Reputation and perverse transparency under two concerns*

Ascensión Andina-Díaz† and José A. García-Martínez‡

October 11, 2022

Abstract

This paper proposes a novel argument to explain why transparency can reduce information transmission in decision making processes. We build on the literature of career concerns and consider an agent (judge) with two dimensions of private information, ability and bias, and two reputational concerns, one for ability and the other for bias. We show that, in equilibrium, the agent can take actions that go against her private information so as to signal she is unbiased, and that this incentive can be exacerbated when the consequences of the agent's action are more likely to be learned. We provide necessary and sufficient conditions for this perverse effect of transparency to emerge and discuss variations of the model. As an application, the result provides a rationale for recent empirical evidence showing that media coverage of judicial cases increases sentence length.

Keywords: Ability; bias; career concerns; transparency; perverse effects

JEL: D82, D83, K40

1 Introduction

The literature on career concerns distinguishes two types of transparency, transparency on actions and transparency on consequences, each with different effects on information transmission and social welfare. Ever since Prat (2005), we know that *transparency on actions*, i.e., providing the principal information on an agent's action, can induce a career concerned agent to disregard an informative private signal about the state of the world, which reduces the principal's expected welfare. The posterior work by Fox and Van Weelden (2012) shows that *transparency on consequences*, i.e., providing the principal information on the state of the world (hence, on the consequences of the agent's action), can also reduce the principal's

*We gratefully acknowledge financial support from the Ministerio de Ciencia, Innovación y Universidades (MCIU/AEI/FEDER, UE) through projects RTI2018-097620-B-I00 and PGC2018-097965-B-I00, the Junta de Andalucía through project P18-FR-3840, and the Universidad de Málaga through project UMA18-FEDERJA-243. The usual disclaimer applies.

†Dpto. Teoría e Historia Económica, Universidad de Málaga, Spain. E-mail: aandina@uma.es

‡Dpto. Estudios Económicos y Financieros, Universidad Miguel Hernández, Spain. E-mail: jose.garciam@umh.es

expected welfare. None of these works, however, expose an argument why transparency on consequences can induce an agent to go against an informative (i.e., decision relevant) private signal.

In the present work, we question this last idea and argue that an increase in transparency on consequences *may* induce a career concerned agent to disregard an informative private signal more often and to take the (ex-ante) incorrect action with a higher probability. By *correct* action, we mean the action that matches the state of the world; the action is incorrect otherwise. This result, which is new in the literature, constitutes our main contribution. As argued later, key to the result is to consider an agent with two dimensions of private information: the standard *ability* dimension and a second *bias* dimension; and two reputational concerns: one for ability, the other for bias.

The new perverse effect of transparency that we identify in this paper speaks to a variety of decision making contexts, from political and financial decision making situations, to passing-sentence decisions in the judicial context. For illustrative purposes, we frame the paper into the latter scenario. We do so because the judicial system provides recent evidence of transparency on consequences (media coverage of judicial cases) influencing decision making (sentencing practices).

The first piece of evidence in this respect comes from a recent paper by Lim et al. (2015), showing that for the US state court judges, period $1986-2006$, newspaper coverage of judicial cases significantly increased sentence length. Their estimates show that eight more newspaper articles per judge per year in a judicial district increases average sentence length for violent crimes by nonpartisan elected judges by 3.4%, which means a sentence of about 5.7 months longer. They show that the effect is only significant for nonpartisan elected judges (nor for partisan judges or appointed judges); and that it is still significant after controlling for more severe crimes attracting more media attention. Talking about the possible mechanisms behind the result, Lim et al. (2015) argue that if voters have a preference for harsher sentences (something supported by numerous surveys, as referred by the authors), then "an increase in voter information about candidates may also induce incumbents to avoid decisions that are disliked by voters". Though at first sight the argument is correct, if more information also facilitates judges' accountability (better monitoring of the quality of sentences), it is not clear why transparency might induce judges to go more often for sentences that might be proven wrong. Our paper proposes a rationale to explain this observation.

Outside the US, anecdotal evidence suggests similar effects of media coverage. An example (not without controversy) is the "*Procés*", the trial judging former leaders of the Catalan independence movement for their involvement in the failed 2017 secession attempt in Catalonia. The defendants, accused of sedition and misuse of public funds during the preparation and declaration of the independence of Catalonia, received up to 13 years in-prison sentences. The lengthly sentences sparked extensive and hot debate, ending with the pardons granted by the Spanish government in June 2021, by which the nine Catalan separatists jailed were released from prison. Although we agree it is difficult to assess and measure the effect of international attention and media coverage of the case, in light of our results we cannot but

wonder whether length sentences would have been different in the absence of media coverage. For more stories, we refer the reader to "*Caso Pantoja*" and "*Nut rage*".[1]

The model we propose builds on the literature of career concerns and considers an agent (judge) that has to sentence a convicted offender. There are two possible sentences: a lenient sentence and a harsh sentence. Accordingly, there are two possible states of the world: one in which the appropriate sentence is lenient and the other one in which the appropriate sentence is harsh. The judge has two dimensions of private information. Specifically, the judge is informed about her ability to interpret the law - she can be either wise or normal - and her bias in passing a sentence - she can be either biased (in favor of the lenient sentence) or unbiased. Prior to taking an action, i.e., passing sentence, the judge receives a private signal about the correct interpretation of the law; with the quality of the signal depending on the judge's ability to interpret the law. We consider that the judge maximizes her expected reputational rent, which is a function of the principal's beliefs about the judge's type in both the ability dimension (the higher the ability, the higher the reputational rent) and the bias dimension (the smaller the bias, the higher the reputational rent). The principal is the audience that evaluates the judge's performance. We consider that the principal is concerned about justice and receives a positive payoff when the sentence matches the state of the world; a zero payoff otherwise. The principal observes the sentence and, with some probability, the true state of the world. We refer to this probability as the degree of transparency on consequences (or media coverage).

Our analysis shows that the equilibrium behavior of the agent depends on how numerous biased agents in the population are. In the extremes, when the probability of being biased is sufficiently low (high), the agent always follows (contradicts) a lenient signal, trying to show ability (unbiasedness) (Proposition 3). In contrast, when the probability that the agent is biased is neither high nor too low, the equilibrium behavior of the agent depends on the degree of transparency on consequences. In particular, in this case we obtain that the incentive of the agent to go against a lenient signal -so as to signal she is unbiased- can be exacerbated when transparency on consequences increases (Proposition 4). This result, new in the literature, gives a rational to explain why media coverage of judicial cases may lead to harsher sentencing

[1]The story about '*Caso Pantoja*" goes back to 2013, when Isabel Pantoja, an internationally renowned Spanish *copla* singer, accused of laundering money proceeding from public funds profited by her partner during Spain's property boom, was sentenced to two-year in prison. Though initially she was not expected to go to jail (no prior convictions and minimum prison-sentence, in which case most defendants avoid jail), the court rejected all her appeals and she finally entered prison. In the sentence, the court explicitly said "*it was making an example of Pantoja*" at a time of economic crisis and rampant corruption by Spanish public figures. See "Una sentencia ejemplarizante", *El País*, November 8, 2014. The story about "*Nut rage*" goes back to John F. Kennedy Airport NY, December 2014, when Cho Hyun-ah, executive of Korean Air and daughter of the company's chief, required the plane to return to gate before takeoff, after she was served nuts in a bag. The apparently minor air rage sparked public outrage in South Korea, where *"the incident was seen as emblematic of a generation of spoilt and arrogant offsprint of owners of family-run conglomerates that dominate the economy"*. Two months later, Cho Hyun-ah was sentenced by a Seoul court to one-year prison for obstructing aviation safety. See " 'Nut rage' trial: Korean Air executive treated crew 'like slaves' ", *The Guardian*, February 2, 2015.

practices. It proposes a signalling mechanism, consistent with Lim et al. (2015), which further requires the existence of both biased judges and judges with career concerns. Existing literature supports both premises.[2]

Then, we explore the forces behind this perverse effect of transparency and provide necessary and sufficient conditions for the result to hold (Proposition 5). We show that we need variation in both ability and bias. The intuition for the result is as follows. Suppose a judge that only cares about appearing unbiased. This itself encourages her to pass harsher sentences. With increased transparency, this incentive increases as a mistaken lenient sentence will be attributed to bias, whereas a mistaken harsh sentence will not. Thus, an increase in transparency strengthens the incentive for harsh actions. This sole effect, however, is not sufficient for the perverse effect to appear, as a judge with a sole concern for bias will never pass lenient sentences. If the judge has rather an additional concern for ability, the incentive to pass lenient sentences appears, as wise judges do pass lenient sentences. This allows the perverse effect of transparency to appear.

The rest of the paper is organized as follows. In Section 2 we review the related literature and in Section 3 we present the model. Section 4 contains the analysis and the main results, which are structured in three parts: preliminaries, main result, and forces behind the perverse effect. Section 5 discusses variations of the model and extends results to alternative frameworks. Finally, Section 6 concludes. All proofs are in the Appendix.

2 Literature review

The main topic in the paper is the effect of transparency. In this sense, our work is particularly related to three papers in the literature. Prat (2005) shows that transparency on actions may induce a career-concerned agent to disregard an informative private signal, which reduces the principal's expected welfare.[3] A crucial idea in Prat (2005) is the relative smartness of the realizations of the agent's signal, i.e., how similar or different the posteriors on the agent's type are for each realization of the agent's signal. The author shows that when one realization is much smarter than the other, and the agent's action is informative of the signal, then the agent has an incentive to take the action that corresponds to the smart realization of the signal. This incentive increases in transparency on actions but decreases

[2]For evidence showing that people perceive justice as biased (lenient), we refer to the study conducted by Princeton Survey Research Associates International in 2006 for the US National Center for State Courts, finding that *"More Americans are inclined to say sentencing practices in their state generally are too lenient than believe they are too harsh (*48% *vs.* 8%*)"*. Evidence for Australia, Canada, Germany, Spain or UK shows similar patterns. For judges with career concerns, see Miceli and Cosgel (1994), Levy (2005), Iossa and Jullien (2012), Cohen et al. (2015), and Ferrer (2016).

[3]He considers a model in which the principal can always observe his payoff and compares a situation where the action (hence the state of the world) is observed or not. In contrast to this, we consider a model in which the principal always observes the agent's action and compare a situation where the state of the world (hence the principal's payoff) is observed or not.

in transparency on consequences. Hence, the main contribution of his work is to show that while transparency on consequences is beneficial to the principal, transparency on actions can be detrimental to the principal's welfare. Fox and Van Weelden (2012) consider a game in which transparency on consequences can have detrimental effects on the principal's welfare. In line with Prat (2005), Fox and Van Weelden (2012) consider a game in which transparency on consequences induces the agent to take the action that most likely corresponds to the state of the world. The key assumption in Fox and Van Weelden (2012) is, however, that the costs to the principal of the agent's mistakes are asymmetric across the states, in the sense that some mistakes are more costly than others. Under this set-up, the authors show that an increase in the probability that the principal learns the consequences of the agent's action may reduce the principal's expected welfare. None of these papers, however, identify conditions under which transparency on consequences can induce an agent to go against an informative private signal. This is our contribution. Finally, Canes-Wrone et al. (2001) consider a game in which the probability that there is transparency on consequences depends on the agent's action, i.e., the probability of observing the state is asymmetric across actions. They show that when the asymmetry in the uncertainty resolution is sufficiently high, the agent faces an incentive to disregard some informative private signals, which may reduce the principal's welfare.[4] It is interesting to note that these papers, as well as our paper, play on an asymmetry. The difference, however, between these models and ours is the source of the asymmetry.

The literature on career concerns has also identified other cases in which transparency may be detrimental to the principal's welfare. See Li and Madarász (2008) and Bourjade and Jullien (2011) for models in which an increase in transparency on the preferences of the agent may reduce the principal's welfare. The present paper is also related to the works by Holmström (1999) and Dewatripont et al. (1999), who identify conditions under which more information may induce the agent to exert less effort; Levy (2007), Sibert (2003), and Gersbach and Hahn (2008), who show that secretive procedures may be better than transparent mechanisms in decision making in committees; and to Morris (2001), Ottaviani and Sørensen (2001), Hörner (2002), and Ely and Välimäki (2003), who focus on the perverse effect of reputation.

Finally, the present work considers an agent with two dimensions of private information and two concerns. This relates our work to other papers in the literature. See Austen-Smith and Fryer (2005), Bénabou and Tirole (2006), Esteban and Ray (2006), Bagwell (2007), and Frankel and Kartik (2019) for papers considering information transmission in the presence of an agent with two-dimensional types and Austen-Smith and Fryer (2005), Fox and Shotts (2009), Bar-Isaac and Deb (2014), and Feller and Schäfer (2020) for papers considering information transmission in the presence of two audiences, i.e., two concerns. None of these papers analyze the effect of transparency on the quality of the decision-making process.

[4]See Levy (2005), Leaver (2009), Liu and Sanyal (2012) and Andina-Díaz and García-Martínez (2020) for other papers with asymmetric monitoring.

3 The model

We consider a principal-agent model in which the agent has two reputational concerns: ability and bias. For expositional purposes, we will refer to the agent (she) as the judge, and to the principal (he) as the general public, i.e., the evaluator that the judge seeks to impress.

The judge has to sentence a convicted offender. There are two possible sentences: a lenient sentence, denoted by $\hat{l}$, and a harsh sentence, denoted by $\hat{h}$. Accordingly, there are two states of the world, $\omega \in \{L, H\}$. We assume that the lenient sentence $\hat{l}$ is the appropriate sentence in state L and the harsh sentence $\hat{h}$ is the appropriate one in state H. Each state occurs with equal probability.[5]

Prior to passing sentence, the judge receives a private signal $s \in \{l, h\}$ about the correct interpretation of the law and takes an action $a \in \{\hat{l}, \hat{h}\}$, which is refereed to as the sentence. The judge has two dimensions of private information: her *ability* to interpret the law, t, and her *bias* (or non-bias) in passing sentence, b. Judges know both their own ability and bias. According to ability, the judge can be either wise W or normal N, i.e., $t \in \{W, N\}$. We assume that a wise type judge receives a signal that perfectly reveals the state of the world (it has quality 1), whereas a normal type judge receives an imperfect but informative (i.e., decision relevant) signal of quality γ, with $\gamma = P(l/L) = P(h/H)$ and $\gamma \in (\frac{1}{2}, 1)$. According to bias, the judge can be either biased B or unbiased $\bar{B}$, i.e., $b \in \{B, \bar{B}\}$. For simplicity, in the main body of the paper we assume that a biased type judge is soft on crime and always passes the lenient sentence.[6] In contrast, unbiased judges can pass both lenient and harsh sentences.

This set-up produces four types of judges that can be subsumed under just three. With some abuse of terminology, we denote the biased type judge by B, the unbiased-wise type judge by W (hereafter, referred to as the wise type judge), and the unbiased-normal type judge by N (hereafter, referred to as the normal type judge). Let α_B, α_W and α_N denote the prior probability that the judge is a biased, wise and normal type, respectively, with $\alpha_t > 0$ for $t \in \{B, W, N\}$ and $\sum_t \alpha_t = 1$. Since biased type judges always take the lenient action, the focus of the paper is on the behavior of the unbiased types, i.e., the strategic types. We denote the strategy of a judge by $\sigma_t(s) \in [0, 1]$, describing the probability that a judge of type $t \in \{W, N\}$ takes action $a = \hat{l}$ after signal $s \in \{l, h\}$.

Apart from observing the judge's action and before forming a belief about the type of the judge, the public may observe the state of the world. We consider that this occurs with probability $\mu \in [0, 1]$, and refer to μ as the probability that there is *transparency on consequences*. For expositional purposes, we consider that this probability μ is increasing in media coverage of the judicial case, and so talk of media coverage and transparency on consequences indistinctively. We denote by $X \in \{0, L, H\}$ the feedback received by the public, with L (alternatively, H) indicating that the public learns that the state is L (alternatively, H), and 0 indicating that there is no feedback, i.e., the public in normal partisan elections

[5] This assumption guarantees that herding motives are not behind our results. We relax this assumption in Section 5 and show that our main results are robust to it.

[6] In Section 5 we relax this assumption and consider a strategic biased type judge.

after being nominated by political parties receives no additional information.

The general public observes sentence $a \in \{\hat{l}, \hat{h}\}$ and feedback $X \in \{0, L, H\}$ and, based on this information, updates his belief about the type of the judge. Let $\lambda_W(a, X)$, $\lambda_N(a, X)$ and $\lambda_B(a, X)$ denote the public's belief that the judge is a wise, normal and biased type, respectively.

We consider that the general public is concerned about justice and receives utility 1 when the sentence matches the state, i.e., $a = \omega$; it receives utility 0 otherwise. We refer to action $a = \omega$ as the *correct* action or correct sentence, in the sense that it is the action that maximizes the (expected) welfare of the public. Note that because the judge's signal is informative, i.e., $\gamma > 1/2$, the public's expected welfare is maximized when the judge follows her signal, i.e., $a = s$. Based on this, we will say that transparency on consequences is detrimental to the public, i.e., it has a perverse effect, when it refrains the judge from following her signal and induces her to take action $a \neq s$.

Let R_θ be the reputational rent that the general public assigns to each possible type $\theta \in \{W, N, B\}$. We consider that the judge has career concerns and seeks to maximize her expected reputational rent. For a given action $a \in \{\hat{l}, \hat{h}\}$ and feedback $X \in \{0, L, H\}$, we define the expected reputational rent $\Pi(a, X)$ as follows:

$$\Pi(a, X) = \lambda_W(a, X)R_W + \lambda_N(a, X)R_N + \lambda_B(a, X)R_B.$$

Assuming $R_W > R_N > R_B$ and $R_B = 0$, and writing $\lambda_{\bar{B}}(a, X) = 1 - \lambda_B(a, X)$ as the belief of the public that the judge is unbiased, the objective function of the judge, representing her expected reputational rent, simplifies to the following:[7]

$$\Pi(a, X) = \eta\lambda_W(a, X) + \beta\lambda_{\bar{B}}(a, X), \tag{1}$$

with $\eta = R_W - R_N$ and $\beta = R_N$; hence, $\eta, \beta > 0$.

Equation (1) describes the objective function of the judge as a linear combination of two terms: one term represents the judge's expected payoff for being perceived as wise, the other term represents her expected payoff for being perceived as unbiased. It is worth noting that each of these terms corresponds to a reputational concern for each of the two dimensions of heterogeneity that we consider, ability and bias, where η and β are the weights that the general public assigns to each concern.

Prior to the analysis of the game, let us introduce some additional definitions. Given the probability $\mu \in [0, 1]$ that there is transparency on consequences, let $\Pi^\mu_{t,s}(a)$ be the expected payoff to a judge of type $t \in \{W, N\}$ when she receives signal $s \in \{l, h\}$ and passes sentence $a \in \{\hat{l}, \hat{h}\}$. Thus:

$$\Pi^\mu_{t,s}(a) = (1 - \mu)\Pi^0_{t,s}(a) + \mu\Pi^1_{t,s}(a), \tag{2}$$

where $\Pi^0_{t,s}(a)$ denotes the payoff to a judge of type t who receives signal s and takes action a when the state ω is not publicly observed (hence, $X = 0$), and $\Pi^1_{t,s}(a)$ denotes her expected payoff when the state

[7] For exposition purposes, the results in the main body of the paper consider the payoff function described in (1). However, all of these results, except for Propositions 4 and 5, are more general and hold for any function $f(\cdot)$ that is increasing in the public's beliefs $\lambda_W(a, X)$ and $\lambda_{\bar{B}}(a, X)$. This is indicated in the Appendix.

is publicly observed (hence, $X \neq 0$). By substituting, expression (2) is as follows:

$$\Pi^{\mu}_{t,s}(a) = (1-\mu)\Pi(a,0) + \mu[P(\omega = L|s;t)\Pi(a,L) + P(\omega = H|s;t)\Pi(a,H)]. \tag{3}$$

Additionally, let $\Delta^{\mu}_{t,s}$ be the expected gain to a judge of type t for passing sentence $\hat{h}$ rather than $\hat{l}$ after signal s when the level of transparency is μ:

$$\Delta^{\mu}_{t,s} = \Pi^{\mu}_{t,s}(\hat{h}) - \Pi^{\mu}_{t,s}(\hat{l}), \tag{4}$$

which can be rewritten as follows:

$$\Delta^{\mu}_{t,s} = (1-\mu)\Delta^{0}_{t,s} + \mu\Delta^{1}_{t,s}, \tag{5}$$

with $\Delta^{0}_{t,s} = \Pi^{0}_{t,s}(\hat{h}) - \Pi^{0}_{t,s}(\hat{l})$ and $\Delta^{1}_{t,s} = \Pi^{1}_{t,s}(\hat{h}) - \Pi^{1}_{t,s}(\hat{l})$.[8]

Our equilibrium concept is Perfect Bayesian Equilibrium. We will denote an equilibrium strategy by $\sigma^{\mu *} = (\sigma^{\mu *}_{W}, \sigma^{\mu *}_{N})$, with $\sigma^{\mu *}_{W} = (\sigma^{\mu}_{W}(l)^{*}, \sigma^{\mu}_{W}(h)^{*})$ and $\sigma^{\mu *}_{N} = (\sigma^{\mu}_{N}(l)^{*}, \sigma^{\mu}_{N}(h)^{*})$.

4 Results

For expositional purposes, the presentation of the results in this section considers two simplifications. Firstly, we skip the limit case of $\mu = 0$ and focus instead on the general and more interesting case of $\mu > 0$. Briefly, when $\mu = 0$, there is multiplicity of equilibria, which introduces many particularities in the results and a need for equilibrium selection. To facilitate the reading process, we relegate the results of the case $\mu = 0$ to the Appendix.[9] Secondly, we restrict our analysis to non-perverse equilibria. We say that an equilibrium is *non-perverse* if for a given μ, $\sigma^{\mu}_{t}(l)^{*} \geq \sigma^{\mu}_{t}(h)^{*}$ for all $t \in \{W, N\}$, i.e., the two strategic types use non-perverse strategies.

Other concepts that we use are the following. For a given type $t \in \{W, N\}$, we say that an equilibrium strategy is *informative* if $\sigma^{\mu}_{t}(l)^{*} \neq \sigma^{\mu}_{t}(h)^{*}$, i.e., the strategy is signal dependent, which means that the agent uses the information she has about the state of the world; and non-informative otherwise. An equilibrium will be informative when the two strategic types use informative strategies. Additionally, for a given type $t \in \{W, N\}$, an informative equilibrium strategy is *honest* if $(\sigma^{\mu}_{t}(l)^{*}, \sigma^{\mu}_{t}(h)^{*}) = (1, 0)$.

4.1 Preliminaries

The first result characterizes the behavior of the wise type judge. It distinguishes two cases, according to whether the strategy of the normal type judge is informative or not.

[8] Note that $\Delta^{\mu}_{t,s}$ is a function of the strategy profile (σ_W, σ_N), with $\sigma_W = (\sigma_W(l), \sigma_W(h))$ and $\sigma_N = (\sigma_N(l), \sigma_N(h))$. For the sake of simplicity, this dependence is sometimes omitted.

[9] To see the multiplicity, note that without transparency, the judge's expected gain for passing sentence $\hat{h}$ rather than $\hat{l}$ is the same, independent of the signal and the type, as the public never learns the state of the world. Mathematically, in this case, there are four variables, $\sigma_W(l), \sigma_W(h), \sigma_N(l)$, and $\sigma_N(h)$, and only one equation, as $\Delta^{0}_{W,l}[\sigma_W, \sigma_N] = \Delta^{0}_{W,h}[\sigma_W, \sigma_N] = \Delta^{0}_{N,l}[\sigma_W, \sigma_N] = \Delta^{0}_{N,h}[\sigma_W, \sigma_N]$.

Proposition 1. *For any $\mu > 0$:*

1. *If in equilibrium $\sigma_N^\mu(l)^* \neq \sigma_N^\mu(h)^*$, then $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1,0)$ is the unique equilibrium strategy of the wise type judge.*

2. *If in equilibrium $\sigma_N^\mu(l)^* = \sigma_N^\mu(h)^*$, there exists $\widetilde{\alpha}_B \in (0,1)$ such that:*

   (a) *If $\alpha_B > \widetilde{\alpha}_B$, then $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0,0)$ is the unique equilibrium strategy of the wise type judge.*

   (b) *If $\alpha_B < \widetilde{\alpha}_B$, then in equilibrium either $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0,0)$, $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1,0)$ or $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (x_1, 0)$, with $x_1 \in (0,1)$.*

For expositional purposes, we first discuss point *2.* of the Proposition, which characterizes the behavior of the wise type judge when the normal type uses a non-informative strategy. It identifies the existence of threshold $\widetilde{\alpha}_B$ such that when the probability that the judge is a biased type is higher than this threshold, in equilibrium, the wise type judge always uses a non-informative strategy, which consists in always passing the harsh sentence. It also states that when $\alpha_B < \widetilde{\alpha}_B$, in equilibrium, the wise type judge uses one of the following three strategies: $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0,0)$, $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1,0)$ or $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (x_1, 0)$, with $x_1 \in (0,1)$.

The first point of Proposition 1 characterizes the behavior of the wise type judge when the normal type uses an informative strategy. It states that, in equilibrium, the wise type judge always follows her signal and takes action $a = s$. A first implication of this result is that condition $\sigma_N^\mu(l)^* \neq \sigma_N^\mu(h)^*$, i.e., the normal type judge uses an informative strategy, is necessary and sufficient for the equilibrium to be informative. A second implication is that in any informative equilibrium the wise type judge uses the honest strategy. In other words, there is no informative equilibrium in which $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1,0)$. Corollary 1 below formally states (and elaborates a bit further) this idea, by describing the behavior of the normal type judge when $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1,0)$.

Corollary 1. *For all $\mu > 0$, if in equilibrium $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1,0)$, then $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0,0)$ is the unique equilibrium strategy of the normal type judge.*

Corollary 1 states that if the wise type judge does not use the honest strategy, i.e., $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1,0)$, then, in equilibrium, the normal type judge always takes the harsh action, for any level of transparency on consequences. A straightforward implication of Proposition 1 and Corollary 1 is that there is always an equilibrium in which the two strategic types always take the harsh action, independently of the signal and the level of transparency. In other words, the strategy profile $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*; \sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0,0;0,0)$ is always an equilibrium strategy profile. Proposition 3 will further elaborate on this idea.

The next result characterizes the behavior of the normal type judge after the harsh signal ($s = h$). It states that, in equilibrium, after the harsh signal, and independently of the strategy of the wise type

judge, the normal type judge always takes the harsh action; then, $\sigma_N^\mu(h)^* = 0$ for all $\mu > 0$. To see the intuition, simply note that, in this case, both the incentives to look wise and to look unbiased point to the same direction: passing the harsh sentence.

Proposition 2. *For all $\mu > 0$, in any equilibrium, $\sigma_N^\mu(h)^* = 0$.*

Given the result of Proposition 2, in the rest of the paper we focus our attention on the equilibrium behavior of the normal type judge when she receives the lenient signal ($s = l$), which constitutes the interesting case. We start by presenting the beliefs of the public about the type of the judge under the consideration that the normal type judge plays $\sigma_N^\mu(h)^* = 0$, and the wise type judge plays the honest strategy, i.e., $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$.[10] For a given sentence $a \in \{\hat{l}, \hat{h}\}$ and feedback $X \in \{0, L, H\}$, we obtain that the public's beliefs about the type of the judge, $\lambda_W(a, X)$ and $\lambda_{\bar{B}}(a, X)$, which determine the judge's expected reputational rent $\Pi(a, X)$, as described by (1), are given by:

[Tables 1 and 2 about here]

$\lambda_W(a,X)$			
	0	L	H
$\hat{l}$	$\frac{\alpha_W}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N}$	$\frac{\alpha_W}{\alpha_B+\alpha_W+\gamma\sigma_N(l)\alpha_N}$	0
$\hat{h}$	$\frac{\alpha_W}{\alpha_W+(2-\sigma_N(l))\alpha_N}$	0	$\frac{\alpha_W}{\alpha_W+(\gamma+(1-\gamma)(1-\sigma_N(l)))\alpha_N}$

Table 1: Principal's belief that the judge is a wise type agent.

$\lambda_{\bar{B}}(a,X)$			
	0	L	H
$\hat{l}$	$\frac{\alpha_W+\sigma_N(l)\alpha_N}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N}$	$\frac{\alpha_W+\gamma\sigma_N(l)\alpha_N}{\alpha_B+\alpha_W+\gamma\sigma_N(l)\alpha_N}$	$\frac{(1-\gamma)\sigma_N(l)\alpha_N}{\alpha_B+(1-\gamma)\sigma_N(l)\alpha_N}$
$\hat{h}$	1	1	1

Table 2: Principal's belief that the judge is an unbiased type agent.

The observation of the public's beliefs in Tables 1 and 2 yields a number of interesting comments. At this point, we focus on the effect that the consideration of a biased type judge has on the behavior of the normal type judge and, in particular, on the incentive of the normal type judge to pass the harsh sentence after the lenient signal. To see this, it is useful to compare two extreme scenarios, corresponding to cases $\alpha_B \to 1$ and $\alpha_B \to 0$.

[10] This is the interesting case as, from Corollary 1, if $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1, 0)$, the equilibrium behavior of the normal type judge after the lenient signal will always be $\sigma_N^\mu(l)^* = 0$, for any $\mu > 0$.

Consider first that the prior probability that the judge is biased is very high, i.e., $\alpha_B \to 1$. In this case, we can observe that both $\lambda_W(\hat{l}, X)$ and $\lambda_{\bar{B}}(\hat{l}, X)$ tend to 0 for any $X \in \{0, L, H\}$ and, therefore, for any μ. Hence, the first idea that we draw is that when the prior probability that the judge is biased is sufficiently high, the optimal action of a normal type judge who receives signal l is $a = \hat{h}$, for any level of transparency. Moreover, we can show that, in this case, there is no equilibrium in which $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$, and that, in the unique equilibrium, $\sigma_t^\mu(l)^* = 0$ for all $t \in \{W, N\}$ and $\mu > 0$. This means the unique equilibrium in this case is non-informative. This result, which constitutes the first point of Proposition 3 below, suggests that when the prior probability α_B is higher than a certain threshold, the incentive to avoid looking biased is so strong that, in the unique equilibrium of the game, judges disregard (even perfect) informative signals and always pass the harsh sentence, seeking to avoid being tarred as soft on crime.[11]

Suppose now that the prior probability that the judge is biased is very low, i.e., $\alpha_B \to 0$. In this case, we can observe that $\lambda_{\bar{B}}(\hat{l}, X) \to \lambda_{\bar{B}}(\hat{h}, X) = 1$ for any $X \in \{0, L, H\}$ and, therefore, for any μ. This means that, in the limit, the public's belief that a judge is unbiased is invariant to the judge's action and to the level of transparency. As a result, the normal type judge's decision is, here, exclusively driven by her concern for ability, $\lambda_W(a, X)$. In this case, we obtain that there is an equilibrium in which the normal type judge always passes the lenient sentence after the lenient signal, for any level of transparency. This result constitutes the second point of Proposition 3.

Proposition 3. *For any $\mu > 0$, there exists $\alpha_B^{max} < 1$ and $\alpha_B^{min} > 0$ such that the following holds:*

1. *If $\alpha_B > \alpha_B^{max}$, in the unique equilibrium, $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0, 0)$ and $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0, 0)$. The equilibrium is non-informative.*

2. *If $\alpha_B < \alpha_B^{min}$, in the unique informative equilibrium, $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$ and $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (1, 0)$.*

Two conclusions are drawn from Proposition 3. Firstly, the consideration of a biased type judge has important effects on the behavior of the strategic types and, especially, on the behavior of the normal type judge.[12] In particular, we show that a sufficiently high α_B induces the normal type judge (and even the wise type judge) to always pass the harsh sentence, for fear of looking biased; whereas a sufficiently low α_B induces the normal type judge to always follow her signal, trying to look wise. Secondly, when the proportion of biased type judges in the population is either too high or too low, transparency on consequences does not affect the equilibrium behavior of the strategic types.

[11] A sufficient condition for $\sigma_N^\mu(l)^* = 0$ is $\alpha_B > 1/2$. See Lemma 7 in Appendix A.3.

[12] The consideration of $\alpha_B > 0$ has a first effect on the equilibrium behavior of the normal type judge. The proof of point *1.* of Proposition 3 shows that for α_B higher than a certain threshold, the unique equilibrium strategy of the normal type judge is non-informative. Additional increases in α_B can also affect the equilibrium behavior of the wise type judge. In particular, we show that for α_B that is high enough, in equilibrium, the wise type judge also uses a non-informative strategy.

4.2 The perverse effect of transparency

This section presents the main result of the paper, stated in Proposition 4. It shows that there exists a region of parameters in which an increase in the probability that the public learns the state of the world induces the normal type judge to disregard an informative private signal more often and to pass the harsh sentence after the lenient signal with a higher probability. The conditions on the parameters refer to the prior probability that the judge is biased, which can be neither high nor too low, and the quality of the signal of the judge, which cannot be very high. The expressions of these thresholds are given in the proof of the result, in Appendix A.3. For simplicity, the result assumes $\eta = \beta$. However, this assumption is not crucial to the result, as Proposition 5 in the next section shows.

Proposition 4. *Consider $\eta = \beta$. There exist cutoffs γ', μ', μ'', α'_B, and α''_B, with $\gamma' > \frac{1}{2}$, $0 < \mu' < \mu'' < 1$, and $0 < \alpha'_B < \alpha''_B < \frac{1}{2}$, such that for all $\gamma \in (\frac{1}{2}, \gamma')$ and $\alpha_B \in (\alpha'_B, \alpha''_B)$:*

1. *If $\mu < \mu'$, in the unique informative equilibrium, $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (1, 0)$ and $(\sigma^\mu_N(l)^*, \sigma^\mu_N(h)^*) = (\sigma', 0)$, with $0 < \sigma' < 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$ and σ' decreasing in μ.*

2. *If $\mu \in (\mu', \mu'')$, in the most informative equilibrium, $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (1, 0)$ and $(\sigma^\mu_N(l)^*, \sigma^\mu_N(h)^*) = (\sigma'', 0)$, with $0 < \sigma'' < \sigma'$ and σ'' decreasing in μ.*

3. *If $\mu > \mu''$, there is no informative equilibrium. In this case, $(\sigma^\mu_N(l)^*, \sigma^\mu_N(h)^*) = (0, 0)$ and either $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (0, 0)$, $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (1, 0)$ or $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (x_1, 0)$, with $x_1 \in (0, 1)$.*

Proposition 4 describes the equilibrium behavior of the strategic types for different degrees of transparency on consequences.[13] The first point of the proposition states that for low enough values of μ, there is a unique informative equilibrium, in which a normal type judge that receives a lenient signal takes the lenient action with probability σ', with σ' being decreasing in μ.[14] The second point considers intermediate values of μ, in which case there can be multiple equilibria.[15] Here, we focus on the most informative equilibrium, which is the most interesting equilibrium from the point of view of the general public - it is

[13] In the proof of Proposition 4 (see Lemma 10), in Appendix A.3, we show that $\alpha''_B < \widetilde{\alpha}_B$. Note that, by Proposition 1, if $\alpha_B < \widetilde{\alpha}_B$, then there is an equilibrium in which $(\sigma^\mu_W(l)^*, \sigma^\mu_W(h)^*) = (1, 0)$.

[14] In Appendix A.3, we show that when $\mu = 0$, $\sigma^0_N(l)^* = 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$. Thus, $\sigma' < \sigma^0_N(l)^*$ for all $\mu > 0$.

[15] Note that $\Delta^\mu_{N,l}$ is not necessarily a monotonic function in $\sigma^\mu_N(l)$; hence, there can be multiple equilibria. See Figure 1. To see why $\Delta^\mu_{N,l}$ is not monotonic, first note that when $\sigma_N(l)$ increases, there are two forces at work: i) the public's belief that the judge is wise when taking action $a = \hat{h}$ increases (or does not vary) and the public's belief that the judge is wise when taking action $a = \hat{l}$ decreases (or does not vary). See Table 1. Thus, the expected reputational rent to the judge for taking action $\hat{h}$ ($\hat{l}$) increases (decreases); hence, $\Delta^\mu_{N,l} = \Pi^\mu_{N,l}(\hat{h}) - \Pi^\mu_{N,l}(\hat{l})$ increases. See expression (7) in Appendix A.1. ii) The public's belief that the judge is unbiased when taking action $a = \hat{h}$ does not vary, and the public's belief that the judge is unbiased when taking action $a = \hat{l}$ increases. See Table 2. Thus, the expected reputational rent to the judge for taking action $\hat{l}$ ($\hat{h}$) increases (does not vary); hence, $\Delta^\mu_{N,l} = \Pi^\mu_{N,l}(\hat{h}) - \Pi^\mu_{N,l}(\hat{l})$ decreases. See expression (7) in Appendix A.1. Secondly, note that when $\sigma_N(l)$ is low, effect ii) is stronger than effect i); hence, $\Delta^\mu_{N,l}$ decreases in $\sigma_N(l)$. In contrast, when $\sigma_N(l)$ is high, effect i) is stronger than effect ii); hence, $\Delta^\mu_{N,l}$ increases in $\sigma_N(l)$.

the one that maximizes the public's welfare. We obtain that the equilibrium probability that the normal type judge takes the lenient action after the lenient signal, σ'', also decreases in μ. The third point states that for high enough values of μ, the normal type judge always plays $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0, 0)$; which implies there is never an informative equilibrium in this case.

Two conclusions can be drawn from this result. First, contrary to what the literature has considered so far, increased transparency on consequences can induce an agent to go more often against an informative signal. This reduces the welfare of the public, and so talk about transparency on consequences as having a perverse effect. This effect is, however, different from the perverse effects of transparency previous identified in the literature (see Prat (2005) and Fox and Van Weelden (2012)). Second, this result speaks to the motivating stories in the Introduction, providing a theoretical foundation to explain why media coverage of judicial cases may lead to harsher sentencing practices. In the next section we will identify the driving forces behind this perverse effect and will shed light on the intuition for it. It will help us rationalize why media coverage of judicial cases results in lengthly sentences.

4.3 The sources of the perverse effect

To identify the driving forces behind the perverse effect, it is useful to consider each of the two concerns separately and, for each of them, to analyze the effect that transparency on consequences has on the incentive of a normal type judge to take the harsh action after the lenient signal.

The concern for bias. We start by considering $\eta \to 0$, in which case the objective function of the judge, as described by (1), is simply $\Pi(a, X) = \beta\lambda_{\bar{B}}(a, X)$. The analysis of this case yields the following important result.[16]

Remark 1. *Consider $\eta \to 0$. Then, $\Delta_{N,l}^0 < \Delta_{N,l}^1$ if and only if $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$.*

This result says that when the proportion of biased type judges in the population is below a certain threshold, the incentive of the judge to take the harsh action, rather than the lenient action, after the lenient signal, is higher with transparency than without it. In other words, the incentive to contradict an informative signal increases with transparency. To give an intuition for this result, it is important to note that transparency on consequences has very different effects on the payoffs from taking each action. Thus, whereas the payoff from taking the harsh action does not depend on μ (see Table 2), the payoff from taking the lenient action may decrease with μ. This is so because only when there is transparency on consequences, a judge that takes the lenient action faces the risk of mismatching the state and prove wrong. Interestingly, even if the judge does not care about ability, mismatching the state is a bad signal, as it clearly signals that the judge is not wise, which indirectly implies she is biased with

[16] To prove the result, simply note that when $\eta \to 0$, $\Delta_{N,l}^0 = \frac{2\alpha_B}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N}$ and $\Delta_{N,l}^1 = \frac{\gamma\alpha_B}{\alpha_B+\alpha_W+\gamma\sigma_N(l)\alpha_N} + \frac{(1-\gamma)\alpha_B}{\alpha_B+(1-\gamma)\sigma_N(l)\alpha_N}$ (see Table 2). From here, the result follows.

a higher probability. As a result, an increase in μ reduces the incentive to take the lenient action and increases the incentive to take the harsh one. Additionally, the higher the probability and/or the cost of mismatching the state, the smaller the incentive to take the lenient action and the higher the incentive to contradict the signal.

This argument helps explain why condition $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$ is easier to satisfy the higher α_W and the smaller α_B. To see it, note that the higher α_W, the higher the cost of mismatching the state, as the higher it is the increase in the probability of being biased when being shown wrong. Regarding α_B, a similar idea applies, as when α_B is very high, the cost of mismatching the state is very low. Last, condition $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$ is also easier to satisfy the smaller γ, as the smaller γ, the higher the probability of mismatching the state and being shown wrong when taking the lenient action; hence, the higher the incentive to take the harsh action.

Last, note that the result of Remark 1, though crucial, does not suffice to explain the perverse effect of transparency on consequences. This is clear from the fact that when $\eta \to 0$, in any equilibrium, $a = \hat{h}$ for all μ (see Table 2). The idea is that when there is a sole concern for bias, the incentive to take the harsh action is so strong that it completely hides the asymmetric effect that transparency on consequences has on the incentives to take the two actions. In this sense, Remark 1 describes a (first) necessary condition, although not sufficient, for the perverse effect to appear.

The concern for ability. Let us now consider $\beta \to 0$, in which case the objective function of the judge, as described by (1), is simply $\Pi(a, X) = \eta\lambda_W(a, X)$. Quite intuitively, adding a concern for ability (on top of a concern for bias), reduces the incentive to take the harsh action and, eventually, allows the perverse effect to come to light. For this to occur, the following condition must hold.[17]

Remark 2. *Consider $\beta \to 0$. Then, $\sigma_N^0(l)^* > 0$ if and only if $\alpha_B < \frac{1-\alpha_W}{2}$.*

This result says that when the proportion of biased type judges is below a certain threshold, then, in equilibrium without transparency, the normal type judge follows her signal and takes the lenient action with positive probability. This remark describes a (second) necessary condition for the perverse effect to appear, i.e, $\alpha_B < \frac{1-\alpha_W}{2}$. To see why this is a necessary condition, note that if the judge were rather to always take the harsh action when $\mu = 0$, i.e., $\sigma_N^0(l)^* = 0$, in equilibrium without transparency, a normal type judge that cares both about ability and bias would always disregard the lenient signal. In such a case, transparency could never have a perverse effect.

Note also that condition $\alpha_B < \frac{1-\alpha_W}{2}$ (which is equivalent to $\alpha_B < \alpha_N$) is easier to satisfy the smaller both α_W and α_B (alternatively, the higher α_N). To see this, note that the smaller α_B, the smaller the incentive to pass the harsh sentence. On the other hand, the smaller α_W, the more likely is that a harsh

[17] To prove the result, first note that when $\beta \to 0$, $\Delta^0_{N,l} = \frac{\alpha_W}{\alpha_W+(2-\sigma_N(l))\alpha_N} - \frac{\alpha_W}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N}$ (see Table 1), with $\Delta^0_{N,l}$ being increasing in $\sigma_N^0(l)$ and $\Delta^0_{N,l}\Big|_{\sigma_N^0(l)=1} > 0$. Additionally, note that $\Delta^0_{N,l}\Big|_{\sigma_N^0(l)=0} = \frac{\alpha_W}{\alpha_W+2\alpha_N} - \frac{\alpha_W}{2\alpha_B+\alpha_W}$, with $\Delta^0_{N,l}\Big|_{\sigma_N^0(l)=0} < 0$ iff $\alpha_B < \frac{1-\alpha_W}{2}$. From here, the result follows.

sentence comes from a normal judge. This reduces the payoff from sending $\hat{h}$ and increases the payoff from sending $\hat{l}$.

In addition to this result, and in line with what the reader might expect, Proposition 9 in Appendix A.3 shows that when there is a sole concern for ability, transparency on consequences never decreases the probability that the normal type judge takes the lenient action after the lenient signal. Because transparency disciplines in this case, when adding a concern for ability (on top of a concern for bias), we must be careful that the positive effect of transparency (that originates in the concern for ability) does not offset the negative effect of transparency (that originates in the concern for bias). The next result considers $\beta = 1$ and establishes sufficient conditions on η and the other parameters of the model for the perverse effect of transparency to come to light. The expressions of these thresholds are given in the proof of the result, in Appendix A.3.

Proposition 5. *Consider $\beta = 1$. There exists cutoffs $\bar{\alpha}_B$, $\bar{\gamma}$, $\bar{\mu}'$, $\bar{\mu}''$, $\bar{\eta}'$, and $\bar{\eta}''$, with $\bar{\alpha}_B \in (0,1)$, $\bar{\gamma} \in (\frac{1}{2},1)$, $0 < \bar{\mu}' < \bar{\mu}'' < 1$, and $0 < \bar{\eta}' < \bar{\eta}''$, such that for all $\alpha_B < \bar{\alpha}_B$, $\gamma < \bar{\gamma}$ and $\eta \in (\bar{\eta}', \bar{\eta}'')$:*

1. *If $\mu < \bar{\mu}'$, in the unique informative equilibrium $(\sigma_W^\mu(l)^*, \sigma_W^\mu(l)^*) = (1,0)$ and $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (\bar{\sigma}', 0)$, with $\bar{\sigma}' > 0$.*

2. *If $\mu > \bar{\mu}''$, there is no informative equilibrium. In this case, $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0,0)$ and either $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0,0)$, $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1,0)$ or $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (x_1, 0)$, with $x_1 \in (0,1)$.*

As expected, we observe that for the perverse effect of transparency to come to light, the weight of the concern for ability, measured by η, cannot be too high nor too small. We also observe that the smaller γ and α_B, the higher the range of values for which the perverse effect appears. These effects are in line with previous discussion of the parameters.[18] Two further comments on the role of these parameters are worth mentioning. First, despite the perverse effect is easier to sustain the smaller γ, we can find parameter configurations for which the perverse effect exits for γ very high.[19] Second, for the perverse effect to appear, we do not need a large proportion of biased and wise type judges in the population. However, these groups could be very small (e.g., the example in footnote 19). This result suggests that what matters for our result is that these groups exist, so that there is a reason for having a concern both for ability and bias; and not that the size of these groups is large as compared to the size of the whole population.

Top panels of Figure 1 illustrate the result of Proposition 5 with a numerical example. Given parameters α_B, α_W and γ, we observe that starting from $\eta = 0, \beta = 1$ and increasing η, reduces the incentive

[18] The effect of α_W, which is implicit in some of the expressions of the thresholds above, can neither be low (from Remark 1) nor high (from Remark 2).

[19] For example, if $\gamma = 0.9$, $\eta = 0.000143$, $\beta = 1$, $\alpha_B = 0.00001$, and $\alpha_W = 0.175$, we obtain $\sigma_N^0(l)^* = 0.2$ and $\sigma_N^1(l)^* = 0$. This result suggests that even when the signal of the judge is very informative ($\gamma = 0.9$), an increase in transparency on consequences may induce the judge to contradict it more often.

to take the harsh action and eventually makes the perverse effect of transparency come to light. We also observe that increasing η too much results in transparency having the positive desirable effect, for the concern for ability outweighing the concern for bias.[20]

[Figure 1 about here]

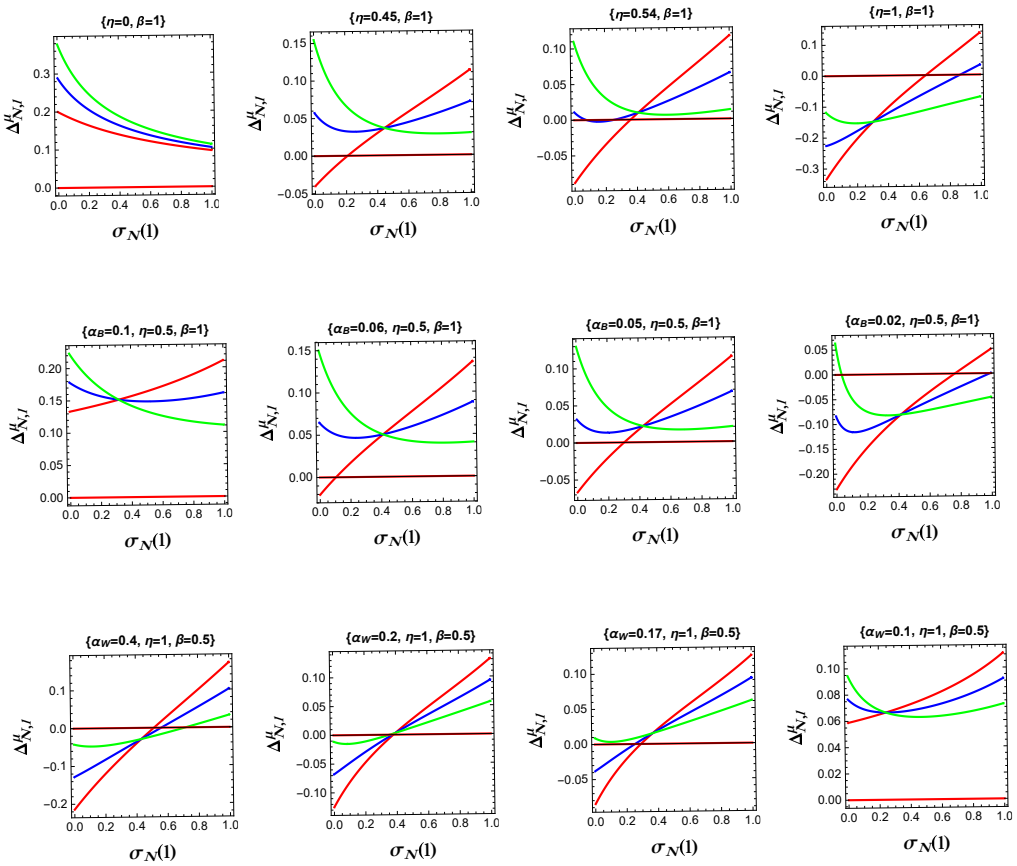

Figure 1: We represent $\Delta^{\mu}_{N,l}$ for different degrees of transparency on consequences and different values of parameters η, α_B and α_W. Red functions represent $\Delta^{0}_{N,l}$, blue functions represent $\Delta^{0.5}_{N,l}$, and green functions represent $\Delta^{1}_{N,l}$. Upper panels represent variations in η, with $\eta = 0, 0.45, 0.54, 1$, from left to right; setting $\beta = 1$, $\alpha_W = 0.4$, $\alpha_B = 0.05$, and $\gamma = 0.7$. Middle panels represent variations in α_B, with $\alpha_B = 0.1, 0.06, 0.05, 0.02$, from left to right; setting $\eta = 0.5$, $\beta = 1$, $\alpha_W = 0.4$ and $\gamma = 0.7$. Bottom panels represent variations in α_W, with $\alpha_W = 0.4, 0.2, 0.17, 0.1$, from left to right; setting $\eta = 1$, $\beta = 0.5$, $\alpha_B = 0.1$ and $\gamma = 0.6$.

Middle and bottom panels of Figure 1 propose a different exercise. They play with parameters α_B and α_W. To understand the purpose of this exercise, note that both α_B and β affect the behavior of the

[20] A similar exercise can be done for β, starting from $\eta = 1, \beta = 0$, and increasing β. It is straightforward to see that an increase in β will increase the incentive to take the harsh action and, eventually, will make that the perverse effect of transparency appears. For β sufficiently high, the incentive to take the harsh action will be so high that the judge will always pass the harsh sentence, irrespective of the level of transparency. This will occur when the concern for bias outweighs the concern for ability.

judge in a similar way, as an increase in either parameter increases the incentive to take the harsh action, for the concern for bias becoming relatively more important. Similarly, an increase in either α_W or η plays also similar roles, as both make the concern for ability relatively more salient. The substitutability between β and α_B on the one hand, and η and α_W on the other hand, although imperfect, has one important implication: it allows us to compensate effects and sustain the perverse effect of transparency on consequences even for quite different values of η and β. This is what middle and bottom panels of Figure 1 show.[21] Middle panels propose a situation where the concern for bias is twice stronger than the concern for ability, i.e., $\eta = 0.5$ and $\beta = 1$, and vary α_B. Here, we observe that despite the strong concern for bias, the perverse effect may come to light provided that the proportion of biased type judges is sufficiently small, for it reducing the incentive to pass the harsh action. Bottom panels propose the opposite exercise. They consider a situation where the concern for ability is twice stronger than the concern for bias, i.e., $\eta = 1$ and $\beta = 0.5$, and vary α_W. Similarly to the previous case, we observe that the perverse effect may come to light provided that the proportion of wise type judges is sufficiently small, for it reducing the incentive to pass the lenient action.

5 Discussion

This section discusses some of the assumptions of the model and shows the robustness of our results to these variations.

5.1 Popular belief about the state of the world

In this section, we relax the assumption about the two states of the world being equally likely. Let θ be the prior probability that the state is L, with $\theta \in (0, 1)$. Note that considering one state to be more likely than the other will introduce an incentive for the judge to go for the popular belief, as in herding models (see Avery and Chevalier (1999), Ottaviani and Sørensen (2006), and Gentzkow and Shapiro (2006), among others). Quite intuitively, this incentive to herd (on the popular belief) will reinforce or counterbalance the judge's incentive to pass the harsh sentence, depending on what the popular belief is. In other words, it may help the perverse effect to come to light, or hide it instead, depending on the strength of the other forces at work - just in line with the aforementioned effects that the other parameters of the model, η, β, α_B, and α_W, have.

Figure 2 below presents an exercise in this line. It considers $\eta = 1$, $\beta = 1$, and $\gamma = 0.75$, and propose a situation in which the proportion of wise type judges is so high as compared to the proportion of biased type judges ($\alpha_W = 0.7$ versus $\alpha_B = 0.01$), that the incentive to follow the lenient signal (seeking to show

[21] Another idea is the relationship between parameters α_B and γ. Remember that when $\eta \to 0$, $\Delta^0_{N,l} < \Delta^1_{N,l}$ if and only if $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$, which can be rewritten as $\gamma < \frac{\alpha_B+\alpha_W}{2\alpha_B+\alpha_W}$. Note that the expression on the right-hand side of the inequality is decreasing in α_B. Hence, the lower α_B is, the higher the range of values of γ for which the perverse effect of transparency on consequences appears.

ability) outweighs the incentive to pass the harsh sentence (seeking to show unbiasedness). As a result, in equilibrium, when $\theta = 1/2$, transparency has a positive desirable effect. Decreasing θ, however, increases the incentive to take the harsh action; hence, the incentive to go against the signal. This moves functions upward and, eventually, brings to the equilibrium the perverse effect of transparency on consequences - as we observe in the right panel, with $\theta = 0.26$. A similar exercise can be done for the opposite case: describing a status quo scenario with $\theta = 1/2$, where the proportion of biased types is high as compared to that of wise types, so that the incentive to go against the lenient signal and pass the harsh sentence outweighs the incentive to follow the signal. In such a case, we might expect the perverse effect to originate but not to appear in equilibrium when $\theta = 1/2$; however, an increase in θ, which increases the incentive to take the lenient action, might eventually make the perverse effect come to light.

[Figure 2 about here]

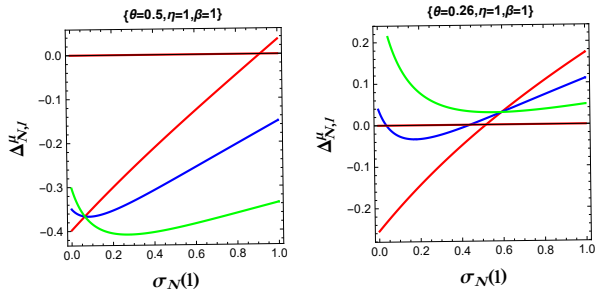

Figure 2: We represent $\Delta^{\mu}_{N,l}$ for different degrees of transparency on consequences and different values of parameter θ. Red functions represent $\Delta^{0}_{N,l}$, blue functions represent $\Delta^{0.5}_{N,l}$, and green functions represent $\Delta^{1}_{N,l}$. All panels consider $\eta = 1$, $\beta = 1$, $\gamma = 0.75$, $\alpha_W = 0.7$, and $\alpha_B = 0.01$. Left panel assumes $\theta = 0.5$, and right panel assumes $\theta = 0.26$.

The logic behind this exercise and the continuity of the judge's payoff function further suggest that if for a particular θ, say $\hat{\theta}$, the perverse effect appears in equilibrium, it would also appear for values of θ sufficiently close to $\hat{\theta}$. For the case of $\hat{\theta} = 1/2$, this argument suggests the robustness of our results to the consideration of a popular state of the world. The next result formally states this idea.

Proposition 6. *If $\sigma^1_N(l)^* < \sigma^0_N(l)^*$ when $\theta = 1/2$, there exists θ', θ'', with $\theta' < 1/2 < \theta''$, such that for all $\theta \in (\theta', \theta'')$, $\sigma^1_N(l)^* < \sigma^0_N(l)^*$.*

5.2 Strategic biased type

In the main body of the paper, we take the simplifying approach of considering that the biased type judge is not strategic and does not care about reputation. In this section, we relax this assumption. A simple way of analyzing a strategic biased type is to consider that the objective of the biased type judge

is to maximize $\Pi(a, X) + \phi I$, where $\Pi(a, X)$ represents the career concerns, as described by (1), $I = 1$ if $a = l$ and $I = 0$ otherwise, and $\phi > 0$ is a parameter that measures how strong the ideological concerns of the biased type are.[22] Apart from this, everything else in the model remains the same.

Quite intuitively, the analysis of this case suggests that we can consider a biased type judge with career concerns and guarantee that our main result still holds, provided that the ideological concerns of this type are sufficiently high. The idea is that for ϕ sufficiently high, the strategic biased judge will always pass the lenient sentence, which suffices to guarantee that all our results hold.[23] The next proposition formally states the result.

Proposition 7. *For all $\mu > 0$, there exists $\hat{\phi} > 0$ such that for all $\phi > \hat{\phi}$, $(\sigma_B^\mu(l)^*, \sigma_B^\mu(h)^*) = (1, 1)$.*

6 Conclusion

This paper analyzes the effects of transparency on consequences on the incentives of a careerist agent to act on her information. Our analysis shows that when we consider an agent with two concerns, ability and bias, increasing transparency on consequences may induce the agent to disregard an informative private signal and to take the (ex-ante) incorrect action with a higher probability. This result provides an argument to explain recent empirical evidence showing that media coverage of judicial cases lead to harsher sentencing practices.

The results in our paper have important policy implications. In terms of the judicial system, a first policy implication is that the public's perception of a bias in the judicial system may have important consequences. In particular, a perceived bias towards lenient sentences, as worldwide opinion studies seem to suggest (see footnote 5), would result in lengthly sentences.[24] Second, one should be cautious about media coverage of judicial cases, as it may even exacerbate the incentive for lengthly sentences. If media coverage of certain cases is something we cannot prevent, our policy recommendation would be to assign those cases to highly experienced and/or able judges (i.e., judges with higher values of γ). According to our results, they would be the judges less 'susceptible' to the perverse effect.

Beyond the courtrooms, our paper also helps rationalize the behavior of career concerned agents in other situations. Say physicians, financial forecasters, governments, and so; in the presence of a bias,

[22] See Maskin and Tirole (2004) and Besley (2006), for papers where career concerned agents receive a benefit from taking *themselves* their preferred action.

[23] A different approach would be to consider that the biased type aims to influence the action of the decision maker and persuades him to take her preferred action, as in Morris (2001). Though more complex than the alternative we analyze, we conjecture that for the appropriate discount factor (i.e., a sufficiently patient biased type), it would it an equilibrium in which the strategic biased type judge would always pass the lenient sentence.

[24] This argument might help explain why the proportion of jail inmates in US has sharply increased in the last forty years. Studies show that while in 1970 the proportion of US prisoners was below one in 400, in 2010 it was one in 100. From 2010 onwards, the number of US inmates has gone down. However, numbers are still much higher than in the past: 0.5 millions in the 80s, and 2.2 millions in 2016. See "Too many laws, too many prisoners", *The Economists*, July 22, 2010; and "America's incarceration rate is at a two-decade low", *Pew Research Center*, May 2, 2018.

for example, gender or racial biases. To those cases, our results suggest we should be cautious about policies that promote transparency on consequences, as more transparency might not be necessarily good. Speaking to a present issue, the Covid-19 pandemic, we consider that our paper might help explain the reaction of many countries around the world, restricting travels and closing borders with South Africa and other souther African countries, just after the new coronavirus variant Omicron was detected in South Africa in late November 2021. Of course we agree there are important health, safety, and economic reasons to support such a quick reaction and strong measure. In this sense, we do not argue that the only mechanism behind this decision was a pure reputational mechanism. However, we consider that if governments around the world care about their image and they perceive public opinion to doubt about governments' ability to manage the pandemic (see Lazarus et al. (2020)), our paper may have something to say to the story, and our result about the perverse effect of transparency may help explain part of the observed reaction.

A Appendix

The Appendix consists of three parts. In Appendix A.1, we introduce some relevant definitions that are useful for the posterior analysis. The analysis of the game and the proofs of the results are presented in Appendixes A.2 and A.3. In Appendix A.2, we analyze the equilibrium behavior of the wise type judge. This part includes the proof of Proposition 1. In Appendix A.3, we analyze the equilibrium behavior of the normal type judge. This part includes the proofs of Corollary 1 and Propositions 2, 3 and 4.

The analysis in the Appendix considers a more general set-up than the one described in the main body of the paper. In particular, all the results in the Appendix, except for Propositions 4 and 9, are proven for the case in which the objective function of the judge is as follows:

$$\Pi\left(a, X\right) = f\left(\lambda_W(a, X), \lambda_{\bar{B}}(a, X)\right), \tag{6}$$

with $f(\cdot)$ denoting an increasing function in $\lambda_W(a, X)$ and $\lambda_{\bar{B}}(a, X)$. This includes the linear specification of equation (1) as a particular case.

Additionally, the analysis in the Appendix considers $\mu \in [0, 1]$, which includes the limit case $\mu = 0$.

A.1 Part I: Definitions

We first detail the expression of the expected gain to a judge of type $t \in \{W, N\}$ for passing sentence $\hat{h}$ rather than $\hat{l}$ after observing signal $s \in \{l, h\}$ when the level of transparency is $\mu \in [0, 1]$. Given equations

(1)-(5), expressions $\Delta^{\mu}_{t,l}$ and $\Delta^{\mu}_{t,h}$ are given by equations (7) and (8) below:

$$\begin{aligned}\Delta^{\mu}_{t,l}[\sigma_W,\sigma_N] &= (1-\mu)\Delta^{0}_{t,l}[\sigma_W,\sigma_N]+\mu\Delta^{1}_{t,l}[\sigma_W,\sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(\hat{h},0),1\right)-f\left(\lambda_W(\hat{l},0),\lambda_{\bar{B}}(\hat{l},0)\right)\right)+ \\ &\mu\left(\left(\gamma f\left(0,1\right)+(1-\gamma)f\left(\lambda_W(\hat{h},H),1\right)\right)-\left(\gamma f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right)+(1-\gamma)f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right)\right)\right),\end{aligned} \tag{7}$$

$$\begin{aligned}\Delta^{\mu}_{t,h}[\sigma_W,\sigma_N] &= (1-\mu)\Delta^{0}_{t,h}[\sigma_W,\sigma_N]+\mu\Delta^{1}_{t,h}[\sigma_W,\sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(\hat{h},0),1\right)-f\left(\lambda_W(\hat{l},0),\lambda_{\bar{B}}(\hat{l},0)\right)\right)+ \\ &\mu\left(\left(\gamma f\left(\lambda_W(\hat{h},H),1\right)+(1-\gamma)f\left(0,1\right)\right)-\left(\gamma f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right)+(1-\gamma)f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right)\right)\right).\end{aligned} \tag{8}$$

Next, we introduce two definitions.

Definition 1. *Given $\mu \in [0,1]$, a strategy profile $\sigma^{\mu*} = (\sigma^{\mu*}_W, \sigma^{\mu*}_N)$, with $\sigma^{\mu*}_W = (\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*)$ and $\sigma^{\mu*}_N = (\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*)$ is a Perfect Bayesian equilibrium strategy profile if for each type $t \in \{W,N\}$:*

1. *When $s = l$, either $\Delta^{\mu}_{t,l}[\sigma^{\mu*}_W, \sigma^{\mu*}_N] = 0$ or $\Delta^{\mu}_{t,l}[\sigma^{\mu*}_W, \sigma^{\mu*}_N] > 0$ (< 0) holds. In the latter case, $\sigma^{\mu}_t(l)^* = 0$ (1).*
2. *When $s = h$, either $\Delta^{\mu}_{t,h}[\sigma^{\mu*}_W, \sigma^{\mu*}_N] = 0$ or $\Delta^{\mu}_{t,h}[\sigma^{\mu*}_W, \sigma^{\mu*}_N] > 0$ (< 0) holds. In the latter case, $\sigma^{\mu}_t(h)^* = 0$ (1).*

Definition 1 defines an equilibrium strategy profile. To stress the fact that, in equilibrium, the strategies of the wise type judge and the normal type judge may depend on μ, we make this dependence explicit and write the superscript μ.

Definition 2. *Consider $\mu = 0$. An equilibrium strategy $\bar{\sigma}^0_t(s)^*$ is robust to transparency if there exist $\bar{\mu} > 0$ and an associated equilibrium strategy $\sigma^{\bar{\mu}}_t(s)^*$ such that $\lim_{\bar{\mu}\to 0} \sigma^{\bar{\mu}}_t(s)^* = \bar{\sigma}^0_t(s)^*$.*

The second definition defines a robustness criterion. This definition will be of help when analyzing the limit case $\mu = 0$, where there is a multiplicity of equilibria. To see the multiplicity, note that when $\mu = 0$, the expected gain to the judge for passing sentence $\hat{h}$ rather than $\hat{l}$ is the same, independent of the signal and the type, as the principal never learns the state of the world and, hence, the correct sentence. Mathematically, there are four variables in this case, $\sigma_W(l), \sigma_W(h), \sigma_N(l)$, and $\sigma_N(h)$, and only one equation, as $\Delta^0_{W,l}[\sigma_W,\sigma_N] = \Delta^0_{W,h}[\sigma_W,\sigma_N] = \Delta^0_{N,l}[\sigma_W,\sigma_N] = \Delta^0_{N,h}[\sigma_W,\sigma_N]$.

A.2 Part II: Analysis of the wise type judge

In this section, we analyze the equilibrium behavior of the wise type judge. Prior to this, note that for the wise type judge, $\gamma = 1$. Hence, equations (7)-(8) above simplify to the following:

$$\begin{aligned}\Delta^{\mu}_{W,l}[\sigma_W, \sigma_N] &= (1-\mu)\Delta^{0}_{W,l}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{W,l}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(\hat{h}, 0), 1\right) - f\left(\lambda_W(\hat{l}, 0), \lambda_{\bar{B}}(\hat{l}, 0)\right)\right) + \mu\left(f(0,1) - f\left(\lambda_W(\hat{l}, L), \lambda_{\bar{B}}(\hat{l}, L)\right)\right),\end{aligned} \tag{9}$$

$$\begin{aligned}\Delta^{\mu}_{W,h}[\sigma_W, \sigma_N] &= (1-\mu)\Delta^{0}_{W,h}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{W,h}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(\hat{h}, 0), 1\right) - f\left(\lambda_W(\hat{l}, 0), \lambda_{\bar{B}}(\hat{l}, 0)\right)\right) + \mu\left(f\left(\lambda_W(\hat{h}, H), 1\right) - f\left(0, \lambda_{\bar{B}}(\hat{l}, H)\right)\right).\end{aligned} \tag{10}$$

Proof of Proposition 1.

The results of Proposition 1 hold for the more general objective function described in (6), which is assumed in the proof.

Lemmas 1-4 prove the result. Lemma 1 introduces a preliminary technical result. Lemma 4 analyzes the behavior of the *normal* type judge. This lemma shows that when $\sigma^{\mu}_N(l)^* > 0$, the equilibrium strategy of the normal type judge is always informative, i.e., $\sigma^{\mu}_N(l)^* \neq \sigma^{\mu}_N(h)^*$; and that when $\sigma^{\mu}_N(l)^* = 0$, the equilibrium strategy of the normal type judge is always non-informative, i.e., $\sigma^{\mu}_N(l)^* = \sigma^{\mu}_N(h)^*$. Considering, now, the behavior of the *wise* type judge, Lemma 2 characterizes the equilibrium strategy of the wise type judge when $\sigma^{\mu}_N(l)^* > 0$ and obtains that, in this case, $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1, 0)$. Given Lemma 4, Lemma 2 thus proves point *1.* of Proposition 1. Finally, Lemma 3 characterizes the equilibrium strategy of the wise type judge when $\sigma^{\mu}_N(l)^* = 0$. In this case, we show that there is a threshold, $\widetilde{\alpha}_B$, such that if $\alpha_B > \widetilde{\alpha}_B$, the wise type always takes action $a = \hat{l}$. However, if $\alpha_B < \widetilde{\alpha}_B$, there are three equilibrium strategies, one of them being the honest strategy $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1, 0)$. Given Lemma 4, Lemma 3 thus proves point *2.* of Proposition 1.

Lemma 1. *Consider $\mu > 0$. For any increasing function $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X))$, we have $\Delta^{\mu}_{W,l}[\sigma_W, \sigma_N] < \Delta^{\mu}_{N,l}[\sigma_W, \sigma_N] < \Delta^{\mu}_{N,h}[\sigma_W, \sigma_N] < \Delta^{\mu}_{W,h}[\sigma_W, \sigma_N]$.*

Proof

Firstly, we prove that $\lambda_{\bar{B}}(\hat{l}, L) \geq \lambda_{\bar{B}}(\hat{l}, H)$. Note the following:

$$\begin{aligned}\lambda_B(\hat{l}, L) &= \frac{P(\hat{l}|B,L)P(L)P(B)}{P(\hat{l}|B,L)P(L)P(B)+P(\hat{l}|N,L)P(L)P(N)+P(\hat{l}|W,L)P(L)P(W)} = \frac{\alpha_B}{\alpha_B+(\gamma\sigma_N(l)+(1-\gamma)\sigma_N(h))\alpha_N+\sigma_W(l)\alpha_W}, \\ \lambda_B(\hat{l}, H) &= \frac{P(\hat{l}|B,H)P(H)P(B)}{P(\hat{l}|B,H)P(H)P(B)+P(\hat{l}|N,H)P(H)P(N)+P(\hat{l}|W,H)P(H)P(W)} = \frac{\alpha_B}{\alpha_B+((1-\gamma)\sigma_N(l)+\gamma\sigma_N(h))\alpha_N+\sigma_W(h)\alpha_W},\end{aligned}$$

with $\lambda_B(\hat{l}, L) \leq \lambda_B(\hat{l}, H) \iff (1-\gamma)\sigma_N(l)+\gamma\sigma_N(h) \leq \gamma\sigma_N(l)+(1-\gamma)\sigma_N(h) \iff (1-2\gamma)\sigma_N(l) \leq (1-2\gamma)\sigma_N(h) \iff \sigma_N(l) \geq \sigma_N(h)$. This is always the case in a non-perverse equilibrium. For the same reason, $\sigma_W(l) \geq \sigma_W(h)$. Hence, $\lambda_B(\hat{l}, L) \leq \lambda_B(\hat{l}, H)$, and consequently, $\lambda_{\bar{B}}[\hat{l}, L] \geq \lambda_{\bar{B}}(\hat{l}, H)$.

Secondly, note that $\lambda_{\bar{B}}(\hat{l},L) \geq \lambda_{\bar{B}}(\hat{l},H)$ implies $f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right) > f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right)$. Finally, from equations (7)-(8) and (9)-(10), it is straightforward to show the following:

$$f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right) > f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right) \Longrightarrow \Delta^{\mu}_{W,l}[\sigma_W,\sigma_N] < \Delta^{\mu}_{N,l}[\sigma_W,\sigma_N],$$
$$f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right) > f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right) \Longrightarrow \Delta^{\mu}_{N,l}[\sigma_W,\sigma_N] < \Delta^{\mu}_{N,h}[\sigma_W,\sigma_N],$$
$$f\left(\lambda_W(\hat{l},L),\lambda_{\bar{B}}(\hat{l},L)\right) > f\left(0,\lambda_{\bar{B}}(\hat{l},H)\right) \Longrightarrow \Delta^{\mu}_{N,h}[\sigma_W,\sigma_N] < \Delta^{\mu}_{W,h}[\sigma_W,\sigma_N]. \blacklozenge$$

Lemma 2. *Consider $\mu > 0$ and $\sigma^{\mu}_N(l)^* > 0$. For any increasing function $f(\lambda_W(a,X),\lambda_{\bar{B}}(a,X))$, the unique equilibrium strategy of the wise type judge is $(\sigma^{\mu}_W(l)^*,\sigma^{\mu}_W(h)^*) = (1,0)$.*

Proof

Firstly, we prove that $\sigma^{\mu}_W(h)^* = 0$. We prove it by contradiction, so let us assume $\sigma^{\mu}_W(h)^* > 0$. In this case, $\Delta^{\mu}_{W,h}[\sigma^*_W,\sigma^*_N] \leq 0$; hence, by Lemma 1, $\Delta^{\mu}_{W,l}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{N,l}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{N,h}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{W,h}[\sigma^*_W,\sigma^*_N] \leq 0$, and therefore, $\sigma^{\mu}_W(l)^* = 1$, $\sigma^{\mu}_N(l)^* = 1$ and $\sigma^{\mu}_N(h)^* = 1$. This implies that if the wise type judge takes action $a = \hat{h}$, the principal's equilibrium beliefs assign probability one to the judge being wise and zero probability to her being biased. Thus, the payoff to the judge for taking action $a = \hat{h}$ is always higher than the payoff for taking action $a = \hat{l}$. Hence, $\sigma^{\mu}_W(h)^*$ cannot be greater than zero in equilibrium, which is a contradiction. Consequently, in equilibrium, $\sigma^{\mu}_W(h)^* = 0$ always.

Secondly, we prove that $\sigma^{\mu}_W(l)^* = 1$. Note that if $\sigma^{\mu}_N(l)^* > 0$, then $\Delta^{\mu}_{N,l}[\sigma^*_W,\sigma^*_N] \leq 0$. Now, by Lemma 1, $\Delta^{\mu}_{W,l}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{N,l}[\sigma^*_W,\sigma^*_N] \leq 0$. Then, in equilibrium, $\sigma^{\mu}_W(l)^* = 1$ always. $\blacklozenge$

Lemma 3. *Consider $\mu > 0$ and $\sigma^{\mu}_N(l)^* = 0$. For any increasing function $f(\lambda_W(a,X),\lambda_{\bar{B}}(a,X))$, in equilibrium $\sigma^{\mu}_N(h)^* = \sigma^{\mu}_W(h)^* = 0$. In the case in which $s = l$, there exists threshold $\widetilde{\alpha}_B$ such that if $\alpha_B > \widetilde{\alpha}_B$, then $\sigma^{\mu}_W(l)^* = 0$; and if $\alpha_B < \widetilde{\alpha}_B$, then in equilibrium either $\sigma^{\mu}_W(l)^* = 0$, $\sigma^{\mu}_W(l)^* = 1$ or $\sigma^{\mu}_W(l)^* \in (0,1)$. The described equilibrium strategies are the unique ones.*

Proof

Firstly, note that if $\sigma^{\mu}_N(l)^* = 0$, then $\Delta^{\mu}_{N,l}[\sigma^*_W,\sigma^*_N] \geq 0$. Hence, by Lemma 1, $0 \leq \Delta^{\mu}_{N,l}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{N,h}[\sigma^*_W,\sigma^*_N] < \Delta^{\mu}_{W,h}[\sigma^*_W,\sigma^*_N]$. It implies $\sigma^{\mu}_N(h)^* = 0$ and $\sigma^{\mu}_W(h)^* = 0$.

To obtain the equilibrium value $\sigma^{\mu}_W(l)^*$, we analyze equation (9). Firstly, note that when $\sigma^{\mu}_N(l)^* = 0$, $\sigma^{\mu}_N(h)^* = 0$ and $\sigma^{\mu}_W(h)^* = 0$, beliefs are as follows:

$$\lambda_W(\hat{h},0) = \tfrac{(2-\sigma_W(l))\alpha_W}{(2-\sigma_W(l))\alpha_W+2(1-\alpha_W-\alpha_B)},$$
$$\lambda_W(\hat{l},0) = \lambda_{\bar{B}}(\hat{l},0) = \tfrac{\sigma_W(l)\alpha_W}{\sigma_W(l)\alpha_W+2\alpha_B},$$
$$\lambda_W(\hat{l},L) = \lambda_{\bar{B}}(\hat{l},L) = \tfrac{\sigma_W(l)\alpha_W}{\sigma_W(l)\alpha_W+\alpha_B},$$

with $\frac{\partial\lambda_W(\hat{h},0)}{\partial\sigma_W(l)} < 0$, $\frac{\partial\lambda_W(\hat{l},0)}{\partial\sigma_W(l)} > 0$ and $\frac{\partial\lambda_W(\hat{l},L)}{\partial\sigma_W(l)} > 0$. Consequently, $\frac{\partial\Delta^{\mu}_{W,l}[\sigma_W,\sigma_N]}{\partial\sigma_W(l)} < 0$.

Additionally, $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=0} = (1-\mu)\left(f\left(\frac{\alpha_W}{\alpha_W+(1-\alpha_W-\alpha_B)},1\right) - f(0,0)\right) + \mu\left(f(0,1) - f(0,0)\right) > 0.$

Then, the sign of $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1}$ determines whether there is either one equilibrium, $\sigma^{\mu}_W(l)^* = 0$, or three equilibria: $\sigma^{\mu}_W(l)^* = 0$, $\sigma^{\mu}_W(l)^* = 1$ and $\sigma^{\mu}_W(l)^* \in (0,1)$. Note that if $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1} \leq 0$, there are three equilibria, and if $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1} > 0$, there is only one equilibrium.

Note that

$$\begin{aligned}\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1} = (1-\mu)\left(f\left(\frac{\alpha_W}{\alpha_W + 2(1-\alpha_W-\alpha_B)}, 1\right) - f\left(\frac{\alpha_W}{\alpha_W+2\alpha_B}, \frac{\alpha_W}{\alpha_W+2\alpha_B}\right)\right) \\ + \mu\left(f(0,1) - f\left(\frac{\alpha_W}{\alpha_W+\alpha_B}, \frac{\alpha_W}{\alpha_W+\alpha_B}\right)\right). \qquad (11)\end{aligned}$$

This expression is increasing in α_B. Hence, when α_B is close enough to zero, we have $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1} < 0$. However, when α_B is high enough, the expression is positive. Consequently, there exists a threshold for α_B, denoted by $\widetilde{\alpha}_B$, such that above the threshold, there is a unique equilibrium, and below the threshold, there are three equilibria. ♦

Lemma 4. *In equilibrium,*

1. *If $\sigma^{\mu}_N(l)^* = 0$, then the unique equilibrium strategy of the normal type judge is non-informative, i.e., $\sigma^{\mu}_N(l)^* = \sigma^{\mu}_N(h)^*$.*
2. *If $\sigma^{\mu}_N(l)^* > 0$, then the equilibrium strategy of the normal type judge is informative, i.e., $\sigma^{\mu}_N(l)^* \neq \sigma^{\mu}_N(h)^*$.*

Proof

Case 1. From Lemma 3, we know that if $\sigma^{\mu}_N(l)^* = 0$, then in equilibrium, $\sigma^{\mu}_N(h)^* = 0$ always. The strategy of the normal type judge is thus non-informative.

Case 2. From Lemma 2, we know that if $\sigma^{\mu}_N(l)^* > 0$, then in equilibrium, $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$ always. Additionally, Proposition 2 in Appendix A.3 below shows that if $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$, then in equilibrium, $\sigma^{\mu}_N(h)^* = 0$ always. Combining these results, we have that the equilibrium strategy of the normal type judge is informative if and only if $\sigma^{\mu}_N(l)^* > 0$.

Hence, in our model, there is an equivalence between stating $\sigma^{\mu}_N(l)^* > 0$ and saying that the normal type judge uses an informative strategy, and between stating $\sigma^{\mu}_N(l)^* = 0$ and saying that the normal type judge uses a non-informative strategy. ♦

This completes the proof of Proposition 1. **QED**.

Remark on Proposition 1 When $\mu = 0$, then $\Delta^0_{W,l}[\sigma_W, \sigma_N] = \Delta^0_{N,l}[\sigma_W, \sigma_N] = \Delta^0_{N,h}[\sigma_W, \sigma_N] = \Delta^0_{W,h}[\sigma_W, \sigma_N]$, which implies that there are multiple equilibria. However, by Definition 2, the equilibrium strategies of the wise type judge that Proposition 1 describes are the only ones that can be robust to transparency when $\mu = 0$.

A.3 Part III: Analysis of the normal type judge

First, note that Proposition 1 above shows that if the normal type judge uses an informative strategy, then the unique equilibrium strategy of the wise type judge is $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$. Hence, if the analysis that follows shows that under this premise the equilibrium strategy of the normal type judge is informative, then we have an informative equilibrium. However, if the best response of the normal type judge is to use a non-informative strategy, then to know whether there is a (non-informative) equilibrium in which the normal type uses a non-informative strategy, we need to analyze the following: i) the best response of the wise type judge to a normal type using a non-informative strategy and ii) the best response of the normal type judge to the previous optimal response of the wise type. Note that Proposition 1 completely characterizes the wise type judge's optimal behavior in point i). It shows that if $\alpha_B > \widetilde{\alpha}_B$, the unique equilibrium strategy of the wise type judge is $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (0, 0)$. However, if $\alpha_B < \widetilde{\alpha}_B$, the wise type judge has three equilibrium strategies that can be subsumed into the following two: $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$ and $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (x_1, 0)$, with $x_1 < 1$. The next two results, Proposition 8 and Corollary 1, characterize the equilibrium strategy of the normal type judge when the wise type does not always follow her signal, i.e., when $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1, 0)$. Hence, these results answer point ii).

Proposition 8. *Consider $\sigma_W^\mu(l)^* < 1$ and $\mu \in [0, 1]$. For any increasing function $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X))$, in equilibrium, $\sigma_W^\mu(h)^* = 0$ and $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0, 0)$. Furthermore, if $\mu > 0$, the equilibrium strategy of the normal type judge is unique.*

Proof

The proof restricts attention to the case $\mu > 0$. Note that if $\mu = 0$, there is a multiplicity of equilibria, one of which prescribes $\sigma_W^\mu(h)^* = 0$ and $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0, 0)$.

Now, if $\sigma_W^\mu(l)^* < 1$, then $\Delta_{W,l}^\mu[\sigma_W^*, \sigma_N^*] \geq 0$. By Lemma 1, $0 < \Delta_{N,l}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,h}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{W,h}^\mu[\sigma_W^*, \sigma_N^*]$, which implies $\sigma_W^\mu(h)^* = 0$, $\sigma_N^\mu(l)^* = 0$, and $\sigma_N^\mu(h)^* = 0$. **QED**.

Proof of Corollary 1.

Note that if $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) \neq (1, 0)$, then necessarily either $\sigma_W^\mu(l)^* < 1$ or $\sigma_W^\mu(h)^* > 0$. By Proposition 1, there is no equilibrium in which $\sigma_W^\mu(h)^* > 0$; hence, $\sigma_W^\mu(l)^* < 1$ must hold. Now, by Proposition 8, if $\sigma_W^\mu(l)^* < 1$, then $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (0, 0)$. **QED**.

Proof of Proposition 2.

The result of Proposition 2 holds for the more general objective function described in (6), which is assumed in the proof.

We start noting that, by Corollary 1, if the wise type does not use the honest strategy, then $\sigma_N^\mu(h)^* = 0$. Hence, we next consider that the wise type uses the honest strategy, i.e., $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$.

In this case, we first prove that, in equilibrium, the normal type judge cannot disregard both signals l and h with positive probability. We prove this by contradiction. Hence, suppose that $\sigma_N^\mu(l)^* < 1$ and $\sigma_N^\mu(h)^* > 0$. Since $\sigma_N^\mu(h)^* > 0$, then $\Delta_{N,h}^\mu \leq 0$. Additionally, since $\sigma_N^\mu(l)^* < 1$, then $\Delta_{N,l}^\mu \geq 0$. This contradicts Lemma 1, which states $\Delta_{N,l}^\mu < \Delta_{N,h}^\mu$. Then, the equilibrium strategy profile of the normal type judge is either $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (y_1, 0)$ with $y_1 < 1$, or $(\sigma_N^\mu(l)^*, \sigma_N^\mu(h)^*) = (1, y_2)$ with $y_2 > 0$.

Next, we show that $\sigma_N^\mu(h)^* > 0$ is not possible. Again, we prove it by contradiction. Suppose that $\sigma_N^\mu(h)^* > 0$. Then, $\sigma_N^\mu(l)^* = 1$. Now, we use the results (12) and (13), which we prove below:

$$\lambda_W(\hat{h}, H) > \lambda_W(\hat{l}, L) \Longleftrightarrow (1 - \sigma_N(l) - \sigma_N(h))\alpha_N < \alpha_B, \tag{12}$$

$$\lambda_W(\hat{h}, 0) > \lambda_W(\hat{l}, 0) \Longleftrightarrow (1 - \sigma_N(l) - \sigma_N(h))\alpha_N < \alpha_B. \tag{13}$$

In the case that $\sigma_N(h) > 0$ and $\sigma_N(l) = 1$, then $\lambda_W(\hat{h}, H) > \lambda_W(\hat{l}, L)$ and $\lambda_W(\hat{h}, 0) > \lambda_W(\hat{l}, 0)$. Substituting in expression (8), we obtain $\Delta_{N,h}^\mu > 0$. However, $\Delta_{N,h}^\mu > 0$ implies $\sigma_N^\mu(h)^* = 0$, which contradicts $\sigma_N^\mu(h)^* > 0$. As a result, in equilibrium, $\sigma_N^\mu(h)^* = 0$.

To complete the proof, we prove conditions (12) and (13). Given the equilibrium strategy $(\sigma_W^\mu(l)^*, \sigma_W^\mu(h)^*) = (1, 0)$, Bayes' rule determines the following:

$$\lambda_W(\hat{h}, H) = \frac{\alpha_W}{\alpha_W + (\gamma(1-\sigma_N(h)) + (1-\gamma)(1-\sigma_N(l)))\alpha_N} \qquad \lambda_W(\hat{h}, 0) = \frac{\alpha_W}{\alpha_W + (2 - \sigma_N(h) - \sigma_N(l))\alpha_N}$$

$$\lambda_W(\hat{l}, L) = \frac{\alpha_W}{\alpha_B + \alpha_W + ((1-\gamma)\sigma_N(h) + \gamma\sigma_N(l))\alpha_N} \qquad \lambda_W(\hat{l}, 0) = \frac{\alpha_W}{2\alpha_B + \alpha_W + (\sigma_N(h) + \sigma_N(l))\alpha_N}$$

First, we observe that $\lambda_W(\hat{h}, H) > \lambda_W(\hat{l}, L) \Longleftrightarrow \alpha_B + \alpha_W + ((1-\gamma)\sigma_N(h) + \gamma\sigma_N(l))\,\alpha_N > \alpha_W + (\gamma(1-\sigma_N(h)) + (1-\gamma)(1-\sigma_N(l)))\,\alpha_N \Longleftrightarrow (1 - \sigma_N(l) - \sigma_N(h))\alpha_N < \alpha_B$.

Second, we observe that $\lambda_W(\hat{h}, 0) > \lambda_W(\hat{l}, 0) \Longleftrightarrow 2\alpha_B + \alpha_W + (\sigma_N(l) + \sigma_N(h))\alpha_N > \alpha_W + (2 - \sigma_N(l) - \sigma_N(h))\alpha_N \iff (1 - \sigma_N(l) - \sigma_N(h))\alpha_N < \alpha_B$. **QED**.

Remark on Proposition 2. When $\mu = 0$, there are multiple equilibria. According to Definition 2, in all equilibria that are robust to transparency, $\sigma_N^\mu(h)^* = 0$.

Proof of Proposition 3.

The results of Proposition 3 hold for the more general objective function described in (6), which is assumed in the proof.

First, we prove point *1.* of the proposition. To show that when $\alpha_B > \alpha_B^{max}$, $\Delta_{N,l}^\mu > 0$ for any $\mu \in [0, 1]$, we show that there exist two thresholds, $\breve{\alpha}_B < 1$ and $\breve{\alpha}_B < 1$, such that $\Delta_{N,l}^0 > 0$ for all $\alpha_B > \breve{\alpha}_B$ and $\Delta_{N,l}^1 > 0$ for all $\alpha_B > \breve{\alpha}_B$.

Let us start with $\breve{\alpha}_B$. Note that $\Delta_{N,l}^1 = \Pi_{N,l}^1(\hat{h}) - \Pi_{N,l}^1(\hat{l})$, with $\Pi_{N,l}^1(\hat{h}) = \gamma f(0,1) + (1-\gamma) f\left(\lambda_W(\hat{h}, H), 1\right) > \gamma f(0,1) + (1-\gamma) f(0,1) = f(0,1) > 0$ and $\Pi_{N,l}^1(\hat{l}) = \gamma f\left(\lambda_W(\hat{l}, L), \lambda_{\bar{B}}(\hat{l}, L)\right) + (1-\gamma) f\left(0, \lambda_{\bar{B}}(\hat{l}, H)\right)$. It can be shown that $\Pi_{N,l}^1(\hat{l})$ is decreasing in α_B (either with α_W or α_N constant), with $lim_{\alpha_B \to 1} \Pi_{N,l}^1(\hat{l}) \to 0$ (as beliefs go to zero when α_B goes to one). Hence, there always exists $\breve{\alpha}_B < 1$ such that for any $\alpha_B > \breve{\alpha}_B$, $\Pi_{N,l}^1(\hat{h}) > \Pi_{N,l}^1(\hat{l})$, and consequently, $\Delta_{N,l}^1 > 0$.

Now, to prove the existence of $\breve{\alpha}_B$, first note that $\Delta^0_{N,l} = f\left(\lambda_W(\hat{h},0),1\right) - f\left(\lambda_W(\hat{l},0),\lambda_{\bar{B}}(\hat{l},0)\right)$. If the wise type judge plays the honest strategy, i.e. $(\sigma^\mu_W(l)^*,\sigma^\mu_W(h)^*) = (1,0)$, a sufficient condition for $\Delta^0_{N,l} > 0$ is $\lambda_W(\hat{h},0) > \lambda_W(\hat{l},0)$, which is satisfied if and only if $(1-\sigma(l)-\sigma_N(h))\alpha_N < \alpha_B$ (see condition (13)). Now, since $(1-\sigma_N(l)-\sigma_N(h))\alpha_N < 1$, there always exists $\breve{\alpha}_B < 1$ for which $(1-\sigma_N(l)-\sigma_N(h))\alpha_N < \breve{\alpha}_B$. Consequently, for any $\alpha_B > \breve{\alpha}_B$, $\Delta^0_{N,l} > 0$.

Therefore, if the wise type judge plays the honest strategy, then for all $\alpha_B > \max\{\breve{\alpha}_B,\breve{\alpha}_B\}$, $\Delta^\mu_{N,l} = (1-\mu)\Delta^0_{N,l} + \mu\Delta^1_{N,l} > 0$ for any $\mu \in [0,1]$, which implies that $\sigma^\mu_N(l)^* = 0$. Thus, in this case, the unique equilibrium strategy of the normal type judge is $(\sigma^\mu_N(l)^*,\sigma^\mu_N(h)^*) = (0,0)$. Note also that if the wise type judge does not play the honest strategy, i.e. $(\sigma^\mu_W(l)^*,\sigma^\mu_W(h)^*) \neq (1,0)$, then by Corollary 1, the unique equilibrium strategy of the normal type judge is $(\sigma^\mu_N(l)^*,\sigma^\mu_N(h)^*) = (0,0)$. Consequently, in this case, if $\alpha_B > \max\{\breve{\alpha}_B,\breve{\alpha}_B\}$, then for any strategy of the wise type judge, the unique equilibrium strategy of the normal type judge is $(\sigma^\mu_N(l)^*,\sigma^\mu_N(h)^*) = (0,0)$.

To complete the proof of point *1.* note that, by Lemma 3, if the strategy of the normal type judge is $(\sigma^\mu_N(l)^*,\sigma^\mu_N(h)^*) = (0,0)$, we know that there exists threshold $\widetilde{\alpha}_B$ for the wise type judge such that, for all $\alpha_B > \widetilde{\alpha}_B$, the wise type always takes action $a = \hat{h}$. Let $\alpha^{max}_B = \max\{\breve{\alpha}_B,\breve{\alpha}_B,\widetilde{\alpha}_B\}$, and note that $\alpha^{max}_B = \widetilde{\alpha}_B$ (otherwise, we could have a situation in which the normal type judge uses an informative strategy and the wise type judge uses a non-informative strategy, which contradicts point *1.* of Proposition 1). Therefore, for all $\alpha_B > \alpha^{max}_B$, there is a unique equilibrium in which both the normal type judge and the wise type always take action $a = \hat{h}$.

Second, we prove point *2.* of the proposition. First note that by Corollary 1, in any informative equilibrium the wise type judge always uses the honest strategy, i.e. $(\sigma^\mu_W(l)^*,\sigma^\mu_W(h)^*) = (1,0)$. Thus, we consider $(\sigma^\mu_W(l)^*,\sigma^\mu_W(h)^*) = (1,0)$ and show that for all $\mu \in [0,1]$ and $\alpha_B < \alpha^{min}_B$, $\Delta^\mu_{N,l} < 0$. First, note that condition $\Delta^\mu_{N,l} = (1-\mu)\Delta^0_{N,l} + \mu\Delta^1_{N,l} < 0$ is equivalent to $\Delta^0_{N,l} < \frac{\mu}{1-\mu}(-\Delta^1_{N,l})$. We use two lemmas to prove the result: Lemma 5 and Lemma 6, which are stated and proven below. Lemma 5 shows that for $\varepsilon_1 > 0$, there always exists $\alpha_{B(\varepsilon_1)} > 0$ such that for any $\alpha_B < \alpha_{B(\varepsilon_1)}$, in equilibrium $\Delta^0_{N,l} < \varepsilon_1$. Lemma 6 shows that there exist $\varepsilon_2 > 0$ and $\alpha_{B(\varepsilon_2)} > 0$ such that for any $\alpha_B < \alpha_{B(\varepsilon_2)}$, $\Delta^1_{N,l} < -\varepsilon_2$. Therefore, if we take $\varepsilon_1 < \frac{\mu}{1-\mu}\varepsilon_2$ and $\alpha^{min}_B = \min\{\alpha_{B(\varepsilon_1)},\alpha_{B(\varepsilon_2)},\widetilde{\alpha}_B\}$, then for any $\alpha_B < \alpha^{min}_B$, we have $\Delta^\mu_{N,l} = (1-\mu)\Delta^0_{N,l} + \mu\Delta^1_{N,l} < 0$, and consequently, $\sigma^\mu_N(l)^* = 1$.

Lemma 5. *For any $\varepsilon_1 > 0$, there always exists $\alpha_{B(\varepsilon_1)} > 0$ such that $\forall \alpha_B < \alpha_{B(\varepsilon_1)}$, in equilibrium, $\Delta^0_{N,l} < \varepsilon_1$.*

Proof

Note that $\Delta^0_{N,l} = f\left(\lambda_W(\hat{h},0),1\right) - f\left(\lambda_W(\hat{l},0),\lambda_{\bar{B}}(\hat{l},0)\right)$. From the beliefs in Table 2, we obtain $\frac{\lambda_{\bar{B}}(\hat{l},0)}{\alpha_B} < 0$, with $lim_{\alpha_B\to 0}\lambda_{\bar{B}}[\hat{l},0] \to 1$. From the beliefs in Table 1, we obtain $lim_{\alpha_B\to 0}\lambda_W(\hat{h},0) \leq lim_{\alpha_B\to 0}\lambda_W(\hat{l},0) \iff \alpha_W + \sigma_N(l)\alpha_N \leq \alpha_W + (2-\sigma_N(l))\alpha_N \iff \sigma_N(l) \leq 1$. Hence, for any $\varepsilon_1 > 0$, there exists $\alpha_{B(\varepsilon_1)} > 0$, such that for any $\alpha_B < \alpha_{B(\varepsilon_1)}$, $\Delta^0_{N,l} < \varepsilon_1$. ♦

Lemma 6. *There exists* $\varepsilon_2 > 0$ *and* $\alpha_{B(\varepsilon_2)} > 0$*, such that* $\forall \alpha_B < \alpha_{B(\varepsilon_2)}$, $\Delta^1_{N,l} < -\varepsilon_2$.

Proof

Note that $\Delta^1_{N,l} = \gamma f(0,1) + (1-\gamma) f\left(\lambda_W(\hat{h}, H), 1\right) - \left(\gamma f\left(\lambda_W(\hat{l}, L), \lambda_{\bar{B}}(\hat{l}, L)\right) + (1-\gamma) f\left(0, \lambda_{\bar{B}}[\hat{l}, H]\right)\right)$. By Proposition 2, $\sigma^{\mu}_N(h)^* = 0$. Assuming $\sigma^{\mu}_N(h)^* = 0$ and taking limits on some of the beliefs in Table 1, we obtain $lim_{\alpha_B \to 0} \lambda_W(\hat{h}, H) \geq \alpha_W > 0$ and $lim_{\alpha_B \to 0} \lambda_W(\hat{l}, L) \geq \alpha_W > 0$. Note also that $lim_{\alpha_B \to 0} \lambda_W(\hat{h}, H) > lim_{\alpha_B \to 0} \lambda_W(\hat{l}, L) \iff \gamma + (1-\gamma)(1 - \sigma_N(l)) > \gamma \sigma_N(l)$. Therefore, if $\sigma_N(l) < 1$, then $lim_{\alpha_B \to 0} \lambda_W(\hat{h}, H) > lim_{\alpha_B \to 0} \lambda_W(\hat{l}, L)$, and if $\sigma_N(l) = 1$, then $lim_{\alpha_B \to 0} \lambda_W(\hat{h}, H) = lim_{\alpha_B \to 0} \lambda_W(\hat{l}, L)$. Let us denote $k = lim_{\alpha_B \to 0} \lambda_W(\hat{h}, H)$ and $K = lim_{\alpha_B \to 0} \lambda_W(\hat{l}, L)$.

Now, taking limits on some beliefs in Table 2, we obtain $lim_{\alpha_B \to 0} \lambda_{\bar{B}}(\hat{l}, L) \to 1$ and $lim_{\alpha_B \to 0} \lambda_{\bar{B}}[\hat{l}, H] \to 1$.

Using these results, we have $lim_{\alpha_B \to 0} \Delta^1_{N,l} = \gamma f(0,1) + (1-\gamma) f(k,1) - (\gamma f(K,1) + (1-\gamma) f(0,1))$. Let $z = lim_{\alpha_B \to 0} \Delta^1_{N,l}$. First, note that $\gamma f(0,1) + (1-\gamma) f(k,1) < \gamma f(K,1) + (1-\gamma) f(0,1)$, as $\gamma > \frac{1}{2}$ and $f(K,1) \geq f(k,1)$. Thus, $z < 0$. Now, let $\varepsilon_2 > 0$ be any number in the interval $(0, |z|)$. Then, $lim_{\alpha_B \to 0} \Delta^1_{N,l} < -\varepsilon_2$. Thus, there always exists $\alpha_{B(\varepsilon_2)} > 0$, such that if $\alpha_B < \alpha_{B(\varepsilon_2)}$, then $\Delta^1_{N,l} < -\varepsilon_2$. ♦

This completes the proof of Proposition 3. **QED**.

Remark 1 on Proposition 3. When $\alpha_B > \alpha_B^{max}$ and $\mu = 0$, there are multiple equilibria. According to Definition 2, in all the equilibria that are robust to transparency, we have $\sigma^{\mu}_N(l)^* = 0$.

Remark 2 on Proposition 3. When $\mu = 0$, in equilibrium, we always have $\sigma^{\mu}_N(l)^* < 1$. Briefly, when $\mu \to 0$, $\alpha_B^{min} \to 0$. Therefore, there is no α_B that is low enough to sustain an equilibrium in which $\sigma^{\mu}_N(l)^* = 1$.

Proof of Proposition 4.

In the proof, we use two simplifications, which are without loss of generality:

- We consider $\eta = \beta = 1$ and prove the result for this case. Note that $\Delta^{\mu}_l \geq 0$ when $\eta = \beta = 1$ if and only if $\Delta^{\mu}_l \geq 0$ when $\eta = \beta$. See expression (7) with $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X)) = \eta \lambda_W(a, X) + \beta \lambda_{\bar{B}}(a, X)$. Hence, the analysis that follows proves the result for any $\eta = \beta > 0$.

- We use the result in Proposition 2 and Definition 2 and, therefore, consider $\sigma^{\mu}_N(h)^* = 0 \ \forall \mu \in [0,1]$.

We start noting that, by Corollary 1, if the wise type judge does not use the honest strategy, then $(\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*) = (0,0)$ for all μ. Hence, if we want to analyze whether an informative equilibrium exists, we have to consider that the wise type judge uses the honest strategy, i.e., $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$. This is what we do in points *1.* and *2.* of Proposition 4. It is also what we assume in order to prove point *3.* of the Proposition. The difference is that, in this case, we obtain that the equilibrium strategy of the normal type judge is $(\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*) = (0,0)$, which is non-informative. To complete the proof of this point, we show that $\alpha''_B < \widetilde{\alpha}_B$ (see Lemma 10 below). Then, by Proposition 1, in equilibrium either $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (0,0)$, $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$ or $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (x_1, 0)$, with $x_1 \in (0,1)$.

Next, we move on to prove Proposition 4, which requires a thorough analysis of expression $\Delta^{\mu}_{N,l} = (1-\mu)\Delta^{0}_{N,l} + \mu\Delta^{1}_{N,l}$. This analysis is structured in three lemmas: Lemmas: 7, 8 and 9.

Lemma 8 analyzes the behavior of $\Delta^{\mu}_{N,l}$ when $\mu = 1$. It shows that $\Delta^{1}_{N,l} > 0$ for any $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in (\frac{1}{2}, \gamma')$, with $\alpha'_B = \frac{\alpha_W\left(2+\alpha_W+\alpha_W^2-2\sqrt{2}\sqrt{\alpha_W(\alpha_W+1)}\right)}{(\alpha_W+2)^2}$, $\alpha''_B = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4}$ and $\gamma' = \frac{\left(\alpha_B^2-\alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W}-(2\alpha_B+\alpha_W)(1-\alpha_B+\alpha_W)}$. Lemma 7 focuses on the analysis of $\Delta^{\mu}_{N,l}$ when $\mu = 0$. It shows that when $\alpha_B < \alpha''_B$, hence when $\alpha_B \in (\alpha'_B, \alpha''_B)$, the following occurs: *1)*: $\frac{\partial \Delta^{0}_{N,l}}{\partial \sigma_N(l)} > 0$, *2)*: $\Delta^{0}_{N,l}\Big|_{\sigma_N(l)=0} < 0$, *3)*: $\Delta^{0}_{N,l}\Big|_{\sigma_N(l)=1} > 0$, and *4)*: $\Delta^{0}_{N,l}\Big|_{\sigma^0_N(l)^*} = 0$, with $\sigma^0_N(l)^* = 1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}$. Consequently, $\Delta^{0}_{N,l} < 0$ if $\sigma_N(l) < \sigma^0_N(l)^*$, and $\Delta^{0}_{N,l} > 0$ if $\sigma_N(l) > \sigma^0_N(l)^*$. Finally, Lemma 9 focuses on $\sigma^{\mu}_N(l)^{Sup*}$, which denotes the highest equilibrium probability with which the normal type judge takes action $a = \hat{l}$ after signal l. Note that in the most informative equilibrium, $\sigma^{\mu}_N(l)^* = \sigma^{\mu}_N(l)^{Sup*}$. We refer to $\sigma^{\mu}_N(l)^{Sup*}$ as the supremum. This lemma shows that $\sigma^{\mu}_N(l)^{Sup*}$ decreases in μ.

The results in these lemmas allow us to state that if $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in \left(\frac{1}{2}, \gamma'\right)$:

i) There always exist $\mu'' < 1$ such that if $\mu > \mu''$, $\Delta^{\mu}_{N,l} = (1-\mu)\Delta^{0}_{N,l} + \mu\Delta^{1}_{N,l} > 0$ and if $\mu < \mu''$, $\Delta^{\mu}_{N,l}$ has, at least, one root. This is so because $\Delta^{1}_{N,l} > 0$ and $\Delta^{0}_{N,l}$ is increasing in $\sigma_N(l)$, with $\Delta^{0}_{N,l}\Big|_{\sigma_N(l)=0} < 0$ and $\Delta^{0}_{N,l}\Big|_{\sigma_N(l)=1} > 0$. This implies that, if $\mu > \mu''$, in the unique equilibrium strategy of the normal type judge, $\sigma^{\mu}_N(l)^* = 0$. This result describes the equilibrium behavior of the normal type judge in point *3.* of Proposition 4.

ii) There always exists $\mu' > 0$ such that if $\mu < \mu'$, then $\sigma^{\mu}_N(l)^*$ is the unique root of $\Delta^{\mu}_{N,l}$ in the open interval $(0,1)$. This result follows from the fact that when μ is small enough, $\Delta^{\mu}_{N,l}$ must have a unique root, because of the described characteristics of $\Delta^{0}_{N,l}$, $\Delta^{1}_{N,l}$ and $\Delta^{\mu}_{N,l}$. Hence, $\Delta^{\mu}_{N,l} = (1-\mu)\Delta^{0}_{N,l} + \mu\Delta^{1}_{N,l} < 0$ for any $\sigma_N(l) < \sigma^{\mu}_N(l)^*$ and $\Delta^{\mu}_{N,l} = (1-\mu)\Delta^{0}_{N,l} + \mu\Delta^{1}_{N,l} > 0$ for any $\sigma_N(l) > \sigma^{\mu}_N(l)^*$. See Figure 1. This implies that in the equilibrium, we necessarily have $0 < \sigma^{\mu}_N(l)^* < \sigma^{0}_N(l)^*$. This result describes the equilibrium behavior of the normal type judge in point *1.* of Proposition 4. This equilibrium strategy of the normal type judge is unique. Note also that in this point of the Proposition, $\sigma^{\mu}_N(l)^* = \sigma'$, with $\sigma' = \sigma^{\mu}_N(l)^{Sup*}$, as σ' is the unique (then, the highest) equilibrium probability with which the normal type judge takes action $a = \hat{l}$ after signal l. Then, by Lemma 9, σ' decreases in μ.

iii) In addition, since $(\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*) = (\sigma'', 0)$ is the most informative equilibrium strategy of the normal type judge, then $\sigma'' = \sigma^{\mu}_N(l)^{Sup*}$. Then, by Lemma 9, σ'' decreases in μ, and $\sigma'' < \sigma'$.

Next, we prove these lemmas.

Lemma 7. *Consider $\mu = 0$. Let $\alpha''_B = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4} < \frac{1}{2}$.*

1. If $\alpha_B < \alpha''_B$, then in the unique equilibrium, $(\sigma^0_N(l)^*, \sigma^0_N(h)^*) = \left(1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}, 0\right)$.

2. If $\alpha_B \geq \alpha''_B$, then $(\sigma^0_N(l)^*, \sigma^0_N(h)^*) = (0,0)$ is the unique equilibrium strategy of the normal type judge.

Proof

The following results prove the lemma (they are proven below).

1) $\Delta^0_{N,l}\Big|_{\sigma_N(l)=1} > 0$.

2) $\Delta^0_{N,l}\Big|_{\sigma_N(l)=0} < 0 \iff \alpha_B < \alpha''_B = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4}$, where $\alpha''_B < \frac{1}{2}\alpha_W$; hence, $\alpha''_B < \frac{1}{2}$.

3) $\frac{\partial \Delta^0_{N,l}}{\partial \sigma_N(l)} > 0$ if $\alpha_B < \frac{1}{2}\alpha_W$.

4) $\Delta^0_{N,l} > 0$ if $\alpha_B > \frac{1}{2}\alpha_W$.

5) $\Delta^0_{N,l} = 0 \iff \sigma_N(l) = 1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}$.

First, note that result 3) implies that $\Delta^0_{N,l}$ is increasing in $\sigma_N(l)$ when $\alpha_B < \alpha''_B$. In addition, $\Delta^0_{N,l}\Big|_{\sigma_N(l)=1} > 0$ and $\Delta^0_{N,l}\Big|_{\sigma_N(l)=0} < 0$ if and only if $\alpha_B < \alpha''_B$. Therefore, if $\alpha_B \in (0, \alpha''_B)$, there is a unique equilibrium, which is $\Delta^0_{N,l}(\sigma^0_N(l)^*) = 0$, with $\sigma^0_N(l)^* = 1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}$. Second, note also that if $\alpha_B \in \left(\alpha''_B, \frac{1}{2}\alpha_W\right)$, then $\Delta^0_{N,l} > 0$ by results 2) and 3), and if $\alpha_B \in \left(\frac{1}{2}\alpha_W, 1\right)$, then $\Delta^0_{N,l} > 0$ by result 4). This second case implies that the optimal strategy of the normal type judge is $(\sigma^0_N(l)^*, \sigma^0_N(h)^*) = (0,0)$ for any $\alpha_B > \alpha''_B$. Now, by Lemma 3, we know that, in this case, the wise type judge could find it optimal to use a strategy that does not always follow the judge's signal. In this case, Proposition 8 guarantees that the normal type judge has a unique equilibrium strategy, which is to always take action $a = \hat{h}$.

Next, we prove the results stated at the beginning of the lemma and that allowed us to prove this lemma.

Proof of result *1)*: from equations (7) and (8) and the beliefs in Tables 1-2, when $\eta = \beta = 1$, we have $\Delta^0_{N,l}\Big|_{\sigma_N(l)=1} = \frac{2\alpha_W+\alpha_N}{\alpha_W+\alpha_N} - \frac{2\alpha_W+\alpha_N}{2\alpha_B+\alpha_W+\alpha_N} > 0$.

Proof of result *2)*: $\Delta^0_{N,l}\Big|_{\sigma_N(l)=0} = \frac{\alpha_W}{\alpha_W+2\alpha_N} + \frac{2\alpha_B-\alpha_W}{2\alpha_B+\alpha_W} = \frac{\alpha_W}{\alpha_W+2(1-\alpha_B-\alpha_W)} + \frac{2\alpha_B-\alpha_W}{2\alpha_B+\alpha_W} > 0 \iff -2\alpha_B^2 + \alpha_B\alpha_W + 2\alpha_B + \alpha_W^2 - \alpha_W > 0 \iff \alpha_B > \alpha''_B = \frac{1}{4}\left(2+\alpha_W - \sqrt{9\alpha_W^2-4\alpha_W+4}\right)$. Note that $\alpha''_B = \frac{1}{4}\left(2+\alpha_W - \sqrt{9\alpha_W^2-4\alpha_W+4}\right) < \frac{1}{2}\alpha_W \iff 2-\alpha_W < \sqrt{9\alpha_W^2-4\alpha_W+4}$, and $\sqrt{9\alpha_W^2-4\alpha_W+4} > \sqrt{\alpha_W^2-4\alpha_W+4} = 2-\alpha_W$.

Proof of result *3)*: $\frac{\partial \Delta^\mu_{N,l}}{\partial \sigma_N(l)} = \frac{\alpha_W\alpha_N}{(\alpha_W+2\alpha_N-\alpha_N\sigma_N(l))^2} + \frac{(\alpha_W-2\alpha_B)\alpha_N}{(2\alpha_B+\alpha_W+\alpha_N\sigma_N(l))^2} > 0$ when $\alpha_B < \frac{1}{2}\alpha_W$.

Proof of result *4)*: $\Delta^0_{N,l} = \frac{\alpha_W}{\alpha_W+(2-\sigma_N(l))\alpha_N} + \frac{2\alpha_B-\alpha_W}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N} > 0$.

Proof of result *5)*: $\Delta^0_{N,l} = \frac{\alpha_W}{\alpha_W+(2-\sigma_N(l))\alpha_N} + \frac{2\alpha_B-\alpha_W}{2\alpha_B+\alpha_W+\sigma_N(l)\alpha_N} = 0 \iff \sigma_N(l) = 1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}$. ♦

Lemma 8. *Consider* $\mu = 1$ *and* $\alpha''_B = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4}$. *For any* $\alpha_W \in (0,1)$, *there always exists* $\alpha'_B = \frac{\alpha_W\left(2+\alpha_W+\alpha_W^2-2\sqrt{2}\sqrt{\alpha_W(\alpha_W+1)}\right)}{(\alpha_W+2)^2} \in (0,1)$, *with* $\alpha'_B < \alpha''_B$, *such that for any* $\alpha_B \in (\alpha'_B, \alpha''_B)$, *there exists* $\gamma' = \frac{\left(\alpha_B^2-\alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W}-(2\alpha_B+\alpha_W)(1-\alpha_B+\alpha_W)} > \frac{1}{2}$. *Then, for any* $\alpha_B \in (\alpha'_B, \alpha''_B)$ *and* $\gamma \in (\frac{1}{2}, \gamma')$, *we have* $\Delta^1_{N,l} > 0$. *Hence, the unique equilibrium strategy of the normal type judge is* $(\sigma^1_N(l)^*, \sigma^1_N(h)^*) = (0,0)$.

Proof

From equation (7) and the beliefs in Tables 1-2, we have the following:

$\Delta^1_{N,l} = \frac{(1-\gamma)\alpha_W}{\alpha_W+(\gamma+(1-\gamma)(1-\sigma_N(l)))\alpha_N} - \frac{\gamma(\alpha_W-\alpha_B)}{\alpha_B+\alpha_W+\gamma\sigma_N(l)\alpha_N} + \frac{(1-\gamma)\alpha_B}{\alpha_B+(1-\gamma)\sigma_N(l)\alpha_N}$.

Note that if $\alpha_B > \alpha_W$, the expression above is positive, in which case $\sigma^1_N(l)^* = 0$. Hence, hereafter, we focus on $\alpha_B < \alpha_W$. Note that $\alpha''_B < \alpha_W$ as shown above.

For the case $\alpha_B < \alpha_W$, it is easy to obtain that $\Delta^1_{N,l} > 0$ if and only if the following second-degree

polynomial $p(1-\sigma_N(l))$ is positive:

$$
\begin{aligned}
p(\cdot) = {} & (1-\sigma_N(l))^2 2\gamma\alpha_N^2 (1-\gamma)^2 (\alpha_W - \alpha_B) \\
& + (1-\sigma_N(l)) \left(\alpha_N(1-\gamma)(-4\gamma^2\alpha_N\alpha_B + 4\gamma^2\alpha_N\alpha_W - 4\gamma\alpha_B\alpha_W + 2\gamma\alpha_N\alpha_B + 2\gamma\alpha_W^2 - 3\gamma\alpha_N\alpha_W + \alpha_B^2 - \alpha_W^2)\right) \\
& + \gamma(\alpha_W + \gamma\alpha_N)(\alpha_B - \alpha_N(\gamma-1))(\alpha_B - \alpha_W) - \alpha_W(\gamma-1)(\alpha_B - \alpha_N(\gamma-1))(\alpha_B + \alpha_W + \gamma\alpha_N) \\
& - \alpha_B(\alpha_W + \gamma\alpha_N)(\gamma-1)(\alpha_B + \alpha_W + \gamma\alpha_N).
\end{aligned} \tag{14}
$$

To facilitate the analysis, in the following, we denote $(1-\sigma_N(l))$ by z and refer to this polynomial, equation (14), as $p(z) = a(z)^2 + b(z) + c$. Note that if $p(z) > 0$, then $\Delta^1_{N,l} > 0$. The first derivative is $p'(z) = 2a(z) + b$, and the second one is $p''(z) = 2a$. Additionally, $p(z)$ is convex in z when $\alpha_W > \alpha_B$, as in this case $a = 2\gamma\alpha_N^2 (1-\gamma)^2 (\alpha_W - \alpha_B) > 0$. The first derivative provides the minimum value of $p(z)$: $p'(z_{\min}) = 2az_{\min} + b = 0 \iff z_{\min} = \frac{-b}{2a}$. Consequently, if $p(z_{\min}) > 0$, then $p(z) > 0$, and therefore, $\Delta^1_{N,l} > 0$.

Next, we determine the conditions for $p(z_{\min}) > 0$. Note that $p(z_{\min}) = a(\frac{-b}{2a})^2 + b(\frac{-b}{2a}) + c = c - \frac{b^2}{4a}$. From equation (14), we can obtain the value of $c - \frac{b^2}{4a}$, which is positive if the next polynomial in γ is positive:

$$
\begin{aligned}
pol(\gamma) = {} & \gamma^2(8\alpha_B^3\alpha_W - 4\alpha_B^2\alpha_N^2 - 8\alpha_B\alpha_W^3 + 4\alpha_B\alpha_W\alpha_N^2 - 4\alpha_W^4 - 4\alpha_W^3\alpha_N - \alpha_W^2\alpha_N^2) \\
& + \gamma\,(2(\alpha_W - \alpha_B)(2\alpha_B + \alpha_W)(\alpha_B + \alpha_W)(2\alpha_W + \alpha_N)) \\
& - (\alpha_B^2 - \alpha_W^2)^2
\end{aligned} \tag{15}
$$

Note that $pol(\gamma) > 0 \iff p(z_{\min}) > 0 \Rightarrow p(z) > 0 \iff \Delta^1_{N,l} > 0$. Then, we determine the conditions for $pol(\gamma) > 0$. We can easily check that under condition $\alpha_B < \alpha_W$:

1. $pol(\gamma)$ is a concave function in γ.
2. $\left.\frac{dpol(\gamma)}{d\gamma}\right|_{\gamma=\frac{1}{2}} < 0$

Now, since $pol(\gamma)$ is concave in $\gamma \in (\frac{1}{2}, 1)$ and decreasing in $\gamma = \frac{1}{2}$, polynomial $pol(\gamma)$ is necessarily decreasing in $\gamma \in (\frac{1}{2}, 1)$. Hence, there are only two possibilities: i) $pol(\gamma)$ is negative for all $\gamma \in \left(\frac{1}{2}, 1\right)$, and ii) $pol(\gamma)$ is positive for all $\gamma \in (\frac{1}{2}, \gamma')$ and negative for all $\gamma \in (\tilde{\gamma}, 1)$, where γ' is the greatest real root of $pol(\gamma) = 0$. If $pol(\gamma = \frac{1}{2}) > 0$, we are in the second case.

Now, from equation (15) and setting $\alpha_N = 1 - \alpha_W - \alpha_B$, we obtain the following:

$$
pol(\gamma = \tfrac{1}{2}) = \left(-\tfrac{1}{4}\alpha_W^2 - \alpha_W - 1\right)\alpha_B^2 + \left(\tfrac{1}{2}\alpha_W^3 + \tfrac{1}{2}\alpha_W^2 + \alpha_W\right)\alpha_B + \left(\tfrac{1}{2}\alpha_W^3 - \tfrac{1}{4}\alpha_W^4 - \tfrac{1}{4}\alpha_W^2\right).
$$

The expression above is concave in α_B and has the following two real roots:

$$
\alpha'_B = \frac{\left(2\alpha_W + \alpha_W^2 + \alpha_W^3 - 2\sqrt{2}\sqrt{\alpha_W^4 + \alpha_W^3}\right)}{(\alpha_W+2)^2} \text{ and } \alpha_B^{*(+)} = \frac{\left(2\alpha_W + \alpha_W^2 + \alpha_W^3 + 2\sqrt{2}\sqrt{\alpha_W^4 + \alpha_W^3}\right)}{(\alpha_W+2)^2}.
$$

Consequently, if $\alpha_B \in \left(\alpha'_B, \alpha_B^{*(+)}\right)$, $pol(\gamma = \frac{1}{2}) > 0$, and then, $pol(\gamma)$ is positive for all $\gamma \in (\frac{1}{2}, \gamma')$ and negative for all $\gamma \in (\gamma', 1)$, where γ' is the greatest real root of $pol(\gamma)$, i.e., $\gamma' = \frac{\left(\alpha_B^2 - \alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W} - (2\alpha_B + \alpha_W)(1 - \alpha_B + \alpha_W)}$.

To conclude the proof of Lemma 8, we need to show that for any α_W, $\alpha'_B < \alpha''_B < \alpha_B^{*(+)}$. To prove this, note that after some algebra, we obtain $\alpha''_B < \alpha_B^{*(+)} \iff \frac{9\alpha_W^6 + 68\alpha_W^5 + 184\alpha_W^4 + 208\alpha_W^3 + 64\alpha_W^2}{(\alpha_W + 2)^4} > 0$, which always holds. We also obtain $\alpha'_B < \alpha''_B \iff 3\alpha_W + 3\alpha_W^2 - 2 < \sqrt{8\alpha_W^2 + 8\alpha_W}$. Note that the expression on the left-hand side of the inequality is a convex function in α_W, whereas the expression on the right-hand side is concave in α_W. Additionally, the value of the left-hand side expression evaluated at $\alpha_W = 0$ is smaller than the value of the right-hand side expression evaluated at $\alpha_W = 0$, and both are equal at $\alpha_W = 1$. Now, since $\alpha_W \in (0, 1)$, the inequality holds.

In summary, if $\mu = 1$, $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in (\frac{1}{2}, \gamma')$, then $pol(\gamma) > 0$. Hence, $p(z) = p(1 - \sigma_N(l)) > 0$ and $\Delta^1_{N,l} > 0$. Consequently, in equilibrium, $\sigma^1_N(l)^* = 0$. Finally, by Lemma 3 and Proposition 8, the proof follows. This concludes the proof of Lemma 8. ♦

Lemma 9. *$\sigma^\mu_N(l)^{Sup*}$ decreases as μ increases.*

We first define $\sigma^\mu_N(l)^{Sup*}$ as the highest equilibrium probability with which the normal type judge takes action $a = \hat{l}$ after signal l. We refer to $\sigma^\mu_N(l)^{Sup*}$ as the supremum (of all $\sigma^\mu_N(l)^*$).

Now, note that if $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in (\frac{1}{2}, \gamma')$, then $\Delta^1_{N,l} > 0$ for all $\sigma_N(l) \in (0, 1)$. In addition, $\Delta^0_{N,l} < 0$ for all $\sigma_N(l) \in (0, \sigma^0_N(l)^*)$, and $\Delta^0_{N,l} > 0$ for all $\sigma_N(l) \in (\sigma^0_N(l)^*, 1)$ (see Lemmas 7 and 8). Therefore, $\Delta^\mu_{N,l} = (1 - \mu)\Delta^0_{N,l} + \mu\Delta^1_{N,l} > 0$ for all $\sigma_N(l) \in (\sigma^0_N(l)^*, 1)$, which implies $\sigma^\mu_N(l)^* < \sigma^0_N(l)^*$ for any $\mu > 0$, hence $\sigma^\mu_N(l)^{Sup*} < \sigma^0_N(l)^*$ for any $\mu > 0$. Then, $\sigma^\mu_N(l)^{Sup*} \in (0, \sigma^0_N(l)^*)$. In addition, $\Delta^\mu_{N,l} > 0$ necessarily for all $\sigma^\mu_N(l) \in (\sigma^\mu_N(l)^{Sup*}, \sigma^0_N(l)^*)$.

Now, note that $\frac{\partial \Delta^\mu_{N,l}}{\partial \mu} = -\Delta^0_{N,l} + \Delta^1_{N,l} > 0$ for all $\sigma_N(l) \in (0, \sigma^0_N(l)^*)$ since $\Delta^1_{N,l} > 0$ for all $\sigma_N(l) \in (0, 1)$ and $\Delta^0_{N,l} < 0$ for all $\sigma_N(l) \in (0, \sigma^0_N(l)^*)$.

Therefore, if μ increases, $\Delta^\mu_{N,l}$ increases, and as $\Delta^\mu_{N,l} > 0$ for all $\sigma_N(l) \in (\sigma^\mu_N(l)^{Sup*}, \sigma^0_N(l)^*)$, $\sigma^\mu_N(l)^{Sup*}$ has to decrease with μ. ♦

Lemma 10. $\alpha''_B < \widetilde{\alpha}_B$.

From the proof of Lemma 3, we know that when the normal type judge uses a non-informative strategy, $\frac{\partial \Delta^\mu_{W,l}}{\partial \sigma_W(l)} < 0$, and $\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=0} > 0$. Additionally, in this case, $\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=1} \geq 0 \iff \alpha_B \geq \widetilde{\alpha}_B$.

Also, from expression (11), when $f\left(\lambda_W(a, X), \lambda_{\bar{B}}(a, X)\right) = \lambda_W(a, X) + \lambda_{\bar{B}}(a, X)$, we have the following:

$$\begin{aligned}\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=1} &= (1 - \mu)\left(\tfrac{2(1-\alpha_B)}{2 - 2\alpha_B - \alpha_W} - \tfrac{2\alpha_W}{\alpha_W + 2\alpha_B}\right) + \mu\left(\tfrac{\alpha_B - \alpha_W}{\alpha_B + \alpha_W}\right) \\ &= -\mu\left(\tfrac{\alpha_W - \alpha_B}{\alpha_B + \alpha_W}\right) - (1 - \mu)\left(\tfrac{2\alpha_W}{\alpha_W + 2\alpha_B} - \tfrac{2(1-\alpha_B)}{2 - 2\alpha_B - \alpha_W}\right),\end{aligned}$$

where

$\frac{2\alpha_W}{\alpha_W + 2\alpha_B} - \frac{2(1-\alpha_B)}{2 - 2\alpha_B - \alpha_W} > 0 \iff \alpha_B \leq \frac{2 + \alpha_W - \sqrt{9\alpha_W^2 - 4\alpha_W + 4}}{4} = \alpha''_B$. In addition, $\alpha''_B < \alpha_W$; hence, $-\mu\left(\frac{\alpha_W - \alpha_B}{\alpha_B + \alpha_W}\right) < 0$ for any $\alpha_B \leq \alpha''_B$.

Consequently, if $\alpha_B \leq \alpha''_B$, then $\Delta^{\mu}_{W,l}\Big|_{\sigma_W(l)=1} < 0$, which implies $\alpha''_B < \widetilde{\alpha}_B$. ♦

This completes the proof of Proposition 4. **QED**.

Proof of Proposition 5

By Proposition 2 and Definition 2, $\sigma^{\mu}_N(h)^* = 0 \ \forall \mu \in [0,1]$.

By Corollary 1, if the wise type judge does not use the honest strategy, then $(\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*) = (0,0)$ for all μ. Hence, if we want to analyze whether an informative equilibrium exists, we have to consider that the wise type judge uses the honest strategy, i.e., $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$. This is what we do in point *1.* of Proposition 5. It is also what we assume in order to prove point *2.* of this proposition. The difference is that, in the latter case, we obtain that the equilibrium strategy of the normal type judge is $(\sigma^{\mu}_N(l)^*, \sigma^{\mu}_N(h)^*) = (0,0)$, which is non-informative. Then, by Proposition 1, in equilibrium either $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (0,0)$, $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (1,0)$ or $(\sigma^{\mu}_W(l)^*, \sigma^{\mu}_W(h)^*) = (x_1,0)$, with $x_1 \in (0,1)$.

To prove Proposition 5 we analyze the expression $\Delta^{\mu}_{N,l} = (1-\mu)\Delta^0_{N,l} + \mu\Delta^1_{N,l}$ in the two extreme cases $\mu = 0$ and $\mu = 1$. We consider expression (7), with $f\left(\lambda_W(a,X), \lambda_{\bar{B}}(a,X)\right) = \eta\lambda_W(a,X) + \lambda_{\bar{B}}(a,X)$.

Let us first define some thresholds:

$\bar{\alpha}_B = 2\sqrt{2}\sqrt{\alpha_W^2 + \alpha_W} - 2\alpha_W$,

$\bar{\gamma} = \frac{1}{2}\frac{\alpha_B+\alpha_W}{2\alpha_B+\alpha_W}\left(1 + \sqrt{2 + \frac{2}{(2\alpha_B+\alpha_W-2)}}\right)$,

$\bar{\eta}' = \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$,

$\bar{\eta}'' = \frac{\alpha_B\left(2\sqrt{2}\gamma\sqrt{(\gamma(2\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W)(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1))}+(\gamma(2\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W)(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1))+2\gamma^2\right)}{\alpha_W\left(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1)+\gamma\right)^2}$.

Next, Lemma 11 analyzes the behavior of $\Delta^{\mu}_{N,l}$ when $\mu = 0$, and Lemma 12 does it when $\mu = 1$.

Lemma 11. *Consider $\sigma_N(l) = 0$. Then, $\Delta^0_{N,l} > 0$ if $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \bar{\eta}'$.*

Proof

From (7), we have:

$\Delta^0_{N,l}\Big|_{\sigma_N(l)=0} > 0 \iff \frac{2\alpha_B(2\alpha_B+\alpha_W-2)-2\alpha_W\eta(2\alpha_B+\alpha_W-1)}{(2\alpha_B+\alpha_W-2)(2\alpha_B+\alpha_W)} > 0.$

Simple algebra shows that this expression is greater than zero when $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$, with $\bar{\eta}' = \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$. ♦

Lemma 12. *$\Delta^1_{N,l} > 0$ if $\alpha_B < \bar{\alpha}_B$, $\gamma < \bar{\gamma}$ and $\eta \in (\bar{\eta}', \bar{\eta}'')$, with $0 < \bar{\eta}' < \bar{\eta}'' < 1$.*

Proof

From (7), we have:

$\Delta^1_{N,l} > 0 \iff \frac{\alpha_W(\gamma-1)\eta}{(\gamma-1)\sigma_N(l)(\alpha_B+\alpha_W-1)+\alpha_B-1} + \frac{\gamma(\alpha_B-\alpha_W\eta)}{\alpha_B+\alpha_W-\gamma\sigma_N(l)(\alpha_B+\alpha_W-1)} + \frac{\alpha_B-\alpha_B\gamma}{(\gamma-1)\sigma_N(l)(\alpha_B+\alpha_W-1)+\alpha_B} > 0.$

This expression can be written as $\frac{Nu}{De}$, with:

$Nu = -2(1-\gamma)^2\gamma(\alpha_B+\alpha_W-1)^2(\alpha_B-\alpha_W\eta)\sigma_N(l)^2 - (\gamma-1)\sigma_N(l)(\alpha_B+\alpha_W-1)(\alpha_W\eta(-\alpha_B(2\gamma+1)+\alpha_W(\gamma-1)+\gamma)+\alpha_B(2\alpha_B\gamma+\alpha_B-(\alpha_W+2)\gamma+\alpha_W))+\alpha_B\alpha_W\eta(\alpha_B-(\alpha_W+1)\gamma+\alpha_W)-(\alpha_B-1)\alpha_B(\alpha_B-\alpha_W\gamma+\alpha_W)$,

$De = ((\gamma-1)\sigma_N(l)(\alpha_B+\alpha_W-1)+\alpha_B-1)((\gamma-1)\sigma_N(l)(\alpha_B+\alpha_W-1)+\alpha_B)(\gamma\sigma_N(l)(\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W)$.

It can be shown that the denominator is always positive. Hence, if the numerator is positive, $\Delta^1_{N,l}>0$.

We note that when $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \bar{\eta}'$, $Nu(\sigma_N(l))$ is a convex function in $\sigma_N(l)$. Thus, we next calculate the minimum of $Nu(\sigma_N(l))$, that we denote by $\sigma^M_N(l)$. The value of this function evaluated at the minimum is $Nu(\sigma^M_N(l)) = \frac{Nu'}{De'}$, with:

$Nu' = -2\alpha_B\alpha_W\eta\left(\gamma^2\left(4\alpha_B^2+\alpha_B(4\alpha_W-2)+(\alpha_W-1)\alpha_W+2\right)-\gamma(4\alpha_B+2\alpha_W-1)(\alpha_B+\alpha_W)+(\alpha_B+\alpha_W)^2\right)+\alpha_B^2(-\gamma(2\alpha_B+\alpha_W-2)+\alpha_B+\alpha_W)^2+\alpha_W^2\eta^2(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1)+\gamma)^2$,

$De' = 8\gamma(\alpha_B-\alpha_W\eta)$.

Note that if $Nu(\sigma^M_N(l))\geq 0$, then $Nu(\sigma_N(l))>0$ for any $\sigma_N(l)\in(0,1)$, because of the convexity of $Nu(\sigma_N(l))$. It can be shown that, on the one hand, $Nu(\sigma^M_N(l))=0$ when $\eta=\bar{\eta}''$, with $\bar{\eta}''$ as defined above. On the other hand, it can be shown that $Nu(\sigma^M_N(l))$ is decreasing in η when $\alpha_B<\frac{1-\alpha_W}{2}$ and $\eta>\bar{\eta}'$. In addition, if $\gamma<\bar{\gamma}$ and $\alpha_B<\bar{\alpha}_B$, with $\bar{\alpha}_B = 2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W}-2\alpha_W$, then $\bar{\eta}'<\bar{\eta}''$. Note that $\bar{\alpha}_B<\frac{1-\alpha_W}{2}$. Concluding, if $\alpha_B<2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W}-2\alpha_W$, $\gamma<\bar{\gamma}$, and $\eta\in(\bar{\eta}',\bar{\eta}'')$, then $Nu(\sigma^M_N(l))>0$, which implies $Nu(\sigma_N(l))>0$ and $\Delta^1_{N,l}>0$. ♦

Now, by the continuity of function $\Delta^\mu_{N,l}=(1-\mu)\Delta^0_{N,l}+\mu\Delta^1_{N,l}$, if $\Delta^0_{N,l}\Big|_{\sigma_N(l)=0}>0$, there exits $\bar{\mu}'>0$ such that for all $\mu\in(0,\bar{\mu}')$, $\Delta^\mu_{N,l}\Big|_{\sigma_N(l)=0}>0$, which implies $\sigma^\mu_N(l)^*>0$ for all $\mu<\bar{\mu}'$. Analogously, if $\Delta^1_{N,l}>0$, then there exits $\bar{\mu}''>0$ such that if $\mu\in(\bar{\mu}'',1)$, then $\Delta^\mu_{N,l}>0$, which implies $\sigma^\mu_N(l)^*=0$ for all $\mu>\bar{\mu}''$.

To complete the proof, Lemma 13 below shows that $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W}-2\alpha_W<\widetilde{\alpha}_B$, which implies that if $\sigma^\mu_N(l)^*=0$ and $\alpha_B<\bar{\alpha}_B$, then the equilibrium behavior of the wise type is described by point *2.b* of Proposition 1.

Lemma 13. $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W}-2\alpha_W<\widetilde{\alpha}_B$.

From the proof of Lemma 3, we know that when the normal type judge uses a non-informative strategy, $\frac{\partial\Delta^\mu_{W,l}}{\partial\sigma_W(l)}<0$, and $\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=0}>0$. Additionally, in this case, $\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=1}\geq 0 \iff \alpha_B\geq\widetilde{\alpha}_B$.

Also, from expression (11), when $f(\lambda_W(a,X),\lambda_{\bar{B}}(a,X)) = \eta\lambda_W(a,X)+\lambda_{\bar{B}}(a,X)$, we have the following:

$$\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=1} = (1-\mu)\left(\frac{\alpha_W\eta}{2-2\alpha_B-\alpha_W}-\frac{\alpha_W\eta}{2\alpha_B+\alpha_W}+\frac{2\alpha_B}{2\alpha_B+\alpha_W}\right)+\mu\left(\frac{\alpha_B-\alpha_W\eta}{\alpha_B+\alpha_W}\right).$$

It can be shown that:

1) $\frac{\alpha_W\eta}{2-2\alpha_B-\alpha_W}-\frac{\alpha_W\eta}{2\alpha_B+\alpha_W}+\frac{2\alpha_B}{2\alpha_B+\alpha_W}<0$ if and only if $\alpha_B<\frac{1-\alpha_W}{2}$ and $\eta>\frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$.

2) $\frac{\alpha_B-\alpha_W\eta}{\alpha_B+\alpha_W}<0$ if and only if $\alpha_B<\frac{1-\alpha_W}{2}$ and $\eta>\frac{\alpha_B}{\alpha_W}$.

3) $\frac{\alpha_B}{\alpha_W}<\frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$ if and only if $\alpha_B<\frac{1-\alpha_W}{2}$.

Therefore, if $\alpha_B<\frac{1-\alpha_W}{2}$ and $\eta>\frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$, then $\Delta^\mu_{W,l}\Big|_{\sigma_W(l)=1}<0$.

To conclude, note that $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W}-2\alpha_W<\frac{1-\alpha_W}{2}$.

This completes the proof of Proposition 5. **QED**.

Proof of Proposition 6 Let $\theta' = \hat{\theta} - \epsilon$ and $\theta'' = \hat{\theta} + \epsilon$, with $\hat{\theta} = 1/2$ and $\epsilon \sim 0$. For a given μ, from the continuity of the payoff function, the result follows. **QED**.

Proof of Proposition 7 For all $\mu > 0$ and $s \in \{l, h\}$, now expression (4) corresponds to $\Delta^{\mu}_{B,s} = \Pi^{\mu}_{N,s}(\hat{h}) - (\Pi^{\mu}_{N,s}(\hat{l}) + \phi)$. This can be rewritten as $\Delta^{\mu}_{B,s} = \Delta^{\mu}_{N,s} - \phi$, with $\Delta^{\mu}_{N,s}$ given by expressions (7)-(8). Then, for all $\phi > 0$, $\Delta^{\mu}_{N,s} > \Delta^{\mu}_{B,s}$. Given $s' \in \{l, h\}$, this means that if $\Delta^{\mu}_{N,s'} \leq 0$, then $\Delta^{\mu}_{B,s'} < 0$; hence $\sigma^{\mu}_{B}(s')^* = 1$. Additionally, if $\Delta^{\mu}_{N,s'} > 0$, then $\Delta^{\mu}_{B,s'} \lesseqgtr 0$. Then, there exists $\hat{\phi} > 0$ such that for all $\phi > \hat{\phi}$, $\Delta^{\mu}_{B,s'} < 0$; hence $\sigma^{\mu}_{B}(s')^* = 1$ for all $s' \in \{l, h\}$. **QED**.

Proposition 9. *Consider $\mu \in [0, 1]$, $\beta = 0$ and $\eta = 1$, i.e., the objective function of the agent is $\Pi(a, X) = \lambda_W(a, X)$. There exists $\bar{\alpha}^a_B$ and $\bar{\alpha}^b_B$, with $0 < \bar{\alpha}^a_B < \bar{\alpha}^b_B < 1$, such that the following holds:*

1. *For any $\alpha_B \leq \bar{\alpha}^a_B$, there exists $\bar{\mu} > 0$ such that in the unique informative equilibrium, $\sigma^{\mu}_{N}(l)^* = 1$ when $\mu > \bar{\mu}$, and $\sigma^{\mu}_{N}(l)^* < 1$ when $\mu < \bar{\mu}$. In the latter case, $\frac{d\sigma^{\mu}_{N}(l)^*}{d\mu} > 0$. Additionally, if $\alpha_B \to 0$, then $\bar{\mu} \to 0$.*

2. *If $\alpha_B \in \left(\bar{\alpha}^a_B, \bar{\alpha}^b_B\right)$, then in the unique informative equilibrium, $\sigma^{\mu}_{N}(l)^* < 1$ for all μ. In this case, $\frac{d\sigma^{\mu}_{N}(l)^*}{d\mu} > 0$.*

3. *If $\alpha_B \geq \bar{\alpha}^b_B$, then in equilibrium, $\sigma^{\mu}_{N}(l)^* = 0$, for all μ.*

Proof

The results of Proposition 1, Corollary 1 and Proposition 2 are derived for any increasing function $f(\lambda_W(a, W), \lambda_{\bar{B}}(a, X))$. Hence, they also apply to this case. Consequently, first note that in any informative equilibria the wise type always follows the honest strategy.[25] Hence, in the proof we consider that the wise type judge uses the honest strategy and focus on the behavior of the normal type judge. Second, by Proposition 2, we can consider $\sigma^{\mu}_{N}(h)^* = 0$ for all $\mu \in [0, 1]$. For the case $\mu = 0$, we use Definition 2.

The proof of Proposition 9 requires three results, which are proven in three lemmas, Lemmas 14, 15 and 16. Lemma 14 characterizes the unique equilibrium when $\mu = 0$. It defines threshold $\hat{\alpha}_B = \frac{1-\alpha_W}{2}$ and shows that if $\alpha_B < \hat{\alpha}_B$, then $\sigma^{0}_{N}(l)^* = 1 - \frac{\alpha_B}{\alpha_N}$, and if $\alpha_B > \hat{\alpha}_B$, then $\sigma^{0}_{N}(l)^* = 0$. Lemma 15 characterizes the unique equilibrium when $\mu = 1$. It defines thresholds $\bar{\alpha}^a_B = \frac{(2\gamma-1)(\gamma(1-\alpha_W)+\alpha_W)}{1-(1-\gamma)2\gamma}$ and $\bar{\alpha}^b_B = \gamma - (1-\gamma)\alpha_W$, with $\bar{\alpha}^a_B < \bar{\alpha}^b_B$, and shows that if $\alpha_B \leq \bar{\alpha}^a_B$, then $\sigma^{1}_{N}(l)^* = 1$; if $\bar{\alpha}^a_B < \alpha_B < \bar{\alpha}^b_B$, then $0 < \sigma^{1}_{N}(l)^* < 1$; and if $\alpha_B \geq \bar{\alpha}^b_B$, then $\sigma^{1}_{N}(l)^* = 0$. Finally, Lemma 16 shows that if $\sigma^{\mu}_{N}(l)^* \in (0, 1)$, then $\sigma^{\mu}_{N}(l)^*$ is increasing in μ.

[25] Therefore, in points *1.* and *2.* of Proposition 9 we can consider that the wise type follows the honest strategy and focus on the normal type judge behavior. In point *3.* the normal type judge plays $(\sigma^{\mu}_{N}(l)^*, \sigma^{\mu}_{N}(h)^*) = (0, 0)$. In that case, if $\alpha_B < \widetilde{\alpha}$, by Proposition 1, the honest strategy of the wise type judge is an equilibrium strategy. When $\alpha_B > \widetilde{\alpha}$, then the unique equilibrium strategy of the wise type judge is $(\sigma^{\mu}_{W}(l)^*, \sigma^{\mu}_{W}(h)^*) = (0, 0)$. Nevertheless, note that in any case the normal type judge always plays $(\sigma^{\mu}_{N}(l)^*, \sigma^{\mu}_{N}(h)^*) = (0, 0)$.

We next study the relationship between thresholds $\hat{\alpha}_B$, $\bar{\alpha}_B^a$ and $\bar{\alpha}_B^b$. Note that $\hat{\alpha}_B < \bar{\alpha}_B^b \iff \frac{1}{2} - \frac{1}{2}\alpha_W < \gamma - \alpha_W(1-\gamma)$, which always holds as $\gamma > \frac{1}{2}$. In addition, $\hat{\alpha}_B < \bar{\alpha}_B^a \iff \frac{1}{2} - \frac{1}{2}\alpha_W < \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2} \iff \frac{\alpha_W-2\gamma^2-4\gamma\alpha_W+2\gamma^2\alpha_W+1}{4\gamma^2-4\gamma+2} < 0 \iff \gamma > 1 - \frac{1}{1-\alpha w} + \frac{1}{\sqrt{2-\frac{4\alpha w}{1+\alpha w^2}}}$. Let $\breve{\gamma} = 1 - \frac{1}{1-\alpha w} + \frac{1}{\sqrt{2-\frac{4\alpha w}{1+\alpha w^2}}}$. Then, if $\gamma < \breve{\gamma}$, $\bar{\alpha}_B^a < \hat{\alpha}_B < \bar{\alpha}_B^b$, and if $\gamma \geq \breve{\gamma}$, $\hat{\alpha}_B \leq \bar{\alpha}_B^a < \bar{\alpha}_B^b$.

Now, there are two cases to consider. First, suppose $\gamma < \breve{\gamma}$, then $\bar{\alpha}_B^a < \hat{\alpha}_B < \bar{\alpha}_B^b$. In the following lemmas, we also prove that $\frac{\partial \Delta_{N,l}^0}{\partial \sigma_N(l)} > 0$, $\frac{\partial \Delta_{N,l}^1}{\partial \sigma_N(l)} > 0$, and $\frac{\partial \Delta_{N,l}^\mu}{\partial \sigma_N(l)} > 0$. In this case, we have the following:

1) If $\alpha_B < \bar{\alpha}_B^a$, then $0 < \sigma_N^0(l)^* < 1$ and $\sigma_N^1(l)^* = 1$. Thus, $\sigma_N^\mu(l)^* \in (0,1]$. Since $\frac{d\sigma_N^\mu(l)^*}{d\mu} > 0$ for any $\alpha_B \in (0, \bar{\alpha}_B^a)$, there always exists $\bar{\mu} > 0$ such that $\sigma_N^\mu(l)^* < 1$ if $\mu < \bar{\mu}$, and $\sigma_N^\mu(l)^* = 1$ if $\mu > \bar{\mu}$. Additionally, note that by Lemma 14, $\sigma_N^0(l)^* = 1 - \frac{\alpha_B}{\alpha_N}$; hence, if $\alpha_B \to 0$, then $\sigma_N^0(l)^* \to 1$, which implies that $\bar{\mu} \to 0$.

2) If $\alpha_B \in (\bar{\alpha}_B^a, \hat{\alpha}_B)$, then $0 < \sigma_N^0(l)^* < 1$ and $\sigma_N^1(l)^* < 1$. Thus, $\sigma_N^\mu(l)^* \in (0,1)\ \forall \mu$, with $\frac{d\sigma_N^\mu(l)^*}{d\mu} > 0$.

3) If $\alpha_B \in \left(\hat{\alpha}_B, \bar{\alpha}_B^b\right)$, then $\sigma_N^0(l)^* = 0$ and $\sigma_N^1(l)^* < 1$. Thus, $\sigma_N^\mu(l)^* \in [0,1)\ \forall \mu$, with $\frac{d\sigma_N^\mu(l)^*}{d\mu} > 0$.

4) If $\alpha_B \geq \bar{\alpha}_B^b$, then $\sigma_N^0(l)^* = 0$ and $\sigma_N^1(l)^* = 0$. Thus, $\sigma_N^\mu(l)^* = 0\ \forall \mu$.

The analysis for $\gamma > \breve{\gamma}$, then $\hat{\alpha}_B < \bar{\alpha}_B^a < \bar{\alpha}_B^b$, is analogous, and thus, it is omitted.

Next, we prove Lemmas 14, 15 and 16.

Lemma 14. *Suppose $\mu = 0$. Let $\hat{\alpha}_B = \frac{1-\alpha_W}{2}$.*

(i) If $\alpha_B < \hat{\alpha}_B$, there is a unique equilibrium. In the equilibrium, $\sigma_N^0(l)^* = 1 - \frac{\alpha_B}{\alpha_N}$.

(ii) If $\alpha_B \geq \hat{\alpha}_B$, then $\sigma_N^0(l)^* = 0$ in the unique equilibrium strategy.

Proof

In this case, expressions (7) and (8) simplify to the following:

$$\Delta_{N,l}^0 = \Delta_{N,h}^0 = \lambda_W(\hat{h}, 0) - \lambda_W(\hat{l}, 0) = \frac{\alpha_W}{\alpha_W + (2 - \sigma_N(l))\alpha_N} - \frac{\alpha_W}{2\alpha_B + \alpha_W + \sigma_N(l)\alpha_N}. \tag{16}$$

After some algebra, we obtain the following:

$\Delta_{N,l}^0 = \Delta_{N,h}^0 \geq 0 \iff \frac{\alpha_B}{\alpha_N} \geq 1 - \sigma_N(l)$.

There are two cases:

(i) If $\alpha_B < \alpha_N$, then $\frac{\alpha_B}{\alpha_N} < 1$, and consequently, $\Delta_{N,l}^0 = \Delta_{N,h}^0 = 0$ if $\frac{\alpha_B}{\alpha_N} = 1 - \sigma_N(l)$. Hence, $\sigma_N^0(l)^* = 1 - \frac{\alpha_B}{\alpha_N}$.

(ii) If $\alpha_B \geq \alpha_N$, then $\frac{\alpha_B}{\alpha_N} > 1$, and consequently, $\Delta_{N,l}^0 = \Delta_{N,h}^0 > 0$ for any $\sigma_N(l)$. Hence, the optimal strategy of the normal type judge is $(\sigma_N^0(l)^*, \sigma_N^0(h)^*) = (0,0)$. Now, by Lemma 3, we know that in this case, the wise type judge could find it optimal to use a strategy that does not always follow the judge's signal. For this case, Proposition 8 guarantees that the normal type judge has a unique equilibrium strategy, which is to always take action $a = \hat{h}$.

To complete the proof, finally note that $\alpha_B \geq \alpha_N \iff \alpha_B \geq \hat{\alpha}_B = \frac{1-\alpha_W}{2}$. ♦

Lemma 15. *Suppose $\mu = 1$. Let $\bar{\alpha}_B^a = \frac{(2\gamma-1)(\gamma(1-\alpha_W)+\alpha_W)}{1-(1-\gamma)2\gamma}$ and $\bar{\alpha}_B^b = \gamma - (1-\gamma)\alpha_W$, with $\bar{\alpha}_B^a < \bar{\alpha}_B^b$.*

1. *If* $\alpha_B \leq \bar{\alpha}_B^a$*, there is a unique equilibrium. In the equilibrium,* $\sigma_N^1(l)^* = 1$.

2. *If* $\alpha_B \in (\bar{\alpha}_B^a, \bar{\alpha}_B^b)$*, there is a unique equilibrium. In the equilibrium,* $\sigma_N^1(l)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)} \in (0,1)$.

3. *If* $\alpha_B \geq \bar{\alpha}_B^b$*, then* $\sigma_N^1(l)^* = 0$ *in the unique equilibrium strategy.*

Proof

First, note that

$$\Delta_{N,l}^1 = (1-\gamma)\lambda_W(\hat{h}, H) - \gamma\lambda_W(\hat{l}, L) = \frac{(1-\gamma)\alpha_W}{\alpha_W+(\gamma+(1-\gamma)(1-\sigma_N(l)))\alpha_N} - \frac{\gamma\alpha_W}{\alpha_B+\alpha_W+\gamma\sigma_N(l)\alpha_N}, \tag{17}$$

with $\frac{\partial\Delta_{N,l}^1}{\partial\sigma_N(l)} > 0$. Consequently, there is either one root for $\sigma_N(l)$ or none. In other words, the equilibrium strategy is unique.

Now, note the following:

$\Delta_{N,l}^1\Big|_{\sigma_N(l)=1} \leq 0 \iff \alpha_B \leq \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2}$. Let $\bar{\alpha}_B^a = \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2}$. Consequently, $\sigma_N^1(l)^* = 1$ if $\alpha_B \leq \bar{\alpha}_B^a$.

$\Delta_{N,l}^1\Big|_{\sigma_N(l)=0} \geq 0 \iff \alpha_B \geq \gamma - \alpha_W(1-\gamma)$. Let $\bar{\alpha}_B^b = \gamma - \alpha_W(1-\gamma)$. Consequently, $\sigma_N^1(l)^* = 0$ if $\alpha_B > \bar{\alpha}_B^b$.

When $\bar{\alpha}_B^a < \alpha_B < \bar{\alpha}_B^b$, note that $\Delta_{N,l}^1\Big|_{\sigma_N(l)=0} < 0$ and $\Delta_{N,l}^1\Big|_{\sigma_N(l)=1} > 0$, which implies that there is one root. We obtain the following:

$\Delta_{N,l}^1 = 0 \iff \sigma_N^1(l)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)}$.

Finally, note that in the case $\alpha_B > \bar{\alpha}_B^b$, $\sigma_N^1(l)^* = 0$, which implies that the normal type judge uses a non-informative strategy. For this case, Lemma 3 says that the wise type judge could find it optimal to use a strategy that does not always follow the judge's signal. Finally, by Proposition 8, we know that in this case, the normal type judge has a unique equilibrium strategy, which is to always take action $a = \hat{h}$. ♦

Lemma 16. *If* $\sigma_N^\mu(l)^* \in (0,1)$*, then* $\sigma_N^\mu(l)^*$ *is increasing in* μ.

Proof

First, note that $\Delta_{N,l}^\mu = (1-\mu)\Delta_{N,l}^0 + \mu\Delta_{N,l}^1$, where from (16) and (17), we have $\frac{\partial\Delta_{N,l}^0}{\partial\sigma_N(l)} > 0$ and $\frac{\partial\Delta_{N,l}^1}{\partial\sigma_N(l)} > 0$. Thus, $\frac{\partial\Delta_{N,l}^\mu}{\partial\sigma_N(l)} > 0$, which means that the equilibrium strategy of the normal type judge is unique.

Now, let us denote by $\bar{\sigma}_N^0(l)^*$ and $\bar{\sigma}_N^1(l)^*$ the interior solutions to equations $\Delta_{N,l}^0 = 0$ and $\Delta_{N,l}^1 = 0$, respectively. From Lemmas 14 and 15 above, we know that these equilibrium strategies are $\bar{\sigma}_N^0(l)^* = 1 - \frac{\alpha_B}{\alpha_N}$ and $\bar{\sigma}_N^1(l)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)}$.

Next, note that $\bar{\sigma}_N^0(l)^* < \bar{\sigma}_N^1(l)^* \iff 1 - \frac{\alpha_B}{\alpha_N} - \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)} < 0 \iff \frac{(2\gamma-1)(\alpha_W(1-\gamma)\alpha_B+\gamma\alpha_N)}{2\gamma\alpha_N(1-\gamma)} > 0$, which is always the case.

Finally, note that $\Delta_{N,l}^0 < 0$ if $\sigma_N(l) < \bar{\sigma}_N^0(l)^*$, and $\Delta_{N,l}^0 > 0$ if $\sigma_N(l) > \bar{\sigma}_N^0(l)^*$. Similarly, $\Delta_{N,l}^1 < 0$ if $\sigma_N(l) < \bar{\sigma}_N^1(l)^*$, and $\Delta_{N,l}^1 > 0$ if $\sigma_N(l) > \bar{\sigma}_N^1(l)^*$. Then, if $0 < \sigma_N^\mu(l)^* < 1$, necessarily, $\sigma_N^\mu(l)^* \in$

$\left[\bar{\sigma}_N^0(l)^*, \bar{\sigma}_N^1(l)^*\right]$. Finally, since $\Delta_{N,l}^{\mu} = (1-\mu)\,\Delta_{N,l}^0 + \mu\Delta_{N,l}^1$, with $\Delta_{N,l}^0 > 0$ and $\Delta_{N,l}^1 < 0$ in the interval $\left[\bar{\sigma}_N^0(l)^*, \bar{\sigma}_N^1(l)^*\right]$, then $\sigma_N^{\mu}(l)^*$ has to be increasing in μ. ♦

This completes the proof of Proposition 9. **QED**.

References

Andina-Díaz, Ascensión and José A. García-Martínez (2020), 'Reputation and news suppression in the media industry', *Games and Economic Behavior* **123**, 240–271.

Austen-Smith, David and Roland G. Fryer (2005), 'An economic analysis of "Acting White"', *Quarterly Journal of Economics* **120**(2), 551–583.

Avery, Christopher N. and Judith A. Chevalier (1999), 'Herding over the career', *Economics Letters* **63**, 327–333.

Bagwell, Kyle (2007), 'Signalling and entry deterrence: A multi-dimensional analysis', *RAND Journal of Economics* **38**(3), 670–697.

Bar-Isaac, Heski and Joyee Deb (2014), '(Good and bad) Reputation of a servant of two masters', *American Economic Journal: Microeconomics* **6**(4), 293–325.

Bénabou, Roland and Jean Tirole (2006), 'Incentives and prosocial behavior', *American Economic Review* **96**(5), 1652–1678.

Besley, Timothy (2006), *Principled Agents? The Political Economy of Good Goverment*, Oxford University Press.

Bourjade, Sylvain and Bruno Jullien (2011), 'The roles of reputation and transparency on the behavior of biased experts', *RAND Journal of Economics* **42**(3), 575–594.

Canes-Wrone, Brandice, Michael C. Herron and Kenneth W. Shotts (2001), 'Leardhip and pandering: A theory of executive policymaking', *American Journal of Political Science* **45**(3), 532–550.

Cohen, Alma, Alon Klement and Zvika Neeman (2015), 'Judicial decision making: A dynamic reputation approach', *Journal of Legal Studies* **44**(1), 133–159.

Dewatripont, Mathias, Ian Jewitt and Jean Tirole (1999), 'The economics of career concerns, part i: Comparing information structures', *Review of Economic Studies* **66**(1), 183–198.

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

Esteban, Joan and Debraj Ray (2006), 'Inequality, lobbying, and resource allocation', *American Economic Review* **96**(1), 257–279.

Feller, Miró and Ulrich Schäfer (2020), Deceiving two masters: The effects of capital and labor market incentives on reporting bias. Working paper.

Ferrer, Rosa (2016), The effect of lawyers' career concerns on litigation. Working paper.

Fox, Justin and Kenneth W. Shotts (2009), 'Delegates or trustees? A theory of political accountability', *Journal of Politics* **71**(4), 1125–1137.

Fox, Justin and Richard Van Weelden (2012), 'Costly transparency', *Journal of Public Economics* **96**(1), 142–150.

Frankel, Alex and Navin Kartik (2019), 'Muddled information', *Journal of Political Economy* **127**(4), 1739–1776.

Gentzkow, Matthew and Jesse M. Shapiro (2006), 'Media bias and reputation', *Journal of Political Economy* **114**(2), 280–316.

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

Holmström, Bengt (1999), 'Managerial incentive problems: A dynamic perspective', *Review of Economic Studies* **66**(1), 169–182.

Hörner, Johannes (2002), 'Reputation and competition', *American Economic Review* **92**(3), 644–663.

Iossa, Elisabetta and Bruno Jullien (2012), 'The market for lawyers and quality layers in legal services', *RAND Journal of Economics* **43**(4), 677–704.

Lazarus, Jeffrey V., Scott Ratzan and Adam Palayew et al. (2020), 'Covid-score: A global survey to assess public perceptions of government responses to covid-19 (covid-score-10)', *PloS one* **15**(10), e0240011.

Leaver, Clare (2009), 'Bureaucratic minimal squawk behavior: Theory and evidence from regulatory agencies', *American Economic Review* **99**(3), 572–607.

Levy, Gilat (2005), 'Careerist judges and the appeals process', *RAND Journal of Economics* **36**(2), 275–297.

Levy, Gilat (2007), 'Decision making in committees: Transparency, reputation, and voting rules', *American Economic Review* **97**(1), 150–168.

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

Lim, Claire S. H., James M. Snyder and David Strömberg (2015), 'The judge, the politician, and the press: Newspaper coverage and criminal sentencing across electoral systems', *American Economic Journal: Applied Economics* **7**(4), 103–135.

Liu, Yaozhou Franklin and Amal Sanyal (2012), 'When second opinions hurt: A model of expert advice under career concerns', *Journal of Economic Behavior and Organization* **84**(1), 1–16.

Maskin, Eric and Jean Tirole (2004), 'The politician and the judge: Accouintability in government', *The American Economic Review* **9**(4), 1034–1054.

Miceli, Thomas J. and Metin M. Cosgel (1994), 'Reputation and judicial decision making', *Journal of Economic Behavior and Organization* **23**(1), 31–5.

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

Ottaviani, Marco and Peter N. Sørensen (2001), 'Information aggregation in debate: Who should speak first?', *Journal of Public Economics* **81**(3), 393–421.

Ottaviani, Marco and Peter N. Sørensen (2006), 'The strategy of professional forecasting', *Journal of Financial Economics* **81**(2), 441–466.

Prat, Andrea (2005), 'The wrong kind of transparency', *American Economic Review* **95**(3), 862–877.

Sibert, Anne (2003), 'Monetary policy committees: Individual and collective reputations', *Review of Economic Studies* **70**(3), 649–665.