $v(a,\omega) = (a-\omega)^2$ and the image payoff function $\hat{w}(p(\eta),\theta) = \mathbb{E}_\eta[\tilde{\theta}]/(1-2f(\theta))$. Thus $v = -u^R$ and we effectively have a game with conflicting interests. Provided that condition (1) holds for $\hat{w}(\cdot)$ – which can be the case, for instance, if the interests of higher types are more aligned with the receiver in the sense that $f'(\theta) > 0$ $\forall\theta \in [0,1]$ – then both Theorem 1 and Theorem 3 apply. Thus, the presence of image concerns will trigger all sender types (except possibly type 0) to share information with the receiver, which they would be reluctant to do otherwise. In addition, because $u^R(a,\omega) = -v(a,\omega)$, the payoff-equivalence implied by the D1 criterion holds not only for the sender, but also for

the receiver.

Next, we use two widely-studied examples to demonstrate that the applicability of Theorem 3 is *not* limited to settings in which the sender would prefer not to share any information out if only the persuasion motive is present. This is an important observation, because it is known that partial disclosure is optimal in many settings without image concerns. For instance, Jehiel (2015) shows that this is typically the case when the information of the sender is higher dimensional than the action space of the receiver; Kolotilin and Wolitzky (2020) and Kolotilin, Corrao and Wolitzky (2022a) provide similar results in a setting that allows utilities of the sender and receiver to be non-linear in the state.[18] The optimality of partial disclosure may be compatible with the premise of Theorem 3, $U^* = \underline{U}$, because the access to additional information does not guarantee that the receiver can do *strictly* better than taking his prior-optimal action *on average*.

Example 6: State-independent sender preferences, I. Suppose that $A = \Omega = \{0, 1\}$, $v(a, \omega) = a$ and $u^R(a, \omega) = \mathbb{1}_{a=\omega}$. Thus, while the receiver wants to match the state, the sender's preference over material outcomes is state-independent: she always prefers the receiver to takes the high action. This persuasion setting is most vividly embodied by the prosecutor-judge example in Kamenica and Gentzkow (2011).

Since the state space is binary, we use μ_0 to denote the prior likelihood of the state being $\omega = 1$. We assume $\mu_0 \in (0, 0.5)$ so that $a = 0$ is the receiver's optimal action given the prior. Clearly, releasing no information minimizes the sender's material payoff. At the same time, Kamenica and Gentzkow (2011) show that partial information disclosure is optimal for the sender when she has no image concerns. However, the receiver's expected payoff under optimal partial disclosure is the same as under no information (i.e., $U^* = \underline{U}$).[19] Hence, all results of Theorems 1 and 3 apply.

We present two simple classes of communication protocols 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

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

[19] This is because the optimal disclosure policy uses a binary signal and after both signal realizations, the sender weakly prefers the prior-optimal action.

signal $s = 1$ is drawn with probability

$$\bar{\rho}(\omega; q) = \begin{cases} \min\left\{\frac{q}{\mu_0}, 1\right\} & \text{if } \omega = 1, \\ \max\left\{\frac{q-\mu_0}{1-\mu_0}, 0\right\} & \text{if } \omega = 0. \end{cases} \tag{12}$$

With the remaining probability $1 - \bar{\rho}(\omega; q)$, the signal $s = 0$ is sent to the receiver. Together with the action-recommendation plan $r(s) = s \; \forall s \in A$, $\bar{\pi}^q$ induces the receiver to choose the action $a = 1$ with exactly probability q. While there can be other information structures that induce the same (unconditional) distribution of actions, all of them will be Pareto-dominated by $\bar{\pi}^q$ (see Appendix A.7 for a formal proof). For instance, consider the information structure $\underline{\pi}^q$ defined as follows: Conditional on the true state, the signal $s = 1$ is drawn with probability

$$\underline{\rho}(\omega; q) = \begin{cases} \frac{q}{2\mu_0} & \text{if } \omega = 1, \\ \frac{q}{2(1-\mu_0)} & \text{if } \omega = 0. \end{cases} \tag{13}$$

With the remaining probability $1 - \underline{\rho}(\omega; q)$, the signal $s = 0$ is sent to the receiver. With this information structure, the sender can also nudge the receiver to choose the high action with probability q. However, the probability that the receiver takes the *right* action is just $1 - \mu_0$ under $\underline{\pi}^q$, which he could also achieve by simply sticking to his prior-optimal action $a = 0$. This is the worst possible outcome for the receiver, so he would clearly prefer $\bar{\pi}^q$ over $\underline{\pi}^q$. In fact, $\bar{\pi}^q$ is part of a Pareto-optimal equilibrium, and $\underline{\pi}^q$ is part of a Pareto-worst equilibrium, as we show in the Appendix. Panel (b) in Figure 3 depicts the receiver weilfare in both equilibria, delineating the whole set of implementable payoff profiles for the receiver.

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

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 when there are no image concerns. Our last formal result confirms that in this case, the effect of image concerns will be ambiguous.

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 image concerns or if he is harmed by it. In particular, if ϕ is sufficiently small, then there always co-exist (i) a D1 equilibrium in which the receiver is strictly better-off and (ii) a D1 equilibrium in which the receiver is strictly worse-off relative to the setting without image concerns.*

We close this section with an example that illustrates the findings of Theorem 4.

Example 7: A partially-conflicting-interest game. Let $A = \Omega = \{-1, 0, 1\}$. The material payoff functions of the papers are

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

Intuitively, the players' interests are perfectly aligned in state $\omega \in \{0, 1\}$. However, there is a conflict of interest when the state is $\omega = -1$. If $\omega = -1$, the receiver prefers action $a = -1$ but this is the worst action from the sender's point of view.

Figure 4 illustrates the Pareto-frontier of this example for the case when the prior distribution of ω is specified by $\Pr(\omega = 1) = \Pr(\omega = 0) = 0.4$ and $\Pr(\omega = -1) = 0.2$.

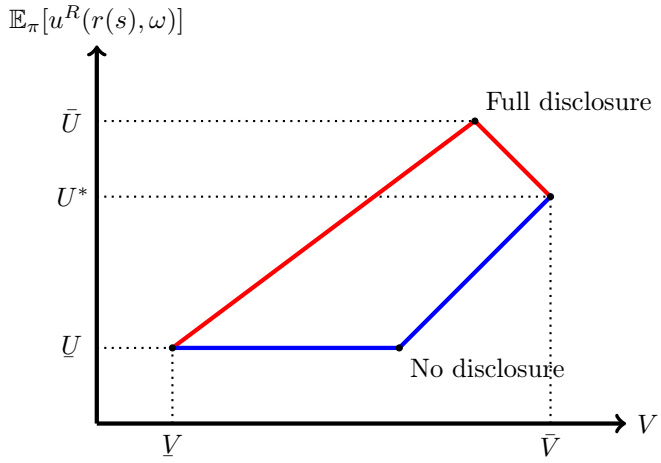


Figure 4: The equilibrium set of implementable payoffs in Example 8, where $U^* \in (\underline{U}, \bar{U})$. The red line depicts the utility-frontier of the Pareto-optimal D1 equilibria, while the blue line corresponds to the Pareto-worst D1 equilibria.

4 Applications

4.1 On Harmful Signaling in Hierarchical Organisations

We provide a model of hierarchical organizations by specializing Example 2 from Section 3.2 through specific choices of the sender's bliss point $a^*(\omega, \theta) = f(\theta)\omega + g(\theta)$ and image concerns $w(p, \theta)$. The model studies the effects of top-down incentive schemes on operational performance. There are three agents: senior management, a mid-manager (the sender), and a subordinate (the receiver).

The hierarchical structure manifests in two ways. First, senior management only observes the mid-manager's behaviour but not the subordinate's actions further down the hierarchy. Second, there is a "chained" moral hazard problem. There is moral hazard on the first hierarchical layer: the subordinate performs operational tasks and prefers to provide systematically lower effort levels than her mid-manager would like her to. There is moral hazard on the second hierarchical layer: The mid-manager can manage the subordinate by ways of communication and information control, thereby affecting her effort.[20] Senior management has issued the guideline that mid-managers should motivate their subordinates to higher effort. However, depending on his type $\theta \in [0, 1]$, the mid-manager has internalized the senior man-

[20]We focus on this management tool. In other situations, it might be reasonable to assume that senior management provides additional monetary funds that can be used to further align the subordinates incentives. However, oftentimes the very function of the direct manager of employees is to increase employee motivation, given a fixed salary scheme.

agement's preferences regarding the subordinate's effort fully or only partially.

Precisely, a state $\omega \in [0,1]$ captures information about the requirements of the subordinates tasks. The higher the state, the more complex the task; thus, requiring more effort. From the point of view of senior management, the optimal effort level is $B \cdot \omega$, for $B > 1$. The subordinates effort bliss point is systematically lower, and given by ω . Mid-managers with more career concerns have internalized the senior management's point of view more. Precisely, the mid-manager's preferences extrapolate between the subordinate and senior management preferences, with a bliss point $f(\theta)\omega \in [1, B]$. The function f is increasing from $f(0) = 1$ to $f(1) = B$. Formally, we set $f(\theta) = (1-\theta) + B\theta = 1 + (B-1)\theta$ for some $B > 1$, and $g(\theta) = 0$.

Consider the interaction between senior management and the mid-manager. Senior management evaluates and rewards the mid-manager internalizing the guideline. Formally, the incentives given by the evaluation translate into image concerns of $w(p, \theta) = p\theta$ for the mid-manager, where $p = \mathrm{E}(\tilde{\theta}|\pi)$.[21] This captures that higher types care more about the evaluation, and squares with the feature that higher types internalize the company guidelines more strongly. It implies that all types (except $\theta = 0$) would like to signal that they have fully internalized the guidelines.

The mid-manager can censor information about the state. This means that he can decide which states to reveal to the subordinate by sending the message $s = \omega$ in ω and which states not to reveal by sending the message $s = \emptyset$ in ω, and this may affect the subordinates effort.[22]

Results. How do the incentives given by senior management translate into behavior of the mid-manager and the subordinate subsequently? Clearly, a manager type who has not internalized the management guideline at all (the type $\theta = 0$ with $f(\theta) = 1$) would fully communicate all information about the state to the subordinate. Managers of higher types will consequently involve in strategies that hide information from subordinates to separate and signal to senior management that they care about their policy. Our results for Example 2 in Section 3.2 imply that this signaling behaviour is harmful to the organizational performance.

[21] Note that the sender preferences have the increasing differences property since $\frac{\theta}{2f(\theta)-1} = \frac{\theta}{1+2(B-1)\theta}$ is increasing in θ).

[22] We make the restriction to censorship communication to be more concrete for the reader. Further, this way, it becomes obvious that the information design game is equivalent to one in which the sender can disclose or withhold hard information. A similar equivalence is derived in Gentzkow and Kamenica (2017). Here, the restriction is without loss. This is for two reasons: first, the whole set of interim payoffs $V \in [\underline{V}, \overline{V}]$ can be implemented with censorship communication. Second, when allowing for all communication protocols, all equilibria will be payoff-equivalent in this application. Both observations together imply that the set of equilibrium outcomes is the same with and without restriction.

This is because without signaling motive, *all* agents would prefer full information disclosure to maximize their material payoffs; for details, see Appendix A.6. To conclude, the hierarchical structure and the incentive scheme of the organization cause harmful upward signaling by the mid-manager.

Related Literature. This application contributes to the literature on incentives in organisations and career concerns Holmström (1999) by providing a simple model of a hierarchical organization.[23]

We show that the hierarchical structure is so that top-down incentives, explicit or implicit, backfire and reduce organizational performance. They create harmful signaling up the hierarchy of the organization. This result relates to recent debates how performance in organizations may suffer from "pleasing your boss"-schemes (see, e.g., Dillon, 2017). To avoid such issues, many companies nowadays conduct performance evaluation not solely by direct superiors but rely on committees.[24] More generally, people have debated downsides of the traditional hierarchical organization in which attention will naturally be directed up the hierarchy; see the diverse theories of non-hierarchical leadership such as flat organizations, or agile practices (Beck, Beedle and Bennekum, 2001).

4.2 Electoral Accountability and Media Influence

We provide a model of media influence and electoral accountability by specializing Example 6 from Section 3.2. The model studies the welfare effects of a politician's influence on politics through media channels in a simple framework of electoral accountability. There is a politician (sender) and a mass 1 of citizens (receiver). In a first stage, the politician communicates about a binary state $\omega \in \{-1, 1\}$ that is pay-off relevant for a political choice $a \in \{-1, 1\}$.

The politician may be incentivized by a third party to use his communication power to move the political choice towards $a = 1$. The third-party incentive is the politician‘s private information. His private type $\theta \in [0, 1]$ measures how "corrupted" the politician is, with $\theta = 1$ being not corrupted at all and lower types being more corrupted. Formally, the sender gets a state-independent payoff of $u(1, \omega) = w_1(\theta)$ when $a = 1$ is chosen and otherwise not, with

[23] See Georgiadis (2022) for a recent review of research on moral hazard, in particular, in organizations.

[24] 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.

$w_1(1) = 0$ and $w_1' < 0$. The receiver is a reduced-form representation of the interests of the "common citizen"; she obtains a utility of 1 if the policy matches the state and otherwise a utility of zero. Upon receiving a signal from the politician, the receiver chooses one of the two policies. At the common prior $\Pr(\omega) < \frac{1}{2}$, the receiver prefers $a = -1$.

The game that only consists of the first stage serves as a benchmark of the persuasive influence of the politician on citizen welfare. This allows to study to which extent a future competitive election may discipline the rhetoric of politicians with third-party incentives.

In the second stage, the politician may be elected into office by the citizens. Citizens are rational in the sense that (on average) they anticipate that more corrupt politicians will act less in the interests of the citizens. Further, they make a Bayesian inference from the communication strategies about the politician's type. Formally, the citizen's payoff from a politician of type θ in office is $\psi(\theta + \epsilon)$ where $\epsilon \in [-1, 1]$ is an idiosyncratic parameter that is drawn independently across citizens, and where ψ varies the pay-off relevance of the election. Thus, a citizen votes for the politician if $p + \epsilon \geq 0$ for $p = \mathrm{E}(\theta|\pi)$. This means that the likelihood of the politician being elected is $G(-p)$ where G is the cumulative distribution function of ϵ. The politician's payoffs from being appointed to office are subsumed into one utility function $\psi w_2(\theta)$.[25] To conclude, the election of the second stage creates incentives for the politician to signal that he is not corrupt because this increases his likelihood of being elected.

The equilibrium of this dynamic game can fully be described by the communication strategy of the politician in stage 1, where each politician type θ maximizes

$$\max_{(\pi,r)} \Pr(r(s) = 1|(\pi, r))w_1(\theta) + w_2(\theta)G(-p)$$

for $p = \mathrm{E}(\theta|(\pi, r))$. This is equivalent to maximizing

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

so that the equilibrium problem maps into our setting by specizializing Example 4 of Section 3.2 with $w(p, \theta) = \frac{w_1(\theta)}{w_2(\theta)}G(-p)$, $u^R(a, \omega) = 1_{x=\omega}$ and $v(a, \omega) = 1_{a=1}$.

[25] These utilities may derive from an office- or a policy-motivation but we remain agnostic about further details.

Populist speech. Without loss, all politician types send a binary recommendation, either to choose policy $a = 1$ or policy $a = -1$. Recall that $V(\theta)$ is decreasing in θ, in equilibrium. Since $V(\theta) = \Pr(r(s) = 1|\pi(\theta))$, this implies that the frequency of policy $a = 1$ being recommended decreases in the type. Putting things together, we see that, in equilibrium, the belief $\mathrm{E}(\theta|\pi)$ is higher when $a = -1$ is recommended less often given π. Motivated by this observation, we interpret the recommendation of policy $a = -1$ as "populist speech". It increases the belief in the alignment of the politician's preferences with the common citizen. This follows the standard definition in political science and political philosophy that defines populism as a political strategy supporting the people in their struggle against the privileged elite; see the American heritage dictionary).[26]

Results. The results from Section 3.2 apply when $\frac{w_2(\theta)'}{w_1(\theta)} > 0$. That is, when less corrupt types have higher incentives to being elected. Note that this includes the scenario in which all types are purely office-motivated, gaining a constant payoff $W = w_2(\theta)$ from being elected, which is the focus of much of the literature on electoral competition.[27]

In this case, the second stage election incentives are effective:[28] They create the desire to impress the voters. As a consequence, higher incentives, as measured by ψ, cause more populist speech (in the sense that all types recommend the policy $a = -1$ more often). The main insight is that this disciplining effect on the politician's speech need not translate monotonically into citizen welfare. Precisely, the results from Section 3.2 show that the citizen's first stage welfare is *non-monotone* in ψ: When ψ is sufficiently small, welfare increases in ψ in the Pareto optimal equilibrium; compare to panel (a) of Figure 1. However, when ψ is sufficiently large, the election incentives "over-discipline" the politician. Then, all types pool and send the "populist" recommendation $a = -1$ with probability 1 in all equilibria; compare to panel (c) of Figure 1.

Related literature This application contributes to the literature on electoral accountability, surveyed thoroughly by Ashworth (2012) and Duggan and Martinelli (2017). The novelty is to analyze the effect of election incentives on a politician's persuasive influence through

[26] See also Mudde (2004), Acemoglu *et al.* (2013), and the recent Vox debate on populism.

[27] See, for example, the book by Persson and Tabellini (2002).

[28] One can show that, when more corrupt types have higher incentives to be elected, that is when $\frac{w_2(\theta)'}{u_w(\theta)} < 0$, then, the election incentives have no effect: the unique equilibrium is so that all types play the equilibrium strategy of the benchmark game with only the first stage.

media channels. We bring forward the insight that the incentives of competitive elections may have non-monotone effects for citizen welfare. On the one hand, incentives can induce more honesty of politicians. In the extreme, however, very high incentives enforce a conformity of the political debate.

The application also contributes to an emerging literature on populism.[29] Our analysis provides an example, in which the implications of more or less populist behaviour are non-trivial. Here, the extent of populist speech (recommendations $a = -1$) increases monotonically with the election incentives. Thus, parallel to the observation that the election incentives have a non-monotone effect on receiver welfare, when political communication becomes more populist, the comparative statics can go both ways. This more broadly relates to Rodrik's point that populism has "many forms" and, as such, sometimes welfare-reducing but also welfare-enhancing consequences.[30]

5 Remarks

5.1 Optimal Information Structures

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 D1 equilibria. The reason is that the abundance of possible information structures allows diverse choices that lead to the same payoff for the sender. Standard refinements are "pay-off based".[31] Thus, intuitively, they do not have any bite in selecting across different choices that are pay-off equivalent to the sender.

We view this multiplicity as a qualification of the information design approach rather than as a drawback. As Schelling (1980) argues, we may view this type of equilibrium multiplicity as a manifestation of different cultures of communication. Or, as Roger Myerson writes, "recognizing the "social" problem of selecting among equilibria can help us to better understand the economic impact of culture."[32]

For one thing, we may learn from observed payoff allocations about external "cultural"

[29] See, e.g. Acemoglu *et al.* (2013); Bernhardt, Krasa and Shadmehr (forthcoming); Guiso, Herrera, Morelli, Sonno *et al.* (2017); Morelli, Nicolò and Roberti (2021), and the recent Vox debate on populism.

[30] See https://new.cepr.org/voxeu/columns/many-forms-populism.

[31] E.g., the D1 criterion rules out that off-path beliefs put mass on types that have lower payoff incentives to deviate to a given off-path action.

[32] See Myerson (2009).

factors that may drive which equilibrium is selected.

Reversely, 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, manipulation of data is constrained by monitoring efforts or plausibility constraints. As a consequence, outright fabrication of new data is infeasible. However, partial omissions and deletion may not be detected easily and feasible. In such settings, one should assume that the sender is restricted to such *censoring* of the information, and consider the corresponding equilibrium. We have exemplified this point within the application in Section 4, in which a mid-manager of a firm can censor the information flow to subordinate employees.

Similarly, our characterization results can be used to study the payoff implications for the receiver when varying the details of a given application. For example, we may ask about the effect of more or less effective monitoring technologies. As discussed, such monitoring may restrict the feasible information structures. These restrictions may yield a prediction by selecting an equilibrium or, further, by constraining the feasible set of the sender's interim payoffs $[\underline{V}, \bar{V}]$. Note here, that versions of our results, given such constraints, are obtained in a straightforward fashion, following the logic layed out in this paper.

Last, this paper answers to 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 *any* information structure. The critique is that optimal information structures sometimes may be infeasible or difficult to implement. This may be because the details of a given application prohibit the use of certain information structures. The information design literature has addressed this issue and identified sufficient conditions for simple information structures to be optimal among all information structures, or, in other work, explicitly incorporated constraints (for a survey of this strand of the literature, see Kamenica, Kim and Zapechelnyuk, 2021). In our setting, if a specific application qualifies a class of simple information structures—e.g. the censoring of available information—this class is consistent with equilibrium (in the unrestricted game) under the weak condition that it can fully implement all possible material payoffs of the sender.

6 Conclusion

In many economic situations, a sender strategically communicates with a receiver with a two-fold goal. First, to influence the decision-making of the receiver. Second, to steer the perception of certain unobserved characteristics of herself, e.g., loyalty, integrity, or unselfishness.

We have provided a model to study the trade-off arising in such situations. The model builds upon the canonical, and general, framework of information design by Kamenica and Gentzkow (2011). Unlike in the standard framework, next to the persuasion motive, the sender has a second motive, namely to signal about her own type. We have characterized the equilibria in this general framework. Our analysis built on previous work of the costly signaling literature but yields several non-standard observations. These include non-monotonicities in the sender behaviour (despite working in a setting with an appropriate single-crossing condition), and a multiplicity of equilibria that survive all standard refinements.

The image concerns can take various interpretations, making this framework a versatile model. First, the sender may be concerned about how his communication reflects on him because these inferences may impress a third party. We have exemplified this within a model of a hierarchical organisation in which mid-managers try to impress their superiors, thereby blundering the organizational performance. This application highlights a potential downside of career concerns and incentives in organisations.

Second, the sender may also worry about her image out of reputational concerns: image may matter for a continuation game. We have exemplified this within a novel model of media influence by politicians and electoral accountability. This application highlights that future elections may be able to discipline the intstrumentalization of media through politicians but effects of such election incentives are non-monotone, unlike in standards results in the literature on electoral accountability (Ashworth, 2012).

Third, the image concerns may be given a behavioural interpretation. This way, our model relates the literature on information design to a broader literature on economics and psychology which has discussed diverse forms of image concerns. We discussed some relations, e.g. how our framework is consistent with the empirical work on self-presentation in social psychology (Schlenker, 2012).

Appendix

A.1 The Single-Crossing Property

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

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

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

Proof. Indifference of type θ means

$$V - V' = \phi \cdot [w(p(\eta'), \theta) - w(p(\eta), \theta)]\,. \tag{14}$$

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 (14) implies

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

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

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

for all $\theta' > \theta$. □

A.2 Proof of Lemma 1

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

$$V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta) \geq V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta) \tag{15}$$

and

$$V(\theta'; \sigma) + \phi \cdot w(p(\theta'; \sigma), \theta') \geq V(\theta; \sigma) + \phi \cdot w(p(\theta; \sigma), \theta'). \tag{16}$$

Summing up (15) and (16), we obtain (with some rearrangement)

$$w(p(\theta;\sigma),\theta) - w(p(\theta';\sigma),\theta) \geq w(p(\theta;\sigma),\theta') - w(p(\theta';\sigma),\theta'). \tag{17}$$

Because $\theta > \theta'$ and $w(\cdot)$ has strictly increasing differences, (17) 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 (16) to hold it is necessary that $V(\theta;\sigma) \leq V(\theta';\sigma)$. □

A.3 Proof of Lemma 2

Take an equilibrium with σ and suppose that it satisfies D1. Suppose that there exists a non-singleton $J \subseteq [0,1]$ such that all types $\theta \in J$ choose the same (π, r) with $\mathbb{E}_{\pi}[v(r(s),\omega)] = V > \underline{V}$. Take a communication protocol $(\pi^{\varepsilon}, r^{\varepsilon}) \in \mathcal{C}^*$ that satisfies $\mathbb{E}_{\pi^{\varepsilon}}[v(r^{\varepsilon}(s),\omega)] = V - \varepsilon$, which must exist for sufficiently small $\varepsilon > 0$ (see footnote 14).

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

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

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

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

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

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

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

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

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

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

Finally, the out-of-equilibrium beliefs $\eta(\pi, r)$ for $(\pi, r) \in \mathcal{C} \setminus \mathcal{C}^*$ can be completed by following the procedure that we described in footnote 9.

Sequential Rationality. Note that, given H, any (π, r) and (π', r') 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 (π, r) that induces a material payoff V with $\bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x,x)}{\partial p} d\, x < V < \underline{V}$ will be sub-optimal for the sender, as she could always obtain a higher material payoff without

undermining her image. Hence, to verify the sequential rationality of the sender's strategy, it suffices to show that no type $\theta \in [0,1]$ of the sender can strictly benefit from mimicking another type $\theta' \in [0,1]$. Since all types $\theta, \theta' < \hat{\theta}$ are separating according to σ, we have

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

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

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

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

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

We distinguish two cases. First, suppose that there is $\theta \in [0,1]$ such that $\mathbb{E}_{\pi'}[v(r'(s),\omega)] = V(\theta;\sigma)$, which implies that $\underline{p}(\pi', r', \theta) = \mathbb{E}[\tilde{\theta}|\pi_\theta, r_\theta; \sigma]$. Consider any type θ' with $(\pi_{\theta'}, r_{\theta'}) \neq (\pi_\theta, r_\theta)$. We have already shown that this type has *strict* incentives *not* to mimic θ. This

implies $\underline{p}(\pi', r', \theta') > \underline{p}(\pi', r', \theta)$, and therefore $D^0(\pi', r', \theta') \subsetneq D(\pi', r', \theta)$. Conversely, for any type θ'' with $(\pi_{\theta''}, r_{\theta''}) = (\pi_\theta, r_\theta)$, clearly $\underline{p}(\pi', r', \theta'') = \underline{p}(\pi', r', \theta)$, and therefore $D^0(\pi', r', \theta'') = D^0(\pi', r', \theta) \supsetneq D(\pi', r', \theta)$. Thus, the D1 criterion requires that the receiver restricts his out-of-equilibrium belief to those types θ'' with $(\pi_{\theta''}, r_{\theta''}) = (\pi_\theta, r_\theta)$. However, our belief system was just chosen this way.

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

$$\underline{V} < \mathbb{E}_{\pi'}[v(r'(s), \omega)] \leq \bar{V} - \phi \int_0^{\hat{\theta}} \frac{\partial w(x, x)}{\partial p} dx. \tag{18}$$

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

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

where the strict inequality follows condition (i), and the weak inequality is jointly implied by (18), the indifference condition of the cut-off type $\hat{\theta}$, and Lemma A1. It is then clear that $\underline{p}(\pi', r', \theta) > \underline{p}(\pi', r', \hat{\theta})$ for all $\theta < \hat{\theta}$. Further, since

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

Lemma A1 and (18) jointly imply that

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

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

A.5 Proof of Theorem 2

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

Let $(\bar{\pi}, \bar{r})$ be a communication protocol that yields the expected payoff $\bar{V}$ to the sender. Without loss of generality, assume that $\bar{\pi}$ and π^N have no overlapping support. For each $\theta \in (0, \check{\theta})$, consider the following communication protocol $(\check{\pi}_\theta, \check{r}_\theta)$: with probability $\lambda(\theta) = (V(\theta; \sigma) - V^N)/(\bar{V} - V^N) < 1$, the protocol generates a signal s according to $\bar{\pi}$ and recommends $\bar{r}(s)$ to the receiver; with the remaining probability $1 - \lambda(\theta)$, the signal and the recommendation are generated according to (π^N, r^N). It is straightforward to check that $(\check{\pi}_\theta, \check{r}_\theta) \in \mathcal{C}^*$, $\mathbb{E}_{\check{\pi}_\theta}[v(\check{r}_\theta(s), \omega)] = V(\theta; \sigma)$, and

$$\mathbb{E}_{\check{\pi}_\theta}[u^R(\check{r}_\theta(s), \omega)] = \lambda(\theta) \cdot \bar{U} + (1 - \lambda(\theta)) \cdot \underline{U} < \bar{U}. \tag{19}$$

Next, define a strategy $\check{\sigma}$ of the sender as follows: for all $\theta \in (0, \hat{\theta})$, $\check{\sigma}(\theta) = (\check{\pi}_\theta, \check{r}_\theta)$; for all other θ, let $\check{\sigma}(\theta) = \sigma(\theta)$. By Theorem 1, $\check{\sigma}$ is part of a D1 equilibrium. Moreover, since the type distribution Γ is continuous and has full support, (19) 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 sender's image concerns.

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

$$\mathbb{E}_{\pi_\theta}[u^R(r_\theta(s), \omega)] < \mathbb{E}_{\pi_{\theta'}}[u^R(r_{\theta'}(s), \omega)]. \tag{20}$$

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

the recommendation are generated according to $(\pi_{\theta'}, r_{\theta'})$. It is straightforward to check that $(\check{\pi}_\theta, \check{r}_\theta) \in \mathcal{C}^*$, $\mathbb{E}_{\check{\pi}_\theta}[v(\check{r}_\theta(s), \omega)] = V(\theta; \sigma)$, and

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

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

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

$$\bar{V} - \phi \cdot \int_0^{\min\{\hat{\theta}, 1\}} \frac{\partial w(x, x)}{dp} dx \geq V^N. \tag{22}$$

We claim that in this case, the receiver's expected payoff must be decreasing everywhere on $[0, \hat{\theta})$. To prove this, suppose by contradiction that there exist $\theta, \theta' \in [0, \hat{\theta})$ such that $\theta < \theta'$ and (20) holds. Then, consider the following communication protocol $(\check{\pi}_{\theta'}, \check{r}_{\theta'})$: with probability $\lambda' = (V(\theta'; \sigma) - V^N)/(V(\theta; \sigma) - V^N) < 1$, the protocol draws a signal s according to π_θ and sends recommendation $r_\theta(s)$ to the receiver; with the remaining probability $1 - \lambda'$, the signal and the recommendation are generated according to (π^N, r^N). It is straightforward to check that $(\check{\pi}_{\theta'}, \check{r}_{\theta'}) \in \mathcal{C}^*$, $\mathbb{E}_{\check{\pi}_{\theta'}}[v(\check{r}_{\theta'}(s), \omega)] = V(\theta'; \sigma)$, and

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

Therefore, it is possible to construct a D1 equilibrium strategy $\check{\sigma}$ that always gives the receiver a weakly higher payoff than σ, and this payoff difference will even be strict when the sender's type is θ'. Hence, if (22) holds, the receiver's expected payoff $\mathbb{E}_{\check{\pi}_\theta}[u^R(\check{r}_\theta(s), \omega)]$ must be monotonically decreasing in θ within the interval $[0, \hat{\theta})$.

Second, suppose that (22) does not hold. Then, there must exist $\theta^N \in [0, \hat{\theta})$, such that

$V(\theta^N;\sigma) = V^N$. Using similar construction of "grand" communication protocols involving (π^N, r^N), it can be shown that the receiver's expected payoff must be first decreasing in θ on $[0, \theta^N]$, and then increasing on $[\theta^N, \hat{\theta})$. □

A.6 Transforming the Quadratic-Loss Games

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

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

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

Now, suppose that $f(\theta) > 0.5\ \forall\theta \in [0,1]$ and compare the following two utility functions of the sender: $u^S(a,\omega,\theta,\eta) = -(a-a^*(\omega,\theta))^2 + \phi\cdot w(p(\eta),\theta)$, and $\hat{u}^S(a,\omega,\theta,\eta) = -(a-\omega)^2 + \phi\cdot\hat{w}(p(\eta),\theta)$, where $\hat{w}(p(\eta),\theta) = w(p(\eta),\theta)/(2f(\theta)-1)$. We claim that, taken the receiver's best response $\hat{a}(\cdot)$ as given, these two utility functions represent the same preference over the pairs (π,η) for all types $\theta\in[0,1]$. This is because, for all $\theta\in[0,1]$ and all (π,η) and (π',η'), we have

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

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

Similarly, if $f(\theta) < 0.5\ \forall \theta \in [0,1]$, then, as described in Example 5, 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 The Sets of Implementable Payoffs in the Examples

Example 3. To be completed.

Example 5. To be completed.

Example 6. To be completed.

Example 7. To be completed.

Example 8. To be completed.

A.8 Proof of Theorem 3

Analogous to Theorem 2. □

A.9 Proof of Theorem 4

The first statement of the theorem can be verified using Example 8 in the main text. To prove the second statement, take any communication protocol $(\pi^N, r^N) \in \mathcal{C}^*$ ($(\pi^F, r^F) \in \mathcal{C}^*$) that conveys no (full) information about the state to the receiver, Let V^N and V^F be the expected material payoffs of the sender under (π^N, r^N) and (π^F, r^F), respectively. Since $U^* \in (\underline{U}, \bar{U})$ is uniquely defined, it is necessary that $\bar{V} > \max\{V^N, V^F\}$.

If ϕ is sufficiently small, all D1 equilibria will be fully separating, and the interim payoff of the highest type will satisfy $V(1;\sigma) > \max\{V^N, V^F\}$ irrespective of the which D1 equilibrium strategy σ is selected. In particular, there exists a D1 equilibrium in which each type θ uses a "grand" communication protocol that mixes appropriately between the sender-optimal protocol $(\bar{\pi}, \bar{r})$ absenting image concerns and the no-disclosure protocol (π^N, r^N). Clearly, the receiver is strictly worse off in this D1 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 the protocol $(\bar{\pi}, \bar{r})$ and the full-disclosure protocol (π^F, r^F). It

is straightforward to see that the receiver must be strictly better off in this D1 equilibrium relative to the equilibrium without image concerns.

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.

Beck, K., Beedle, M. and Bennekum, V. (2001). Manifesto for agile software development.

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. (2006). Incentives and prosocial behavior. *American economic review*, **96** (5), 1652–1678.

Bernhardt, D., Krasa, S. and Shadmehr, M. (forthcoming). Demagogues and the economic fragility of democracies. *American Economic Review*.

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. *The psychology of economic decisions*, **1** (105), 26.

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

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: Journal of the Econometric Society*, pp. 1431–1451.

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

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

Ely, J., Fudenberg, D. and Levine, D. K. (2009). When is reputation bad? In *A Long-Run Collaboration On Long-Run Games*, World Scientific, pp. 177–205.

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

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. MIT press.

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

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

Georgiadis, G. (2022). Contracting with moral hazard: A review of theory & empirics. *Available at SSRN 4196247*.

Guiso, L., Herrera, H., Morelli, M., Sonno, T. *et al.* (2017). Demand and supply of populism.

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.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The review of Economic studies*, **66** (1), 169–182.

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

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

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

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

Koessler, F. and Skreta, V. (2021). Information design by an informed designer.

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. *arXiv preprint arXiv:2206.09164*.

—, 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.

Mailath, G. J. (1987). Incentive compatibility in signaling games with a continuum of types. *Econometrica: Journal of the Econometric Society*, pp. 1349–1365.

—, Samuelson, L. *et al.* (2006). *Repeated games and reputations: long-run relationships*. Oxford University Press.

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

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

Morelli, M., Nicolò, A. and Roberti, P. (2021). A commitment theory of populism.

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

Myerson, R. B. (2009). Learning from schelling's strategy of conflict. *Journal of Economic Literature*, **47** (4), 1109–25.

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

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

Persson, T. and Tabellini, G. (2002). *Political economics: explaining economic policy*. MIT press.

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

Schelling, T. C. (1980). *The Strategy of Conflict: with a new Preface by the Author*. Harvard university press.

Schlenker, B. R. (2012). Self-presentation.

Smolin, A. and Yamashita, T. (2022). Information design in concave games, working paper.

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.