Disclosing Decision Makers' Private Interests *

Antoni-Italo de Moragas†

August 16, 2022

Abstract

I study whether a decision maker would make better decisions if his private interests were disclosed. I focus on a delegation relationship in which a decision maker attempts to build a good reputation. I compare the following two scenarios: a non-disclosure case in which the interest of the decision maker is private information and a disclosure case in which his interest is publicly known. I show that the disclosure of the private interests of the decision maker can only improve the decision made when reputation concerns are intermediate, and decision makers are sufficiently informed and public-spirited. Otherwise, disclosure can be detrimental because it induces decision makers to make decisions against their private interests too often. My framework provides a novel cautionary tale regarding the use of disclosure laws to address conflicts of interest.

*This paper is based on the first chapter of my Ph.D. dissertation at the European University Institute. I am grateful to my advisors Andrea Mattozzi and David K. Levine. I would also like to thank Ciril Bosch, Andrés Carvajal, Agustín Casas, Justin Fox, Moti Michaeli, Amedeo Piolatto, Santiago Sánchez Pagés, Bauke Visser and the participants at the Warwick Ph.D. Conference in Economics 2016, Applied Economics Workshop in Petralia 2016, Leicester Ph.D. Conference in Economics 2016, ASSET 2016, SAEe 2017, ETH Zurich, Max Planck Institute for Tax Law and Public Finance and CUNEF. I acknowledge the financial support received from the PGC2018 – 099415 – B – 100 MICINN/FEDER/UE research project.

†Department of Economics, CUNEF Universidad, C/Leonardo Prieto Castro 2, 28040 Madrid, Spain. E-mail address: antoni.demoragas@cunef.edu.

1 Introduction

When a delegation relationship risks having conflicts of interest, the disclosure of the conflicts of interest is one of the most frequently prescribed remedies. Advocates for disclosure contend that by being aware of potential private interests, the public can form its own judgement regarding whether decision makers place their personal benefits ahead of their duties. These advocates argue that such knowledge forces decision makers to behave according to the public interest instead of their own (Stark, 2003). This belief explains the presence of mandatory disclosure laws in various areas such as politics (Cain, 2014), academic research or medical practice (DeJong et al., 2016).

Despite the popularity of laws that require the disclosure of private interests, how disclosure affects the behavior of decision makers and the decisions that are eventually made remains an open question. In a world where agents such as doctors, politicians, or judges may be biased towards their private interests, would you, as the principal, want to know the direction of such bias or would you be better off ignoring it?

I address this question with a model in which a decision maker receives a private signal concerning a state of the world and must make a decision. The decision maker always has some private interest that can affect his preferred decision.These characteristics capture a setting in which one is uncertain regarding which banking outfit a given political party is indebted to (e.g., Goldman Sachs or Morgan Stanley) or toward which of two competing drug manufacturers a given doctor is biased (e.g., Pfizer or AstraZeneca). Nevertheless, the relevance of conflicts of interest depends on their private type. Good types prefer making the correct decision independent of their private interests, while the preferred decision of bad types depends only on their private interests.

In addition to their intrinsic preference towards the decision made, the decision maker

also has reputation concerns and wants to convince an external evaluator that he is a good type because this can create a reputation that can lead to higher wages, promotions, or reappointments in the future. The assessment of the evaluator depends on the information to which he has access. When the private interests of the decision maker are not disclosed, the evaluator can only observe the state of the world and the decision made. With disclosure, the evaluator also obtains insight into the private interests of the decision maker.

I find that the effect of disclosure on the correctness of the decisions made is ambiguous because it modifies the reputational incentives of the bad types but also distorts the incentives of good types. When the private interests of decision makers are not disclosed, making correct decisions is the only method for decision makers to improve their reputation. With disclosure, making decisions against the private interests of the decision maker also increases their reputation which can induce decision makers to make decisions against their private interest regardless of the signal they receive. I derive the conditions for the existence of the following three different equilibria: an equilibrium in which disclosure has no effect, an equilibrium in which disclosure drives decision makers to make the correct decisions more often, and an equilibrium in which disclosure decreases the probability of making correct decisions.

The existence of this last equilibrium shows that the disclosure of the private interests of decision makers can cause what Gersen and Stephenson (2014) calls over-accountability: a situation in which accountability worsens rather than improves the decision making. Specifically, I find that disclosure can only be beneficial if the reputation concerns of the decision maker are moderate and decision makers are sufficiently informed and public spirited.

2 Related Literature

This paper relates to three different streams of the literature. The first stream encompasses the literature studying the effects of transparency on decision making pioneered by Prat (2005). This literature studied the effect of making a decision public and can be classified into two broad categories depending on the nature of the heterogeneity of the decision maker. On the one hand, Prat (2005) and Fox and Van Weelden (2012) assume that decision makers differ in the quality of the information available to them.[1] On the other hand, Stasavage (2007) and Fox (2007) suppose that decision makers differ in their preferences. A common feature in my model and this literature is that transparency can be negative because it can give incentives to decision makers to select actions that makes them achieve a better reputation (i.e., being more informed or having more desirable preferences) which is not always the best action. Nevertheless, I contribute to this literature by studying the following novel source of transparency: instead of studying the effect of making a decision public, I study what occurs when the private interests (or potential bias) of decision makers are disclosed.

My model is closely related to Fox (2007) and Stasavage (2007) who also considered that the decision made signals the bias of the preferences of the decision maker. In particular, the game in the transparency case of their models is very similar to the game in my disclosure case. However, a crucial feature distinguishes my model from their models: instead of having an unbiased and a biased decision maker type as assumed in these articles, in my model there is uncertainty regarding the direction of the bias. This feature allows me to capture the effect of disclosing private interests. In a world where decision makers often have private interests,

[1]Instead of focusing on individual decision makers, Levy (2007), Visser and Swank (2007), Swank et al. (2008), Gersbach and Hahn (2008), Fehrler and Hughes (2018) and Mattozzi and Nakaguma (2022) study decisions in committees where decisions are informative regarding the quality of the members of the committee.

the disclosure of these interests cannot determine whether the decision maker is biased but can only inform about the potential direction of the bias. Being unbiased depends on the public-spiritedness of the decision maker which cannot be revealed.

Second, my model also relates to the literature concerning the disclosure of biases in expert advice settings (Li and Madarász, 2008; Ismayilov and Potters, 2013; Kartal and Tremewan, 2018). In particular, Li and Madarász (2008) extend the cheap-talk model described by Crawford and Sobel (1982) and study how the results change when the receiver is informed of the bias of the sender.[2] In their model, disclosure can harm the receiver because it can reduce the quality of the communication. The difference between this paper and Li and Madarász (2008) is that in the model presented here, biased decision makers make a payoff-relevant action (delegation), while in Li and Madarász (2008), the utility of the decision maker exclusively depends on how the receiver interprets his message (advice). Therefore, the logic of the bad equilibrium is qualitatively different, and more importantly, the situations that can be captured by the two models differ. For example, expert advice models are not adequate for capturing the effect of the disclosure of private interests in politics because politicians do not advise but make decisions on our behalf.

Third, as it is common in the literature studying the effects of transparency on decision making, the mechanism by which disclosure can worsen the decisions in my model is related to what Canes-Wrone et al. (2001) defined as pandering: to increase his reputation, the decision maker adopts an action that does not benefit him, the public or the principal's interests.[3] In these models, an increase in reputation concerns can worsen the decision made. My model highlights that the adverse effects of reputation concerns are only present

[2]My model is also related to Sobel (1985), Benabou and Laroque (1992) and Dziuda (2011) who study the reputation concerns of experts when the receiver is uncertain regarding the preferences of the sender.

[3]Other papers showing similar results in different settings include Morris (2001), Ely and Välimäki (2003), (Maskin and Tirole, 2004), Canes-Wrone and Shotts (2007), Ely et al. (2008), Acemoglu et al. (2013), Che et al. (2013) and Smart and Sturm (2013).

when the bias of the decision maker is known by the evaluator. When there is sufficient uncertainty regarding the direction of the bias , higher reputation concerns always increase the probability of making correct decisions.

3 The Model

An agent, i.e., the decision maker, must make a decision, d. For simplicity, we assume that there are only two possible decisions $d \in \{a, b\}$. In a business, this could represent a manager selecting between two job applicants; in politics, a politician may have to choose between contracting with firm a or firm b on a public project; and in medical practice, it could be a doctor prescribing a drug from one of two different pharmaceutical companies. There is a state of the world $\omega \in \{a, b\}$. The state of the world is unknown, but it is common knowledge that both states can occur with equal probability.

Before a decision is made, the decision maker observes a private signal $s \in \{a, b\}$ about the state of the world ω. The signal is only partially informative of the realized state of the world. In particular, we assume that $q = Pr(s = \omega|\omega) > \frac{1}{2}$, and we refer to q as the precision of the signal. The correctness of the decision always depends on the state of the world. Specifically, a decision is correct when it matches the state of the world ($d = \omega$).

The decision maker always has a private interest $\beta \in \{a, b\}$ in one of the alternatives. We should think about this as follows. In a business, one of the job applicants can be a friend of the hiring manager. In politics, one of the firms that the politician can contract might be a campaign donor of his party or a politically connected firm that can offer him a job in the future in exchange. Finally, in medical practice, the doctor might be receiving benefits from the pharmaceutical company of one of the drugs he can prescribe.

It is common knowledge that both private interests are equally likely and uncorrelated

with the state of the world ω.[4] However, the weight that the decision maker attaches to his private interest depends on his type θ which can be either good ($\theta = 1$) with probability $\mu \in (0, 1)$ or bad ($\theta = 0$) with the remaining probability $1 - \mu$. The type of the decision maker can be interpreted as a measure of the public-spiritedness of the decision maker. For simplicity, we assume that good types care about making correct decisions regardless of their private interest and that bad types only care about making the decision that coincides with their private interest regardless of the state of the world.[5] Naturally, decision makers know their type, i.e., they know whether they care about their private interest.

In addition to the decision maker, there is an evaluator. Similar to other reputation concerns models[6], the evaluator does not have any utility function, and his task is simply to rationally update his beliefs regarding the type of decision maker (i.e., using Bayes rule whenever possible). The decision maker gains utility from advancing in his career which will only happen if the evaluator believes that he is a good type.[7] We can consider the evaluator the owner of the firm in business, the electorate in politics or the patient in medical practice. We will refer to these beliefs as the reputation R of the decision maker. Clearly, in all these settings, the decision maker has good reasons to improve his reputation with the evaluator.

Let $\phi > 0$ measure the importance of reputation R. The utility of a decision maker of type θ with private interest β when making decision d in state of the world ω is expressed as follows:

$$u(\beta, \omega, d) = \theta \mathbb{1}_{d=\omega} + (1 - \theta)\mathbb{1}_{d=\beta} + \phi R$$

[4] In Section 6 we briefly show that the main results of the paper hold if we assume that one private interest is more likely than the other. A more detailed explanation can be found in Appendix A.2.

[5] We can relax the assumption such that both types care about making decisions that coincide with their private interest and instead assume that they differ in the intensity of their preferences.

[6] See Levy (2004), Levy (2007) and Mattozzi and Nakaguma (2022).

[7] While we do not explicitly model the career advancement, the model can be interpreted as the first of a two-stage model in which the probability of being reappointed is linear in the reputation of the decision maker. In the second stage of such a model, good decision makers would follow their interest and bad decision makers would follow their private interest as in Fox (2007).

We will study the following two different institutional settings: one with and one without the disclosure of private interests. The only difference between these institutional settings is the information available to the evaluator when he updates his beliefs regarding the type of decision maker. In particular, without disclosure, the evaluator only observes the decision made and the state of the world. Thus, his beliefs are $R(\omega, d) = E[\theta|\omega, d]$. With disclosure, the evaluator also observes the private interest of the decision maker, and $R(\beta, \omega, d) = E[\theta|\beta, \omega, d]$.

The strategy of a decision maker is to select d depending on his type θ, his private interest β and signal s; thus, it is a mapping $\alpha(\theta, \beta, s) \to \{a, b\}$. The following pure strategies will be useful in characterizing the equilibria. We will say that the decision maker follows his signal if $\alpha(\theta, \beta, s) = s$ for $s \in \{a, b\}$, and he contradicts it if $\alpha(\theta, \beta, s) \neq s$. Analogously, the decision maker follows his private interest if $\alpha(\theta, \beta, s) = \beta$ for $\beta \in \{a, b\}$ and contradicts it if $\alpha(\theta, \beta, s) \neq \beta$. In the latter, we will also use the notation $\alpha(\theta, \beta, s) = \beta^c$. To summarize, the timing is as follows:

1. The state ω, the private interest β and the type θ of the decision maker are realized.
2. The decision maker observes the signal s and makes decision d.
3. Without disclosure, the evaluator observes d and ω. With disclosure, the evaluator observes d, ω and β. The evaluator forms a posterior R of the decision maker's type.
4. Payoffs are realized.

We will solve the model using the concept of Perfect Bayesian Equilibrium. We will only consider equilibria such that the good types play pure strategies.[8] When there are multiple equilibria, we will select the equilibrium that maximizes the probability of taking the correct

[8]This rules out unstable equilibria. If the good types play non-degenerate mixed strategies in equilibrium, they can obtain a higher reputation and higher utility from any deviation if the evaluator also updates his beliefs, but this is not the case for the bad types.

decision.[9]

4 Results without the disclosure of private interest

Before analyzing the behavior of decision makers with disclosure, we must study what occurs without it as a benchmark. Without disclosure, the reputation of decision makers cannot depend on their private interest β because the evaluator does not observe it. Thus, the beliefs of the evaluator given the observation (d, ω) and his conjecture of α are as follows:

$$R(\omega, d) = \frac{Pr(\theta = 1) \sum_{\beta', s'} Pr(s = s'|\omega) Pr(\alpha(1, \beta', s') = d)}{\sum_{\theta', \beta', s'} Pr(\theta = \theta') Pr(s = s'|\omega) Pr(\alpha(\theta', \beta', s') = d)}$$

Thus, since both states and both private interests are equally likely and uncorrelated with the state of the world, there is always an equilibrium in which the good types follow the signal.[10]

Proposition 1. *There is always a threshold $\phi_{ND}(q, \mu)$ such that in equilibrium: (i) When $\phi \leq \phi_{ND}(q, \mu)$, the good types follow the signal, and the bad types follow their private interest. (ii) When $\phi > \phi_{ND}(q, \mu)$, the good types follow the signal, and the bad types mix between following the signal and the private interest.*

Without disclosure, the good types follow the signal and maximize the likelihood of making the correct decision because their reputational incentives are aligned with their preferred decision. This is not the case for bad types. When the bad types receive a signal

[9] Similarly to Prat (2005), by selecting the equilibrium that maximizes the probability of a correct decision, we are comparing the best equilibrium of disclosure against the best equilibrium of non-disclosure.

[10] Additionally, under some parameters, there are other equilibria (i.e. equilibria in which good types always make the same decision independently of the signal or "mirror" equilibria (Levy, 2004, 2007) where they contradict the signal they receive). Clearly, in all these alternative equilibria, the probability of making a correct decision is lower than in the equilibria in which good types follow their signal and we ignore them.

opposed to their private interest, they experience a trade-off between their reputation (following the signal) and their present utility (following their private interest). Unsurprisingly, when reputation concerns are sufficiently low, the bad types follow their private interest, but when reputation concerns are higher, they mix between following their private interest and following the signal.

Corollary 1. *When the private interest of the decision maker is not disclosed, the probability of a correct decision is always increasing in the reputation concerns of the decision maker.*

As the previous corollary shows, higher reputation concerns always increase the probability of correct decisions. Because the good types always follow the signal, the results are driven by the behavior of the bad types. In particular, when the bad types mix, the probability of following the signal increases when reputation concerns increase. Higher reputation concerns increase the probability of correct decisions because without disclosure, the only method to increase reputation is to make correct decisions. Notably, however, the bad types never follow the signal with probability one because, if they did all reputational incentives would vanish.

5 Results with the disclosure of private interest

After analyzing the behavior of decision makers without the disclosure of their private interests, we can study the effect of such disclosure. When an evaluator knows the private interest β of a decision maker, making the correct decision is not the only relevant information that the evaluator considers in assessing the reputation of the decision maker. In particular, the evaluator considers whether the decision made coincides with the private interest of the decision maker. The beliefs of the evaluator regarding the type θ given the observation (β, ω, d) and his conjecture of the strategy α are as follows:

$$R(\beta,\omega,d)=\frac{Pr(\theta=1)\sum_{s'}Pr(s=s'|\omega)Pr(\alpha(1,\beta,s')=d)}{\sum_{\theta',s'}Pr(\theta=\theta')Pr(s=s'|\omega)Pr(\alpha(\theta',\beta,s')=d)}$$

As a first step, we derive the reputational incentives that characterize the equilibria with the disclosure of the private interest of the decision maker:

Lemma 1. *In equilibrium,* $R(\beta,\omega,d)$ *satisfies* $R(\beta,\beta^c,\beta)\leq R(\beta,\beta,\beta)\leq R(\beta,\beta,\beta^c)\leq R(\beta,\beta^c,\beta^c)$.

The second inequality of Lemma 1 implies that contradicting one's private interest always signals a higher reputation than following it because the cost of contradicting one's private interest is always higher for bad than for good decision makers. The first and third inequalities indicate that conditional on contradicting or following one's private interest, making the correct decision also increases the reputation because the good types incur a cost from not making correct decisions.

Notice that these reputational incentives distort the incentives of the good types to follow their signal. Without disclosure, good decision makers always maximize their reputation and present utility by following the signal. With disclosure, however, when a good type receives a signal that coincides with his private interest, he experiences a trade-off between following his signal to maximize the probability of a correct decision and contradicting his private interest to maximize his reputation.

The tension between the incentives of bad types to follow the signal and the incentives of good types to contradict their private interest shapes the characterization of the equilibria with disclosure. We will derive the conditions for the existence of three types of equilibrium. A *Nothing Changes Equilibrium*, where the good types follow the signal and the bad types follow their private interest. A *Disciplining Equilibrium*, where the good types follow the signal and the bad types mix between following the signal and following their private interest.

Finally, a *Pandering Equilibrium*, where the good types contradict their private interest and ignore the signal and bad types also ignore their signal.

Let us start with the Nothing Changes Equilibrium:

Proposition 2. *There exists a Nothing Changes Equilibrium such that good decision makers follow the signal and bad decision makers follow their private interest if and only if* $\phi < \phi_D(q, \mu)$. *Moreover,* $\phi_D < \phi_{ND}$.

The first part of the proposition shows that, when reputational incentives are sufficiently low, reputation concerns do not induce either good or bad types to shift from their preferred decision; therefore, the good types follow the signal, and the bad types follow their private interest. Recall that this is also the case in the equilibrium without disclosure. However, as the second part of the proposition shows, disclosure tightens the conditions for its existence because it induces larger reputational incentives to depart from the preferred decision. On the one hand, the good types have incentives to contradict their private interest (they have no incentive to do so without disclosure). On the other hand, the bad types anticipate a larger decrease in their reputation if they make a wrong decision aligned with their private interest.

To study the other two equilibria we need to study which type of decision maker needs lower reputation concerns to deviate from the Nothing Changes Equilibrium. Let ϕ_b and ϕ_g be the lower reputation concerns that make the bad and good decision makers, respectively, deviate from this equilibrium (note that $\phi_D = min\{\phi_b, \phi_g\}$). When $\phi_b < \phi_g$, the bad types require a lower ϕ than the good types to deviate from the equilibrium, i.e., following the signal when it does not coincide with their private interest. When $\phi_b > \phi_g$, the good types are more prone to deviate and contradict the signal when it coincides with their private interest.

Lemma 2. *There exists a $\bar{\mu}(q)$ such that the bad types require lower reputation concerns to deviate from the Nothing Changes Equilibrium (i.e. $\phi_b < \phi_g$) if and only if $\mu > \bar{\mu}(q)$. $\bar{\mu}(q)$ is decreasing in q.*

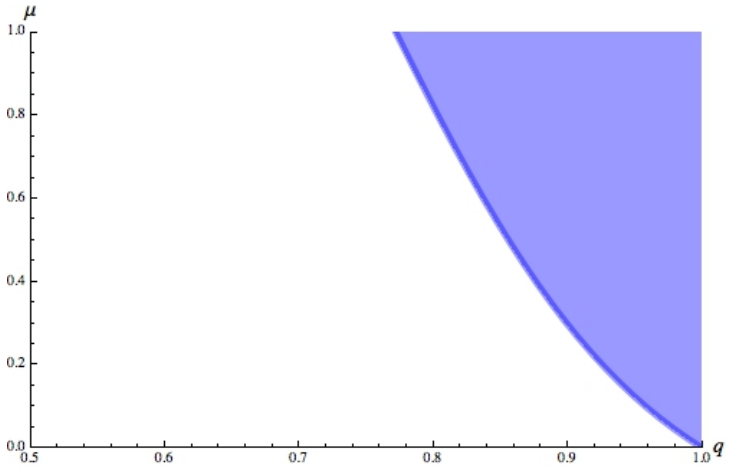

Figure 1: Values such that $\phi_b < \phi_g$.

The previous lemma characterizes the (μ, q) that guarantees that $\phi_b < \phi_g$. Next, we can characterize the conditions for the existence of the Disciplining and Pandering equilibria and at the end of this section, we will discuss the intuition behind Lemma 2.

Proposition 3. *There exists a Disciplining Equilibrium in which the good types follow the signal and the bad types mix between following the signal and their private interest if and only if $\mu \geq \bar{\mu}(q)$ and $\phi \in [\phi_b, \phi_g]$. In the Disciplining Equilibrium, the bad types follow the signal with higher probability than without disclosure and the probability of making a correct decision is higher.*

Proposition 3 shows that when reputation concerns are neither too large nor too low, there exists an equilibrium in which the good types follow the signal, and the bad types mix between following the signal and their private interest. Recall that without disclosure, when reputation concerns are sufficiently high, there is already an equilibrium in which the bad types mix between their private interest and the signal.

However, the last part of the proposition shows that in this equilibrium, the bad types follow the signal more often than without disclosure. Given that the good types do not change their behaviour, this implies that the disclosure of private interests increases the probability that the decision maker makes correct decisions by disciplining the bad types and inducing them to follow the signal more often. The intuition is that when the good types follow the signal, making a decision that aligns with the private interest of the decision maker and does not match the state of the world is a stronger signal of being a bad type. Therefore, the incentives for following the signal are higher in this equilibrium than in the absence of disclosure.

Proposition 4. *There exists a Pandering Equilibrium such that the good types always contradict their private interest if and only if* $2q - 1 < \phi$. *In particular, in this equilibrium:* (i) *if* $\phi \in (2q - 1, 1]$, *the good types contradict their private interest, and the bad types follow it.* (ii) *If* $\phi \in (1, \frac{1}{\mu})$, *the good types contradict their private interest, and the bad types mix between contradicting and following it.* (iii) *If* $\phi \geq \frac{1}{\mu}$, *the good and bad types contradict their private interest.*

When the precision of the signal received by decision makers is low with respect to their reputation concerns, it is optimal for the good types to contradict their private interest in equilibrium. Intuitively, contradicting one's private interest always increases the reputation of the decision maker. Thus, when reputation concerns increase, the benefits of contradicting private interests increase. This strategy comes at the cost of making their preferred decision less often. This cost depends on the precision of the signal. In particular, when the precision of the signal decreases, the cost of contradicting the private interest decreases; therefore, it is more profitable to contradict the private interest. As q tends to $\frac{1}{2}$, this cost vanishes. Regarding the bad types, recall that when the good types do not follow the signal but only contradict their private interest, they have no incentives to follow the signal but, in any case,

to contradict their private interest. Of course, the bad types have fewer incentives than the good types to contradict their private interest.

After deriving all equilibria with disclosure, we can compare the probability of making a correct decision in each equilibrium as follows:

Lemma 3. *Keeping q and μ fixed, from lower to higher probability of making a correct decision, the order of the equilibria with disclosure is as follows: Pandering Equilibrium $\prec$ Nothing Changes Equilibrium $\prec$ Disciplining Equilibrium.*

The intuition behind this order is straight-forward. In the Pandering Equilibrium, both types ignore the signal they receive; in the Nothing Changes-Equilibrium, the good types follow the signal and the bad types ignore it; and in the Disciplining Equilibrium, the good types always follow the signal and the bad types follow the signal with some probability. This order allows us to select an equilibrium with disclosure when multiple equilibria coexist.[11] We are ready to provide the main result of the paper and characterize the conditions under which the disclosure of private interests increases the probability of making a correct decision.

Proposition 5. (i) *Consider $\mu < \bar{\mu}(q)$. When $\phi < \phi_b$, the disclosure has no effect; when $\phi \in [\phi_b, \phi_g]$, the disclosure increases the probability of a correct decision; and when $\phi > \phi_g$, the disclosure decreases the probability of a correct decision.* (ii) *Consider $\mu \geq \bar{\mu}(q)$. When $\phi < \phi_b$, the disclosure has no effect; and when $\phi > \phi_g$, the disclosure decreases the probability of a correct decision.*

The previous proposition summarizes the parameters under which the disclosure of decision makers' private interests can increase or decrease the probability of making correct decisions. When reputation concerns are low, the good and bad types do not renounce to

[11] In Appendix B, we show that the Disciplining Equilibrium always coexist with the Pandering Equilibrium. When this is the case, we select the Disciplining Equilibrium because it maximizes the probability of making a correct decision.

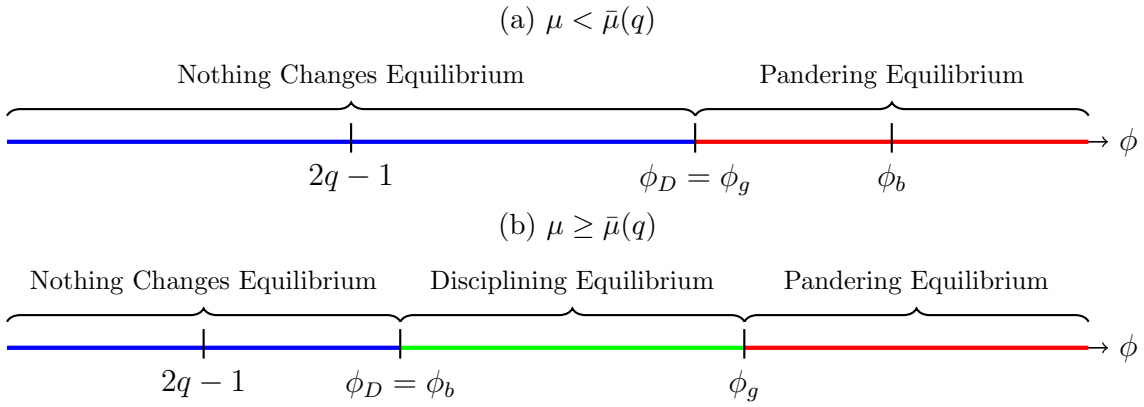

Figure 2: Equilbria characterization. Subfigure 2a shows the equilibria when $\mu < \bar{\mu}(q)$. Subfigure 2b shows the equilibria when $\mu \geq \bar{\mu}(q)$. When two equilibria co-exist, we select the equilibrium that maximizes the probability of a correct decision.

their preferred decision to gain a better reputation; consequently, disclosure has no effect on the decision made (Nothing Changes Equilibrium). When reputation concerns are high enough, all types are willing to relinquish their preferred decision for an increase in reputation. In particular, the good types stop following their signal and contradict their private interest which, in turn, encourages the bad types to do the same. As a consequence, neither of the types follows their signal and the probability of a correct decision decreases with respect to the non-disclosure case (Pandering Equilibrium).

Therefore, the benefits of disclosure can only arise with intermediate values of the reputation concerns. If reputational concerns are high enough for the bad types to follow their signal but low enough to prevent the good types from contradicting their private interest, disclosure can discipline the bad types to follow their signal without inducing the good types to pander and contradict their private interest (Disciplining Equilibrium). However, it could be that with intermediate values, the bad types still do not follow their signal but the good types already contradict their interests. The beneficial effects of disclosure require the condition we derived in Lemma 2 to hold, i.e., the precision of the signal q and the fraction of good types μ must be high enough or, in other words, decision makers have to be sufficiently informed and public spirited.

Why does precision have to be sufficiently high? When the precision of the signal is high enough, good decision makers have high opportunity costs of contradicting their signal and lower reputational incentives to contradict their private interest which increases ϕ_g. The effect on ϕ_b is the opposite because an increase in the precision of the signal increases the reputational incentives for following it. Thus, an increase in q enlarges the region $[\phi_b, \phi_g]$ that guarantees the existence of the Disciplining Equilibrium.

The reason why the fraction of good types μ must also be large enough is that when μ increases, the reputation from making a correct decision increases, which decreases the incentives for good decision makers to contradict their private interest and increases ϕ_g. A large μ also decreases the incentives of the bad types to contradict their interest which increases ϕ_g. However, when q is large enough, the increase in ϕ_g is stronger and an increase in μ enlarges the region where the Disciplining Equilibrium exists.

6 Discussion

In this section I discuss alternative modeling assumptions. The details of the results mentioned in this discussion are provided in Appendix A.

Multiple decisions We considered an environment in which the state of the world, the decision and the private interests were binary. In a richer setting with n alternatives, disclosure is also characterized by the tension between the incentives of the bad types to follow the signal and the incentives of the good types to contradict their private interest. Nevertheless, when n increases, the Disciplining Equilibrium exists under a larger set of parameters because it is more costly for good decision makers to contradict their signal. Moreover, when n is sufficiently high, the Pandering Equilibrium with disclosure can lead to more correct

decisions than the equilibrium without disclosure because, in the Pandering Equilibrium, the good types follow the signal for the $n-1$ signals that do not coincide with their private interest.

Asymmetric likelihood of the private interest We assumed that both private interests were equally likely. The main results of the model are robust to some degree of asymmetry in the distribution of private interests. However, when one private interest is much more likely than the other and the signal of decision makers is sufficiently imprecise, the distortion of the incentives of good decision makers also occurs without disclosure, and they might pander against the most likely private interest.

Observability of the state of the world We assumed that the evaluator can always observe the state of the world. If he cannot, disclosure can only lead to fewer correct decisions because it induces the good types to contradict their interest without incentivizing the bad types to follow their signal.

Opportunistic decision makers We assumed that good types cared about making correct decisions. If we assume, instead, that they are opportunistic and only care about their reputation, disclosure can only decrease the probability of making correct decisions because the good types would always pander with disclosure.

7 Conclusion

We have shown that the disclosure of the private interests of decision makers does not necessarily imply better decision making. The reason is that when private interests are not disclosed, decision makers can be evaluated only based on the correctness of their decisions

and when they are disclosed, they are also appraised by whether these decisions align with their interests. This evaluation can prevent decision makers from placing their interests before their duties because they can be more severely judged. However, it also distorts the accountability process because instead of being only accountable for the correctness of their decisions, decision makers are also accountable for the profits they obtain, and sometimes, the correct decision coincides with the decision that benefits them.

The results of my model imply that the disclosure of private interests can only improve the decision making when decision makers are sufficiently informed and when most of them place common goals ahead of their private interests. Otherwise, disclosure can lead to more incorrect decisions. Therefore, the disclosure of private interests should only be required in good institutions where decision makers are expected to be both competent and public spirited. We end this article with a caveat. Throughout the article we assumed that private interests are exogenous. However, in some contexts (e.g., lobbying), disclosure laws can be beneficial because they can discourage the formation of conflicts of interest. Nevertheless, the results of this paper should be interpreted as a cautionary tale of the indiscriminate use of disclosure laws to address conflicts of interest.

References

D. Acemoglu, G. Egorov, and K. Sonin. A political theory of populism. *Quarterly Journal of Economics*, 771:805, 2013.

R. Benabou and G. Laroque. Using privileged information to manipulate markets: Insiders, gurus, and credibility. *Quarterly Journal of Economics*, 107(3):921–958, 1992.

B. E. Cain. *Democracy More Or Less: America's Political Reform Quandary.* Cambridge University Press, 2014.

B. Canes-Wrone and K. W. Shotts. When do elections encourage ideological rigidity? *American Political Science Review*, 101(2):273–288, 2007.

B. Canes-Wrone, M. C. Herron, and K. W. Shotts. Leadership and pandering: A theory of executive policymaking. *American Journal of Political Science*, pages 532–550, 2001.

Y.-K. Che, W. Dessein, and N. Kartik. Pandering to persuade. *American Economic Review*, 103(1):47–79, 2013.

V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica*, pages 1431–1451, 1982.

C. DeJong, T. Aguilar, C.-W. Tseng, G. A. Lin, W. J. Boscardin, and R. A. Dudley. Pharmaceutical industry–sponsored meals and physician prescribing patterns for medicare beneficiaries. *JAMA Internal Medicine*, 176(8):1114–10, 2016.

W. Dziuda. Strategic argumentation. *Journal of Economic Theory*, 146(4):1362–1397, 2011.

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

J. C. Ely and J. Välimäki. Bad reputation. *Quarterly Journal of Economics*, pages 785–814, 2003.

S. Fehrler and N. Hughes. How transparency kills information aggregation: Theory and experiment. *American Economic Journal: Microeconomics*, 10(1):181–209, 2 2018.

J. Fox. Government transparency and policymaking. *Public Choice*, 131(1-2):23–44, 2007.

J. Fox and R. Van Weelden. Costly transparency. *Journal of Public Economics*, 96(1): 142–150, 2012.

H. Gersbach and V. Hahn. Should the individual voting records of central bankers be published? *Social Choice and Welfare*, 30(4):655–683, 2008.

J. E. Gersen and M. C. Stephenson. Over-accountability. *Journal of Legal Analysis*, 6(2): 185–243, 2014.

H. Ismayilov and J. Potters. Disclosing advisor's interests neither hurts nor helps. *Journal of Economic Behavior & Organization*, 93:314–320, 2013.

M. Kartal and J. Tremewan. An offer you can refuse: the effect of transparency with endogenous conflict of interest. *Journal of Public Economics*, 161:44–55, 2018.

G. Levy. Anti-herding and strategic consultation. *European Economic Review*, 48(3):503–525, 2004.

G. Levy. Decision-making procedures for committees of careerist experts. *American Economic Review*, 97(2):306–310, 2007.

M. Li and K. Madarász. When mandatory disclosure hurts: Expert advice and conflicting interests. *Journal of Economic Theory*, 139(1):47–74, 2008.

E. Maskin and J. Tirole. The politician and the judge: Accountability in government. *American Economic Review*, 94(4):1034–1054, 2004.

A. Mattozzi and M. Y. Nakaguma. Public versus secret voting in committees. *Journal of the Economic European Association*, Forthcoming, 2022.

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

A. Prat. The wrong kind of transparency. *American Economic Review*, 95(2):862–877, 2005.

M. Smart and D. M. Sturm. Term limits and electoral accountability. *Journal of Public Economics*, 107:93–102, 2013.

J. Sobel. A theory of credibility. *The Review of Economic Studies*, 52(4):557–573, 1985.

A. Stark. *Conflict of interest in American public life*. Harvard University Press, 2003.

D. Stasavage. Polarization and publicity: rethinking the benefits of deliberative democracy. *Journal of Politics*, 69(1):59–72, 2007.

J. Swank, O. H. Swank, and B. Visser. How committees of experts interact with the outside world: some theory, and evidence from the fomc. *Journal of the European Economic Association*, 6(2-3):478–486, 2008.

B. Visser and O. H. Swank. On committees of experts. *Quarterly Journal of Economics*, 122(1):337–372, 2007.

A Additional Results

We complete our analysis by relaxing some assumptions of the model to illustrate the robustness of the main results.

A.1 Multiple decisions

Thus far, we assumed that there are only two possible states of the world and, consequently, only two possible private interests, actions and signals. In this section, we study the case of $n > 2$ states of the world. Specifically, the space of states (and private interests) is given by $\Omega = \{\omega_1, \cdots, \omega_n\}$, all states are ex ante equally likely ($Pr(\omega = \omega_i) = \frac{1}{n}$), and the signal is informative, i.e., the precision of the signal is such that $q = Pr(s = \omega|\omega) > \frac{1}{n}$.

The first consequence of expanding the space of states is that it creates higher incentives to follow the signal. From the perspective of good politicians, contradicting the signal is more costly because the probability of making their preferred decision when doing so is $\frac{1-q}{n-1}$, which decreases in n. From the perspective of bad politicians, the cost is also higher because the probability of obtaining a good reputation by contradicting the signal is lower. Thus, a larger space corresponds to a larger set of parameters such that the Disciplining Equilibrium exists as follows:

Proposition 6. *If $\mu > \mu(q, n)$, there always exists a $\phi > 0$ such that a Disciplining Equilibrium exists. Moreover, $\mu(q, n)$ is decreasing in n.*

The second consequence of expanding the set of states is that the Pandering Equilibrium is not as negative as it is with only two states. The reason is that when $n = 2$, in the Pandering Equilibrium, the decision maker completely ignores the signal that he receives because there is only one way to contradict the private interest. In contrast, when $n > 2$,

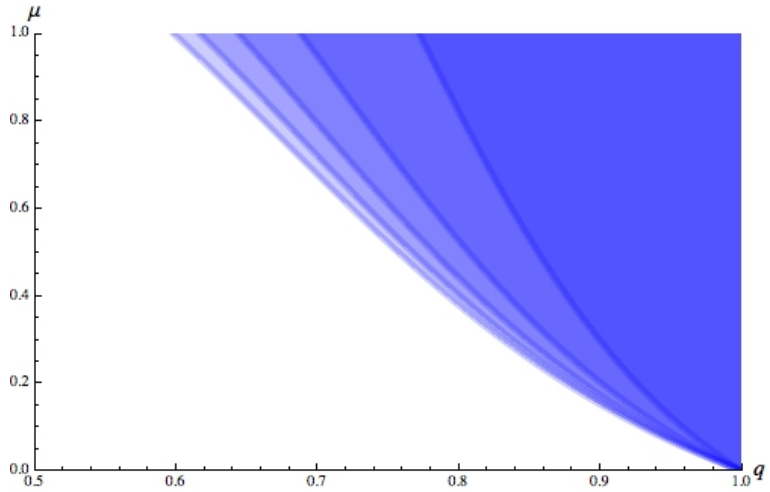

Figure 3: Values such that a Disciplining Equilibrium exists depending on n. From $n = 2$ (dark) to $n = 6$ (light).

the decision maker has $n-1$ decisions that contradict his private interests, and if the signal does not coincide with his private interest, he can always make a decision that matches the signal.

Proposition 7. *When $n > 2$, there exists a $\hat{\phi} > 0$ such that when $\phi > \hat{\phi}$, the probability of a correct decision is greater in the Pandering Equilibrium with disclosure than in the equilibrium without disclosure.*

Thus, interestingly, when n is sufficiently large, disclosure increases the probability of a correct decision even when the good types pander. This case occurs precisely when the reputation concerns are sufficiently high because the bad types need high reputation concerns to contradict their private interest.

A.2 Asymmetric likelihood of the private interest

Assuming that both interests are equally likely creates a scenario in which the disclosure of private interests is more informative because it resolves a larger degree of uncertainty.

Nevertheless, a possible concern is that the results of the model rely on the knife edge assumption that the probabilities of each private interest are exactly the same: $Pr(\beta = a) = Pr(\beta = b) = \frac{1}{2}$. We show that our results are generalizable to a larger range of probabilities.

Without loss of generality, assume that $Pr(a) = p > 1/2$ and $Pr(b) = 1 - p < 1/2$. Notably, the characterization of the equilibria with disclosure remains unchanged because disclosure resolves all previous uncertainty, and the prior probability of each private interest no longer enters the reputation function; we should only focus on the case without disclosure.

Proposition 8. *For every $\mu \in (0, 1)$ and $p > \frac{1}{2}$, there exists a $q' \in (p, 1)$ such that for every $q > q'$, there is a threshold $\phi'_{ND}(q, \mu)$ such that in equilibrium: (i) When $\phi \leq \phi'_{ND}(q, \mu)$, the good types follow the signal, and the bad types follow their private interest. (ii) When $\phi > \phi'_{ND}(q, \mu)$, the good types follow the signal, and the bad types mix between following the signal and their private interest.*

Proposition 8 shows that the main results of the paper are robust to some degree of asymmetry in the private interest. When $p > \frac{1}{2}$, the evaluator assigns a higher reputation to decisions with $d = b$ than decisions with $d = a$ conditional on both being correct or incorrect. Nevertheless, the high precision of the signal relative to the asymmetry of the private interest guarantees that this distortion is not high enough to modify the reputational incentives of the decision maker to follow the signal.

A.3 Observability of the state of the world

If the state of the world is not observed, because the distribution of private interests and states of the world is symmetric, the evaluator can only infer the type of the decision maker based on whether he follows or contradicts his private interest. In particular, because the evaluator cannot assess whether the decision maker made the correct decision, following the

signal cannot lead to a high reputation.

Without disclosure, all types ignore their reputation concerns, the good types follow the signal and make correct decisions, and the bad types follow their private interest.

Disclosure does not change the behavior of the bad types if the good types continue to follow the signal. Suppose that all types play the same strategy as they would in the absence of disclosure. Then, the reputation from following one's private interest is lower than the reputation from contradicting one's private interest. Thus, the good types are more prone than the bad types to contradict their private interest when the signal is aligned with their private interest because their loss of immediate utility is $(2q-1)$, whereas such loss for the bad types is 1. Thus, any equilibrium with disclosure is either a Nothing Changes Equilibrium or a Pandering Equilibrium, and disclosure can only reduce efficiency.

The same reasoning applies if instead of assuming that the state of the world is not observed, we assume that the observation is subject to noise; for example, we could assume that the evaluator receives a signal $s' \in \Omega$ of the state of the world such that $q' = Pr(s' = \omega|\omega) \in (.5, 1)$. When q is sufficiently large, the situation is similar to that studied in sections 4 and 5; when q is sufficiently low, the situation is similar to that studied in the previous paragraph. In summary, if the evaluator does not perfectly observe the state of the world, the bad effects of disclosure are more pronounced than when the observation is perfect.

A.4 Opportunistic decision makers

A critical assumption of the model is that the good types care about making the correct decision. In this section, we study the case in which the good types only have reputation concerns; i.e., their utility is simply ϕR. In the absence of disclosure, we can sustain the equilibrium in which the good types follow the signal and the bad types either follow their

private interest or mix between following the signal and their private interest.

Proposition 9. *When the good types only have reputation concerns, in equilibrium, the good types always contradict their private interest, and disclosure always reduces the probability of a correct decision.*

When we analyze the disclosure case, it is straightforward that the good types always pander and propose the policy that contradicts their private interest. Thus, the only possible equilibria with disclosure is the Pandering Equilibrium described in Section 5, and when the good types do not care about their decision, disclosure always reduces the probability of a correct decision.

B Coexistence of the Pandering and Disciplining equilibria

By construction, the Nothing Changes Equilibrium cannot coexist with the Pandering and Disciplining equilibria but the last two equilibria can coexist. The following corollary characterizes the coexistence of the Disciplining and Pandering equilibria:

Corollary 2. *When a Disciplining Equilibrium exists, a Pandering Equilibrium exists too.*

Since the utility of the good type when following the signal is lower than the utility of the bad type when following his private interest, it is more costly for the bad type to depart from their preferred decision than for the good type.[12]

[12]This would not be the case if the good and bad types had different reputation concerns ϕ.

C Proofs

Proof of Proposition 1

(i) Suppose that the good types follow their signal and the bad types follow their private interest. The consistency of the beliefs requires the following:

$$\begin{cases} R(\omega,\omega) = \frac{\mu q}{\mu q + (1-\mu)\frac{1}{2}} \\ R(\omega,\omega^c) = \frac{\mu(1-q)}{\mu(1-q)+(1-\mu)\frac{1}{2}} \end{cases} \tag{1}$$

Notice that making correct decisions leads to a higher reputation than making wrong decisions:

$$R(\omega,\omega) - R(\omega,\omega^c) = \frac{2(2q-1)(1-\mu)\mu}{1-(2q-1)^2\mu^2} > 0 \tag{2}$$

From $q > \frac{1}{2}$, it is immediate to see that it is optimal for the good types to follow the signal.

Regarding the bad types, when the signal is aligned with their private interest, their problem is analogous to the problem of the good types and they prefer to follow the signal. However, when the signal contradicts the private interest, the utility of following the private interest is $1+\phi((1-q)R(\omega,\omega)+qR(\omega,\omega^c))$ and the utility of contradicting the private interest is $\phi((1-q)R(\omega,\omega^c)+qR(\omega,\omega))$. Thus, the bad types follow their private interest if and only if

$$1+\phi((1-q)R(\omega,\omega)+qR(\omega,\omega^c)) \geq \phi((1-q)R(\omega,\omega^c)+qR(\omega,\omega)) \tag{3}$$

By rearranging and substituting the expression of the reputation functions, we obtain the following:

$$\phi \leq \phi_{ND}(q,\mu) := \frac{1-(2q-1)^2\mu^2}{2(2q-1)^2(1-\mu)\mu} \tag{4}$$

Thus, when $\phi \leq \phi_{ND}(q,\mu)$, there exists an equilibrium in which the good types follow

their signal, and the bad types follow their private interest.

(ii) Next, suppose that the good types follow their signal and the bad types mix between following their signal and their private interest. In particular, let $\alpha(0,\beta,\beta)=\beta$ and $Pr(\alpha(0,\beta,\beta^c)=\beta^c)=x\in(0,1)$. In this equilibrium, $x<1$ because for $x=1$, we have $R(\omega,d)=\mu$, and without reputational incentives, the bad types strictly prefer to follow their private interest over following the signal. The consistency of the beliefs requires the following:

$$\begin{cases} R(\omega,\omega)=\frac{\mu q}{\mu q+(1-\mu)\frac{1}{2}(q+xq+(1-x)(1-q))} \\ R(\omega,\omega^c)=\frac{\mu(1-q)}{\mu(1-q)+(1-\mu)\frac{1}{2}((1-q)+x(1-q)+(1-x)q)} \end{cases} \tag{5}$$

The bad types need to be indifferent between following the signal and the private interest when $s=\beta^c$. The x that renders a bad type indifferent between following the signal and following the private interest when $s=\beta^c$ is as follows:

$$x=\frac{\mu(2q-1)(\phi-1)-\sqrt{1-\mu\phi(1-2q)^2(2-\mu\phi)}}{(1-\mu)(2q-1)} \tag{6}$$

Additionally, $x\in(0,1)$ if and only if $\phi>\phi_{ND}(q,\mu)$.

Proof of Corollary 1 The good types always follow the signal independently of ϕ. When $\phi\leq\phi_{ND}$, the bad types always follow their private interest, and the probability of making a correct decision is constant for all $\phi\leq\phi_{ND}$. When $\phi>\phi_{ND}$, the probability that the bad type follows the signal when it does not coincide with the interest increases with ϕ. Therefore, the probability of making a correct decision is also increasing in ϕ.

Proof of Lemma 1 We will prove that for every (β,s), $Pr(\alpha(0,\beta,s)=\beta)\geq Pr(\alpha(1,\beta,s)=\beta)$. Let $u(\theta,s,d)$ be the expected utility of a decision maker of type θ and private interest β

when he observes signal s and makes decision d.

$$\begin{aligned} u(1,\beta,\beta) - u(1,\beta,\beta^c) &= 2q - 1 + \phi\left(q(R(\beta,\beta) - R(\beta,\beta^c)) + (1-q)(R(\beta^c,\beta) - R(\beta^c,\beta^c)\right) \\ &< 1 + \phi\left(q(R(\beta,\beta) - R(\beta,\beta^c)) + (1-q)(R(\beta^c,\beta) - R(\beta^c,\beta^c)\right. \\ &= u(0,\beta,\beta) - u(0,\beta,\beta^c) \end{aligned} \tag{7}$$

$$\begin{aligned} u(1,\beta^c,\beta) - u(1,\beta^c,\beta^c) &= 1 - 2q + \phi\left((1-q)(R(\beta,\beta) - R(\beta,\beta^c)) + q(R(\beta^c,\beta) - R(\beta^c,\beta^c)\right) \\ &< 1 + \phi\left((1-q)(R(\beta,\beta) - R(\beta,\beta^c)) + q(R(\beta^c,\beta) - R(\beta^c,\beta^c)\right. \\ &= u(0,\beta^c,\beta) - u(0,\beta^c,\beta^c) \end{aligned} \tag{8}$$

Thus, since the bad types always find it more profitable to follow the private interest than the good types, if the good types follow the private interest, the bad types do also follow it. In particular, if $x^s_\theta = Pr(\alpha(\theta,\beta,s) = \beta)$, then $x^s_1 \leq x^s_0$.

$$\begin{cases} R(\beta,\beta,\beta) = \frac{\mu(qx_1^\beta + (1-q)x_1^{\beta^c})}{\mu(qx_1^\beta + (1-q)x_1^{\beta^c}) + (1-\mu)(qx_0^\beta + (1-q)x_0^{\beta^c})} \\ R(\beta,\beta,\beta^c) = \frac{\mu(q(1-x_1^\beta) + (1-q)(1-x_1^{\beta^c}))}{\mu(q(1-x_1^\beta) + (1-q)(1-x_1^{\beta^c})) + (1-\mu)(q(1-x_0^\beta) + (1-q)(1-x_0^{\beta^c}))} \end{cases} \tag{9}$$

By rearranging, we obtain $R(\beta,\beta,\beta) \geq R(\beta,\beta,\beta^c)$ if and only if

$$\frac{qx_1^\beta + (1-q)x_1^{\beta^c}}{qx_0^\beta + (1-q)x_0^{\beta^c}} \geq \frac{q(1-x_1^\beta) + (1-q)(1-x_1^{\beta^c})}{q(1-x_0^\beta) + (1-q)(1-x_0^{\beta^c})} \tag{10}$$

The left-hand-side is smaller than 1, and the right-hand-side is larger. Therefore, $R(\beta,\beta,\beta) \leq R(\beta,\beta,\beta^c)$ must hold. Analogously, we obtain $R(\beta,\beta^c,\beta) \leq R(\beta,\beta^c,\beta^c)$. Finally, $R(\beta,\beta^c,\beta) \leq \mu \leq R(\beta,\beta^c,\beta^c)$.

Proof of Proposition 2 Suppose that $\alpha(1,\beta,s)=s$ and $\alpha(0,\beta,s)=\beta$. The consistency of the beliefs requires the following:

$$\begin{cases} R(\beta,\omega,\beta^c)=1 \\ R(\beta,\beta,\beta)=\frac{\mu q}{\mu q+(1-\mu)} \\ R(\beta,\beta^c,\beta)=\frac{\mu(1-q)}{\mu(1-q)+(1-\mu)} \end{cases} \tag{11}$$

The incentive compatibility condition for the good types is as follows:

$$\phi \leq \phi_g = \frac{(2q-1)\left(\mu+\mu^2(q-1)q-1\right)}{(1-\mu)(\mu(2(q-1)q+1)-1)} \tag{12}$$

The incentive compatibility condition for the bad types is as follows:

$$\phi \leq \phi_b = \frac{1-\mu-\mu^2(q-1)q}{(1-\mu)(2\mu(q-1)q+1)} \tag{13}$$

Let $\phi_D := \min(\phi_b,\phi_g)$. This equilibrium exists if and only if $\phi \leq \phi_D$. Finally, from $\mu \in (0,1)$ and $q \in (.5,1)$, both ϕ_b and ϕ_g are strictly positive.

Finally, we want to prove that $\phi_b < \phi_{ND}$.

$$\begin{aligned} sgn\left(\phi_b-\phi_{ND}\right) &= sgn\left(\frac{\mu(\mu+2(2\mu-3)(q-1)q-2)+1}{2(\mu-1)\mu(1-2q)^2(2\mu(q-1)q+1)}\right) \\ &= -sgn\left(\mu(\mu+2(2\mu-3)(q-1)q-2)+1\right)=-1 \end{aligned} \tag{14}$$

Proof of Lemma 2

$$\phi_g-\phi_b = \frac{\left(\mu+\mu^2(q-1)q-1\right)\left(\mu+2q(2\mu(q-1)q+1)-2\right)}{(\mu-1)(2\mu(q-1)q+1)(\mu(2(q-1)q+1)-1)} \tag{15}$$

By rearranging, we obtain that $\phi_g-\phi_b>0$ if and only if

$$\mu > \bar{\mu} = \frac{2-2q}{4(q-1)q^2+1} \tag{16}$$

In addition, for $q \in \left(\frac{1}{2},1\right)$, $\frac{\partial\bar{\mu}}{\partial q}<0$.

Proof of Proposition 3 Suppose that there is an equilibrium in which the good types follow the signal and the strategy of the bad types is such that $\alpha(0,\beta,\beta)=1$ and $Pr(\alpha(0,\beta,\beta^c)=\beta)=x\in(0,1)$. Let $\Delta R(s)$ be the reputational gains from contradicting the private interest when the signal is s. In particular,

$$\begin{cases}\Delta R(\beta)=q(R(\beta,\beta,\beta^c)-R(\beta,\beta,\beta))+(1-q)(R(\beta,\beta^c,\beta^c)-R(\beta,\beta^c,\beta))\\ \Delta R(\beta^c)=q(R(\beta,\beta^c,\beta^c)-R(\beta,\beta^c,\beta))+(1-q)(R(\beta,\beta,\beta^c)-R(\beta,\beta,\beta))\end{cases}\tag{17}$$

In both signals, the good types contradict their private interest with a higher probability; therefore $R(\beta,\omega,\beta^c)\geq R(\beta,\omega,\beta)$ and $\Delta R(s)\geq 0$. Thus, when $s=\beta^c$, the good types follow the signal because it increases both their reputation and present utility, and for the good types, we must only show that they follow the signal when $s=\beta$. Regarding the bad types, if the good types follow the signal when $s=\beta$, the bad types also follow the signal. Therefore, we only must verify that when $s=\beta^c$, there exists x such that they are indifferent.

Regarding the good types, from $\phi<\phi_g$, we know that when $x=0$, the good types strictly prefer to follow the private interest if $s=\beta$. Here, notice that $\frac{\partial}{\partial x}R(\beta,\omega,\beta)>0$ and $\frac{\partial}{\partial x}R(\beta,\omega,\beta^c)<0$. Therefore, $\frac{\partial}{\partial x}\Delta R(s)<0$, i.e., the reputational gains of contradicting the private interest decrease with x; thus, the good types strictly prefer to follow the private interest when $s=\beta$ for any $x\in(0,1)$.

Finally, regarding the bad types, when $x=0$, their strategy coincides with the strategy of the good types, $\Delta R(\beta^c)=0$, and they have incentives to follow their private interest. In contrast, when $x=1$, only the good types contradict the private interest, $\Delta R(\beta^c)=1$, and they have incentives to contradict their private interest if $\phi>\phi_b$. By continuity, there exists $x\in(0,1)$ such that the bad types are indifferent.

Regarding the probability of making a correct decision, suppose that the bad types follow the signal with identical probability with and without disclosure. The expected reputation

of following the signal when $s = \beta^c$ without disclosure is as follows:

$$qR(\omega, \omega) + (1 - q)R(\omega, \omega^c) \tag{18}$$

The expected reputation of following the signal when $s = \beta^c$ with disclosure is as follows:

$$qR(\beta, \beta^c, \beta^c) + (1 - q)R(\beta, \beta, \beta^c) \tag{19}$$

However, if the bad types follow the signal with an identical probability with and without disclosure, $R(\beta, \beta^c, \beta^c) > R(\omega, \omega)$, and $R(\beta, \beta, \beta^c) > R(\omega, \omega^c)$. We observed that the expected reputation of following the signal when $s = \beta^c$ is higher with disclosure than without. Analogously, the expected reputation of contradicting the signal when $s = \beta^c$ is lower with disclosure than without. Therefore, the bad types follow the signal with a higher probability with disclosure than without.

Proof of Proposition 4

(i) Suppose that $\alpha(1, \beta, s) = \beta^c$ and $\alpha(0, \beta, s) = \beta$. The consistency of the beliefs requires $R(\beta, \omega, d)$ to be such that $R(\beta, \omega, \beta) = 0$ and $R(\beta, \omega, \beta^c) = 1$. With these reputational incentives, the good types only experience a trade-off when the signal coincides with the private interest, and the condition to sustain their strategy in equilibrium is $q < 1 - q + \phi \leftrightarrow 2q - 1 < \phi$. The problem of the bad types does not depend on the signal that they receive, and the condition to sustain their strategy in equilibrium is $\phi < 1$. Therefore, this equilibrium exists when $2q - 1 < \phi < 1$.

(ii) Suppose that $\alpha(1, \beta, s) = \beta^c$, $Pr(\alpha(0, \beta, \beta) = \beta) = x$ and $Pr(\alpha(0, \beta, \beta^c) = \beta) = y$. First, $x = y$. Suppose that without loss of generality $x < y$, then $R(\beta, \beta, \beta^c) - R(\beta, \beta, \beta) < R(\beta, \beta^c, \beta^c) - R(\beta, \beta^c, \beta)$, and the bad types have higher reputational incentives to contradict the private interest when $s = \beta^c$, which contradicts the in-

different condition in both signals. Then, when $x = y$, the consistency of the beliefs requires that $R(\beta, \omega, \beta) = 0$, and $R(\beta, \omega, \beta^c) = \frac{\mu}{\mu+(1-x)(1-\mu)}$. The bad types are indifferent between following and contradicting the private interest if $1 = \phi\frac{\mu}{\mu+(1-x)(1-\mu)}$ and $x = 1 - \frac{\mu(\phi-1)}{1-\mu}$. The reputation incentives of contradicting the private interest are $R(\beta, \omega, \beta^c) = \frac{1}{\phi}$, and the good types always contradict the private interest. This equilibrium exists if $x \in (0, 1)$, which occurs when $1 < \phi < \frac{1}{\mu}$.

(iii) Suppose that $\alpha(1, \beta, s) = \beta^c$ and $\alpha(0, \beta, s) = \beta^c$. The consistency of the beliefs requires that $R(\beta, \omega, \beta^c) = \mu$, and the less restrictive of equilibrium belief is $R(\beta, \omega, \beta) = 0$. The binding incentive compatibility condition of the good types is $2q - 1 < \phi\mu$, and the incentive compatibility condition of the bad types is $1 < \phi\mu$. Therefore, this equilibrium exists when $1 < \frac{1}{\mu} < \phi$.

Proof of Lemma 3 Notice that in the Pandering Equilibrium, both types ignore the signal and the probability of making a correct decision is $1/2$. In the Nothing Changes Equilibrium, the good types follow the signal and the bad types ignore it and the probability of a correct decision is $\mu q + (1 - \mu)\frac{1}{2}) > \frac{1}{2}$. In the Disciplining Equilibrium, the good types follow the signal and the bad types follow the signal with some probability $x > 0$, therefore, the probability of a correct decision is $\mu q + (1 - \mu)(xq + (1 - x)\frac{1}{2})) > \mu q + (1 - \mu)\frac{1}{2})$.

Proof of Proposition 5 Follows directly from Proposition 2, Proposition 3, Proposition 4 and Lemma 3.

Proof of Proposition 6 Analogously to the case of $n = 2$, we compute ϕ_b and ϕ_g and the condition such that $\phi_b < \phi_g$ is given by $\mu > \mu(q, n)$, where:

$$\mu(q, n) := \frac{(n-1)^2 n(1-q)}{n(n(n(2(q-1)q+1) + q((q-5)q+3) - 2) + 2q + 1) - 1} \tag{20}$$

As we previously computed, $\mu(q,2) = \bar{\mu}(q)$, and $\mu(q,n)$ is decreasing in q and n.

Proof of Proposition 7 First, we will derive the equilibrium without disclosure for $n > 2$. Analogously to the case of $n = 2$, it is always sustainable for the good types to follow the signal and the bad types can either follow their private interest or mix between the private interest and the signal. Any other action would be suboptimal because it would reduce the expected reputation of following the signal without increasing the present utility. Similar to the case with $n = 2$, there exists a ϕ_{ND} such that when $\phi < \phi_{ND}$, the bad types follow the private interest; when $\phi \geq \phi_{ND}$, the bad types follow the signal with probability x and the private interest with probability $1 - x$. The expressions of ϕ_{ND} and x are as follows:

$$\begin{cases} \phi_{ND} = \frac{(n-1)(\mu(nq-1)+1)(n(\mu q-1)+1-\mu)}{(1-\mu)\mu n(nq-1)^2} \\ x = \frac{n^2-3n+2+\mu(nq-1)(n(\phi-2)+2)-n\sqrt{\mu^2\phi^2(nq-1)^2-2\mu(n-1)(2q-1)\phi(nq-1)+(n-1)^2}}{2(1-\mu)(n-1)(nq-1)} \end{cases} \tag{21}$$

The probability that a correct decision is made is as follows:

$$W_{ND} := \frac{n-1+\mu\phi(nq-1)-\sqrt{\mu^2\phi^2(nq-1)^2-2\mu(n-1)(2q-1)\phi(nq-1)+(n-1)^2}}{2(n-1)} \tag{22}$$

Regarding the Pandering Equilibrium, we focus on the case in which both types follow the signal when it does not coincide with their private interest and mix among the remaining actions when it does. We can impose the out-of-equilibrium belief $R(\omega = \omega', d = \beta) = 0$, and the consistency of the beliefs requires that $R(\omega = \omega', d \neq \beta) = \mu$. Here, the only binding incentive compatibility condition is for the bad types, which is simply as follows:

$$\phi_D := \frac{1}{\mu} < \phi \tag{23}$$

In terms of welfare, the probability that a correct decision is made is as follows:

$$W_D := \frac{(n-2)nq+1}{(n-1)n} \tag{24}$$

$\phi_D < \phi_{ND}$, i.e., when there is a mixed equilibrium without disclosure, there is also the disclosure Pandering equilibrium. Finally, $W_D > W_{ND}$ if and only if:

$$\phi > \hat{\phi} := \frac{(n^2(1-q)+n(2q-1)-1)\left((n-2)nq+1\right)}{\mu n(nq-1)^2} \tag{25}$$

Notice that $\frac{\partial\hat{\phi}}{\partial n}$, $\frac{\partial\hat{\phi}}{\partial\mu}$ and $\frac{\partial\hat{\phi}}{\partial q} < 0$.

Proof of Proposition 8 Notice that, if the good types follow the signal and bad types follow the private interest, the expression of the reputation function becomes:

$$\begin{cases} R(a,a) = \frac{\mu q}{\mu q+(1-\mu)p} \\ R(a,b) = \frac{\mu(1-q)}{\mu(1-q)+(1-\mu)(1-p)} \\ R(b,a) = \frac{\mu(1-q)}{\mu(1-q)+(1-\mu)p} \\ R(b,b) = \frac{\mu q}{\mu q+(1-\mu)(1-p)} \end{cases} \tag{26}$$

From $p, q > \frac{1}{2}$, we have that the lowest possible reputation is achieved when $d = a$ and $\omega = b$ and the highest reputation is achieved when $d = b$ and $\omega = b$. However, the relationship between $R(a,a)$ and $R(a,b)$ depends on p and q as follows:

$$R(a,a) - R(a,b) = \frac{(1-\mu)\mu(q-p)}{(1-p+\mu(p-q))(p(1-\mu)+\mu q)} \tag{27}$$

Since the denominator is always positive, we have that $R(a,a) \geq R(a,b)$ if and only if $q \geq p$. The utility of a good type when he follows the signal is $q+\phi(qR(s,s)+(1-q)R(s^c,s))$ and his payoff when he contradicts it is $1-q+\phi(qR(s,s^c)+(1-q)R(s^c,s^c))$. The condition

for the good types to follow the signal is:

$$2q - 1 + \phi(q(R(s,s) - R(s,s^c)) + (1-q)(R(s^c,s) - R(s^c,s^c))) \tag{28}$$

When $s = b$, the condition becomes

$$\begin{aligned} &(2q-1) + \phi(q(R(b,b) - R(b,a)) + (1-q)(R(a,b) - R(a,a))) \\ &> (2q-1) + \phi(q(R(a,b) - R(b,a)) + (1-q)(R(b,b) - R(a,a))) > 0 \end{aligned} \tag{29}$$

When $s = a$, the condition becomes

$$(2q-1) + \phi(q(R(a,a) - R(a,b)) + (1-q)(R(b,a) - R(b,b))) \tag{30}$$

Clearly, when $q < p$, the second term is always negative indicating that with a sufficiently large ϕ, the good types will contradict their signal. When $q > p$, $R(a,a) > R(a,b)$ indicating that there exist a $q' \in (p,q)$ such that the second term is positive when $q < q'$.

Once we have shown that the good types follow the signal it is immediate to see that the problem of the bad types is qualitatively equal to the problem with $p = \frac{1}{2}$. However, notice that when $\phi = \phi'_{ND}(q,\mu)$, only bad types with $\beta = a$ start mixing between following their interest and their signal and bad types with $\beta = b$ requires a larger ϕ.

Proof of Proposition 9 Since the good types only maximize their reputation and do not care about the decision, they simply make the decision that gives them higher reputation. We will distinguish between two situations. If R is constant, the bad types always follow the private interest. To satisfy the beliefs, the good types should also follow their private interest. However, this PBE is not stable. Therefore, the reputation from contradicting the private interest must be larger than the reputation from following it; then, the good types

always contradict their private interest.

Proof of Corollary 2 We need to prove that $2q-1<\phi_j$ for $j\in\{0,1\}$.

Notice that $\phi_b<2q-1$ if and only if:

$$\frac{\mu(2q-1)(1-\mu+(3\mu-2)(1-q)q)}{(1-\mu)(1-\mu(2(q-1)q+1))}>0 \tag{31}$$

which always holds because the denominator and numerator are positive. And $\phi_g<2q-1$ if and only if:

$$1<\frac{(1-\mu(1-q))(1-\mu q)}{(1-\mu)(1-\mu(1-2(1-q)q))} \tag{32}$$

which is always satisfied for $\mu\in(0,1)$ and $q\in(.5,1)$.