The Reputational Cost of Truthful Informational Transmission[*]

Stephen Morris
University of Pennsylvania

June 1997

Abstract

An uninformed decision maker repeatedly receives advice from a good informed advisor; that is, the advisor's information is valuable and she only cares about the utility of the decision maker. But the decision maker attaches positive probability to the advisor being stupid or having other objectives. The good advisor has a current incentive to truthfully reveal her information; but she may have a reputational incentive to lie (in order to separate herself from possible "bad" advisors). If the possible bad advisor is informed but biased, and if the current decision problem is relatively unimportant compared to future ones (for the decision maker and thus for the good advisor) this reputational cost of telling the truth ensures that no information is transmitted by either type of advisor in equilibrium. This paper also explores when truth-telling is equilibrium behavior for different kinds of possible bad advisors.

[*] This version is prepared for the 1997 Warwick Summer Research Workshop. It is HIGHLY PRELIMINARY (and should not be referred to w/o permission). In particular, I have not converged on the appropriate notation for making comparisons across a number of models: I hope that ambiguity in notation will eventually be cleared up. I have benefited from valuable conversations on this material with Stephen Coate, George Mailath and Andrew Postlewaite. Financial support from the Alfred P. Sloan foundation is gratefully acknowledged.

1. Introduction

An uninformed decision maker repeatedly receives advice from an advisor. The advisor is of one of two types. A "good" advisor is well (but not perfectly) informed and cares only about the discounted future utility of the decision maker. A "bad" advisor may or may not be informed and has different preferences from the decision maker. Not surprisingly, the bad advisor will lie to the decision maker. But the good advisor may have incentives to lie also. Although telling the truth maximizes the current utility of the good advisor (since it leads to the best decision for the decision maker), the good advisor will also be concerned about her reputation (the probability the decision maker attaches to her being good). Telling the truth need not maximize her reputation, as the following example illustrates.

Consider a social scientist who (repeatedly) advises an uninformed policy maker on alternative social policies. Suppose that the social scientist is not racist, but, on one issue, her policy recommendation coincides with that of many racists. Telling the truth will improve the policy maker's current choice of action (since the policy maker attaches positive probability to the social scientist not being racist). But it would (in equilibrium) increase the policy maker's probability that the social scientist is racist. This is true *independent of whether the policy turns out to be a success*. This means that less weight would be placed on her future advice. Thus she might (in the interests of the policy maker) lie about her current policy recommendation.

The same logic will apply in many contexts. Suppose an investment advisor may either be exclusively concerned with her clients' welfare (good) or may be trying to off-load excess stocks (bad). The good advisor will forego recommending profitable trades in order to enhance her reputation. An academic referee may either be exclusively concerned with the intellectual merits of the work to be refereed (good) or may seek only to enhance her research agenda (bad). The good referee will lie about research that enhances her own agenda.

In all the above examples, the "bad" advisor has some bias in her preferences. Consider instead cases where the bad advisor tells the truth but is stupid (i.e., has noisy signals). An investment advisor is either competent (good) or stupid (bad). The good investment advisor hears good things (i.e., observes a positive signal) about a crazy sounding scheme (i.e., an investment with a low ex ante probability of success). Should she pass on this (useful) information to her client?

On average, people with noisier signals are more likely to hear good things about crazy schemes. The investment advisor may forego passing on the good advice in order to have more conservative future advice taken seriously. Again, this reputational incentive effect works even if the client learns the outcome of the crazy scheme. Finally, a scientist is either competent (good) or stupid (bad). Suppose the competent scientist observes something really surprising in her laboratory. If she reports it, her future observations may not be taken seriously. It is better to establish a reputation with more conservative findings before revealing any radical observations.

The purpose of this paper is to characterize in which circumstances truth-telling is optimal for a "good" advisor and in which circumstances the reputational concerns cited above ensure that no information is conveyed in equilibrium. We address these questions in a repeated cheap talk game, extending the framework of Sobel (1985) and Benabou and Laroque (1992). In particular, suppose that in each period a binary state of the world, 0 or 1, is realized. The advisor observes a noisy signal of that state and may (costlessly) announce that signal to a decision maker. The decision maker chooses an action from a continuum. His optimal action is a continuous increasing function of the probability he attaches (in equilibrium) to state 1. The state is realized (and publicly observed) after the decision maker's action is chosen. In each period, there is a new, independent, state and a new decision to be made. Another independent random variable in each period determines the weight (importance) of the current decision problem; the decision maker maximizes his discounted expected weighted utility from decisions. The good advisor maximizes the same expression. The paper considers alternative specifications of the bad advisor.

I analyze in detail the case where bad advisor and good advisor are equally well-informed but the bad advisor always prefers higher actions, independent of the state of the world. (Recall that the good advisor, like the decision maker, would like a high action in state 1 and a low action in state 0). It is useful to first analyze what might happen in any individual decision problem if we exogenously fix an increasing, continuous, value function in the probability of being good for each advisor. It is possible to provide a general characterization of equilibria in this reduced form game (propositions 1 and 2). There always exist "babbling" equilibria, where no information about the state or the type of advisor is revealed. In addition, there may exist informative equilibria. In such equilibria, the bad advisor always announces signal 1 more often than the good advisor. So independent of the state realized, announcing 1 decreases her reputation and announcing 0 increases her reputation. Thus each advisor always has a reputational incentive to announce 0; the bad advisor always has a current incentive to announce 1 (since she wants a high action realized); while the good advisor has a current incentive to announce 1 only if she observes signal 1. One implication is that the good advisor always announces 0 when she observes signal 0. Now consider what happens to the set of equilibria as the importance of the current decision problem (to the decision maker and thus the good advisor) is varied, holding fixed the reputation value function of the good advisor (proposition 3). If the importance of the current decision problem is sufficiently high, there is an equilibrium where the good advisor always tells the truth. But if the importance of the current decision problem is sufficiently low, no information is conveyed in any equilibrium (i.e., only babbling equilibria exist).

Recall that the above results followed from the *assumption* that players have continuous increasing valuations of reputation. I show that such value functions can arise endogenously from a purely instrumental concern for reputation. In particular, there exists a Markov equilibrium of the infinitely repeated advise game with continuous increasing value functions (proposition 4). All the properties described in the previous paragraph are inherited by this equilibrium.

The above results concerned a particular type of bad advisor (informed but with a specific bias in preferences). We would like to characterize (more generally) which *preferences* and *competence* of the bad advisor will lead the good advisor to lie in equilibrium, and which are consistent with truth-telling by the good advisor. However, it is hard to solve for all possible preferences and competences of the bad advisors. Instead, there is a characterization (lemma 2) of which *strategies* for the bad advisor are consistent with equilibrium truth-telling by the good advisor (however small the importance of the current decision problem). This characterization can be used to solve for various types of bad advisor preferences and competence:

1. *Smart Zealot*: a bad advisor who is informed but has biased preferences. This is the case we already described. The good advisor will always have a reputational incentive to lie given such a bad advisor.

2. *Honest Fool*: a bad advisor who is uninformed but always tells the truth. The good advisor will have a reputational incentive to lie against such an advisor only if his signal is ex ante very surprising.

3. *Foolish Zealot*: a bad advisor who is uninformed *and* has a systematic bias in preferences. The good advisor may have a reputational incentive to always tell the truth if value functions are sufficiently aligned in equilibrium.

4. *Smart Enemy*: a bad advisor who is informed and has *opposite* preferences (with no particular bias) from the decision maker. The good advisor will have a reputational incentive to tell the truth.

This paper builds on a number of earlier literatures and it will be useful to put it in context. The static cheap talk literature (Crawford and Sobel 1982) showed that an advisor (sender) may tell the truth if she is good (her preferences are close to those of the receiver) but must babble is she is bad (her preferences are different from the receiver). Sobel (1985) and Benabou and Laroque (1992) considered *repeated* cheap talk games with uncertainty about the type of the advisor. They *assumed* that the good advisor tells the truth. They showed that a bad advisor will not always babble. Rather, the bad advisor will sometimes tell the truth (investing in reputation) and sometimes lie (exploiting that reputation). This paper endogenizes the behavior of the good advisor in Benabou and Laroque (1992). Just as the bad advisor's current interests are sometimes reversed by reputational concerns, so too for the good advisor. Just as the bad advisor has an incentive to tell the truth (despite a current incentive to lie) in order to enhance her reputation, so the good advisor may have an incentive to lie (despite a current incentive to tell the truth) in order to enhance her reputation. The purpose of this paper is to characterize when this occurs.

Loury (1994) has explored the more general question of "self-censorship in public discourse" for strategic reasons; the classic example is that of so-called political correctness but Loury argues the phenomenon is quite pervasive. Loury's explanation is that participants in public debate are influenced by concerns about how their statements will change listeners' views of their type. In particular, speakers want to be perceived to be respectful of social norms and therefore make statements that respect social norms for expression (he suggests that a formal model along the lines of Bernheim (1994) would be relevant here). There are two problems with this approach. First, there is no explanation of the origin of social norms concerning expression. Second, the cost of the self-censorship is not examined. The model presented here can be seen as a formalization of some of Loury's ideas, which in addition addresses these two problems. In this paper, there is real information content in expression, and therefore social value is lost when self-censorship occurs. Speakers' reputational concerns are not about some reduced form feature such as "respect for social norms," but rather about their perceived preferences and ability. Furthermore, they care about the latter for purely instrumental reasons. This foundation of the model in standard economic and informational assumptions allows the possibility of *explaining* the origin of socially acceptable forms of expression and conducting welfare analysis of political correctness.

Reputational concerns for competence have appeared in a number of papers; Prendergast and Stole (1996) is especially relevant. They consider a manager who takes an investment decision and is concerned both about the output of the investment decision and her reputation for competence. Absent reputational concerns, competent managers would take more extreme investment decisions, because they have more accurate information. Given the reputational concerns, all managers will initially take excessively extreme decisions in order to signal competence. On the other hand, if there is correlation across optimal decisions through time, managers who have been in place for some time will excessively reduce variability, in order to signal their faith in their earlier decisions, and thus their competence. A number of features lead to the different reputation for competence results in this paper. First, there is no correlation in decision problems through time. Second, because advice is not costly, there is no ability to separate by choosing some sufficiently costly action.[1]

2. Separating from a Smart Zealot: A Model of Political Correctness

2.1. Exogenous Reputation

A decision maker's optimal decision depends on the state of the world $\omega \in \{0,1\}$. Each state occurs with probability $\frac{1}{2}$. The decision maker has access to an advisor who may be partially informed about the state of the world. The advisor observes a signal $s \in \{0,1\}$ that is correlated with the true state of the world. In particular, the probability that the signal equals the true state is $\gamma \in \left(\frac{1}{2},1\right)$.

[1] Following Milgrom and Roberts (1986), a number of authors examine incentives of interested parties to reveal information (see also Dewatripont and Tirole (1995) and Shin (1996)). But these are not cheap talk papers as interested parties can prove that their information is accurate (if they choose to reveal it).

With probability λ, the advisor is "good" (type G), and with probability $1-\lambda$, the advisor is "bad" (type B). The type I advisor's strategy is a function $\sigma_I : \{0,1\} \to [0,1]$, where $\sigma_I(s)$ is the probability of announcing message 1 when her signal is s. Given the advisor's message, the decision maker must choose an action $a \in \Re$. After the action is chosen, the state of the world ω is publicly observed.

The decision maker's utility depends on his optimal action and the state of the world: his utility from action a in state ω is $x.u_{DM}(a,\omega)$, where $x > 0$ and $u_{DM}(a,\omega)$ is differentiable and strictly concave in a and attains a maximum for each ω. Write $a^*(m) = \underset{a\in\Re}{\arg\max}\ u_{DM}(a,\omega)$ and assume $a^*(1) > a^*(0)$. The decision maker's *strategy* is a function $\chi : \{0,1\} \to \Re$; $\chi(m)$ is his action if m is the message from his advisor.

The advisor's utility depends on the decision maker's beliefs after observing the state of the world. In particular, write $\Lambda[\sigma_G,\sigma_B](m,\omega)$ for the posterior probability that the advisor is good if she sends message m and state ω is realized.[2] Then

$$\begin{aligned}\Lambda[\sigma_G,\sigma_B](m,\omega) &= \frac{\lambda\phi_G(m|\omega)}{\lambda.\phi_G(m|\omega)+(1-\lambda).\phi_B(m|\omega)} \\ &= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\phi_B(m|\omega)}{\phi_G(m|\omega)}\right)},\end{aligned}$$

where $\phi_I(m|\omega)$is the probability that advisor I sends message m given state ω, i.e.,$\phi_I(1|\omega) = \gamma\sigma_I(\omega) + (1-\gamma)\sigma_I(1-\omega)$ and $\phi_I(0|\omega) = 1-\phi_I(1|\omega)$. (This is well defined only if the denominator is non-zero. I adopt the convention that $\Lambda[\sigma_G,\sigma_B](m,\omega) = \lambda$ if $\sigma_G(m|1) = \sigma_G(m|0) = \sigma_B(m|1) = \sigma_B(m|0) = 0$. This restriction does not effect the analysis of equilibria).

The good advisor cares about the current utility of the DM and her ex post reputation. Her payoff is

$$xu_{DM}(a,\omega) + v_G[\Lambda[\sigma_G,\sigma_B](m,\omega)]$$

[2] A colleague advises me that distinguishing players by sex (the decision maker is male, the advisor is female) will clarify my argument and not distract the reader. My colleague may sincerely believe this. On the other hand, he knows that I would discount his future stylistic advice if I thought it was motivated by sexism. It is possible that he does not believe his own politically correct advice, but that he is sufficiently anxious that I take his future stylistic advice that he is prepared to lie. In this scenario, he is lying precisely because he is concerned about the clarity of my arguments. See proposition 3.

where $x > 0$ and $v_G : [0,1] \to \Re$ is a strictly increasing continuous function. The bad advisor always wants action 1 chosen but also cares about her reputation. In particular, her payoff is

$$yu_B(a) + v_B[\Lambda[\sigma_G, \sigma_B](m, \omega)]$$

where $y > 0$, u_B is a strictly increasing in a on the interval $(a^*(0), a^*(1))$ and $v_B : [0,1] \to \Re$ is a strictly increasing continuous function.

If the advisor follows strategy (σ_G, σ_B), write $\Gamma[\sigma_G, \sigma_B](m)$ for the DM's posterior belief that the actual state is 1 if message 1 is announced. By Bayes' rule,

$$\Gamma[\sigma_G, \sigma_B](m) = \frac{\lambda \phi_G(m|1) + (1-\lambda)\phi_B(m|1)}{\lambda \phi_G(m|1) + (1-\lambda)\phi_B(m|1) + \lambda \phi_G(m|0) + (1-\lambda)\phi_B(m|0)}.$$

(Again, this is well defined only if the denominator is non-zero. I adopt the convention that $\Gamma[\sigma_G, \sigma_B](m) = \frac{1}{2}$ if $\sigma_G(m|0) = \sigma_B(m|0) = \sigma_G(m|1) = \sigma_B(m|1) = 0$.)

Now $(\sigma_G, \sigma_B, \chi)$ is an equilibrium if the advisor's action given his signal maximizes his utility given the decision maker's strategy χ and the type inference function $\Lambda[\sigma_G, \sigma_B]$; and the decision maker's action is optimal given the state inference function $\Gamma[\sigma_G, \sigma_B]$. A more formal statement of the equilibrium concept is given in appendix A.

The decision maker's best response is straightforward to characterize.

Lemma 1. *In any equilibrium $(\sigma_G, \sigma_B, \chi)$, $\chi = \widetilde{\chi}[\sigma_G, \sigma_B]$ where*

$$\widetilde{\chi}[\sigma_G, \sigma_B](m) = \widetilde{a}(\Gamma[\sigma_G, \sigma_B](m))$$

and $\widetilde{a} : [0,1] \to [a^*(0), a^*(1)]$ is the unique continuous, strictly increasing function solving

$$qu'_{DM}(\widetilde{a}(q), 1) + (1-q)u'_{DM}(\widetilde{a}(q), 0) = 0.$$

For example, if $u_{DM}(a, \omega) = -(a-\omega)^2$, $a^*(0) = 0$, $a^*(1) = 1$, and $\widetilde{a}(q) = q$.

Definition 1. *$(\sigma_G, \sigma_B, \chi)$ is a babbling strategy profile if $\chi(0) = \chi(1) = \frac{1}{2}$ and, for some $c \in [0,1]$, $\sigma_G(0) = \sigma_B(0) = \sigma_G(1) = \sigma_B(1) = c$.*

Any babbling strategy is uninformative in two senses: the decision maker receives information neither about the true state of the world, nor about the type of the advisor.

Proposition 1. *Every babbling strategy profile is an equilibrium.*

In analyzing non-babbling equilibria, we will focus on equilibria $(\sigma_G, \sigma_B, \chi)$ where the decision maker takes at least as large an action after state 1 as after state 0, i.e., $\chi(1) \geq \chi(0)$. This assumption is without loss of generality.

Proposition 2. *For any (λ, x, y), any non-babbling equilibrium $(\sigma_G, \sigma_B, \chi)$ has [1] the good advisor always telling the truth when she observes signal* 0 *($\sigma_G(0) = 0$); [2] strictly informative messages ($\chi(1) > \chi(0)$); and [3] a strict reputational incentive for the advisor to announce* 0 *($\Lambda(0,1) \geq \Lambda(0,0) > \lambda > \Lambda(1,1) \geq \Lambda(1,0)$). More specifically, there exist three types of non-babbling equilibria:*

- *Truthful* (the good advisor always tells the truth): $\sigma_G(0) = 0$, $\sigma_G(1) = 1$, $\sigma_B(0) > 0$ and $\sigma_B(1) = 1$. Such equilibria have $\Lambda(0,1) = \Lambda(0,0) > \lambda > \Lambda(1,1) > \Lambda(1,0)$.

- *Politically Correct* (neither advisor says 1 when $s = 0$): $\sigma_G(0) = 0$, $\sigma_G(1) \in (0,1)$, $\sigma_B(0) = 0$ and $\sigma_B(1) > \sigma_G(1)$. Such equilibria have $\Lambda(0,1) > \Lambda(0,0) > \lambda > \Lambda(1,1) = \Lambda(1,0)$.

- *Full Support:* $\sigma_G(0) = 0$, $\sigma_G(1) \in (0,1)$, $\sigma_B(0) \in (0,1)$ and $\sigma_B(1) > 1 - (1 - \sigma_G(1))(1 - \sigma_B(0))$. Such equilibria have $\Lambda(0,1) > \Lambda(0,0) > \lambda > \Lambda(1,1) > \Lambda(1,0)$.

In other words, the good advisor always tells the truth if she observes signal 0 (since announcing 0 tends to show that she is not the bad advisor). But she may lie sometimes if she observes signal 1. The bad advisor may announce 1 all the time, or she may announce 0 almost all the time. But she must announce 1 at least as often as the good advisor does. Under any strategies of this form, announcing 0 always guarantees a reputation strictly greater than λ (whatever the realized state) while announcing 1 always gives the advisor a reputation strictly less than λ. The proof is presented in appendix B. It is clear from the proof that for any strategies satisfying the necessary conditions of the proposition, we can choose x, y, v_G and v_B such that those strategies are played in some equilibrium.

Proposition 3. *There exist continuous functions $\widetilde{x}_1, \widetilde{x}_2 : (0,1) \times \Re_{++} \rightarrow \Re_{++}$ such that [1] if $x \leq \widetilde{x}_1(\lambda, y)$, all equilibria of the (λ, x, y) game are babbling; and [2] there exists a truthful equilibrium if and only if $x \geq \widetilde{x}_2(\lambda, y)$; where the $\widetilde{x}_i$ satisfy (i) $\widetilde{x}_i(\lambda, y) \rightarrow 0$ as $\lambda \rightarrow 1$; (ii) $\widetilde{x}_1(\lambda, y) \rightarrow 0$ as $\lambda \rightarrow 0$; $\widetilde{x}_2(\lambda, y) \rightarrow \overline{x}_2^*(y) \in (0, \infty)$ as $\lambda \rightarrow 0$; (iii) $\widetilde{x}_i(\lambda, y) \rightarrow 0$ as $y \rightarrow 0$; and (iv) $\widetilde{x}_i(\lambda, y) \rightarrow \overline{x}_i^{**}(\lambda)$ as $y \rightarrow \infty$.*

The proposition predicts that political correctness will be most prevalent when the current decision problem is of small importance to the current decision maker.[3]

2.2. Endogenizing Reputation

Now consider the infinitely repeated game where there is a new decision problem in each period. Each period's decision problem is parameterized by (x, y), the importance of the problem for the decision maker (and good advisor) and bad advisor respectively. Assume that x and y are drawn from X and Y respectively, which are discrete subsets of $\Re_{++}$; write $\phi \in \Delta(X \times Y)$ for the probability distribution on $X \times Y$. Assume that ϕ has infinite support but that

$$\sum_{(x,y)\in X\times Y} x.\phi(x,y) < \infty \text{ and } \sum_{(x,y)\in X\times Y} y.\phi(x,y) < \infty.$$

The discount rates of the decision maker and the bad advisor are δ_{DM} and δ_B, both elements of $(0, 1)$. A (Markov) advisor strategy is a pair (σ_G, σ_B), each $\sigma_I : \{0, 1\} \times (0, 1) \times X \times Y \to [0, 1]$; $\sigma_I(s; \lambda, x, y)$ is the probability of sending message 1 if the advisor is of type I, observes signals s, has reputation λ and (x, y) are the values of the current decision problem.

An advisor strategy is a function $\chi : \{0, 1\} \times (0, 1) \times X \times Y \to \Re$, where $\chi(m; \lambda, x, y)$ is the decision maker's action if he receives message m, the advisor has reputation λ and (x, y) are the values of the current decision problem.

Definition 2. *A Markov equilibrium is characterized by a strategy profile $(\sigma_G, \sigma_B, \chi)$ and value functions v_G and v_B for the good and bad advisors such that [1] decision maker strategy χ is optimal given (σ_G, σ_B); [2] advisor strategy (σ_G, σ_B) maximizes current plus reputational utility (given by (v_G, v_B)) after every history; and [3] value functions (v_G, v_B) are generated by strategy profile $(\sigma_G, \sigma_B, \chi)$. A Markov equilibrium is a* monotonic Markov *[MM] if the value functions are continuous and strictly increasing.*

Although my analysis focuses on monotonic Markov equilibria, it is easy to demonstrate the existence of other well-behaved Markov equilibria. Consider the following construction. Suppose the good advisor always told the truth. By

[3] The reported prevalence of political correctness in academia might then be explained by proposition 3 and the dictum (due to Bernard Shaw?) that academic disputes are so heated precisely because there is so little at stake...

a variation on an argument of Benabou and Laroque [1992], there is a unique best response (for any given δ_B) for the bad advisor with a continuous strictly increasing value function. If δ_B is sufficiently small, this best response will have the bad advisor always lying for sufficiently high reputations but sometimes telling the truth for lower reputations. Given this strategy, we can choose δ_{DM} sufficiently small such that truth telling is indeed a best response for the good advisor. Now we can construct the value function for the good advisor corresponding to these strategies. For δ_{DM} sufficiently small, the slope of the value function will be determined by what happens next period. The good advisor will prefer the bad advisor to have a low reputation (so she sometimes tells the truth) rather than a high reputation (so she always lies).

Nonetheless, there do always also exist MM equilibria.

Proposition 4. *A monotonic Markov equilibrium always exists.*

The intuition for existence is straightforward. Suppose some pair of valuations (x, y) occurs with very low probability ε. Consider the strategy profile where the advisor always babbles after all histories where (x, y) is not drawn. If (x, y) is drawn, the good advisor tells the truth and the bad advisor always announces 1. If ε is sufficiently small, these strategies will be best responses to each other (as reputational concerns will become insignificant). But we can choose ε sufficiently small by our choice of (x, y).

Monotonic Markov equilibria inherit all the structure of propositions 1, 2 and 3. In particular, for any given λ and y, there exists x^* such that for all $x \leq x^*$,

$$\sigma_G\left(1\left|\lambda, x, y\right.\right) = \sigma_G\left(0\left|\lambda, x, y\right.\right) = \sigma_B\left(1\left|\lambda, x, y\right.\right) = \sigma_B\left(0\left|\lambda, x, y\right.\right).$$

3. Separating from Other Bad Advisors

Now complicate the model by first allowing bad advisors to be less accurate than good advisors: $\frac{1}{2} < \gamma_G$ and $\frac{1}{2} \leq \gamma_B \leq \gamma_G$. Also allow the probability of state 1 to be any $\pi \in (0, 1)$.

3.1. General Truth-Telling Conditions

In order to deal with a wide range of alternative bad advisors, it is useful to focus on a narrower question. Suppose we knew the bad advisor's strategy (wherever it comes from). When is truth-telling a best response for the good advisor whatever the importance of the current decision problem (i.e., in the model of the previous section, no matter how small x)?

Thus we will write β_ω for the probability that the bad advisor announces 1 when the true state is ω (in the language of the previous section, $\beta_\omega = \phi_B\left(1|\omega\right)$). For which values of $(\beta_0, \beta_1, \lambda, \pi)$ does the good advisor have a reputational incentive to tell the truth? Clearly, this is a necessary condition for truth telling to be optimal for all values of x. It will also be sufficient if $\beta_1 \geq \beta_0$.

We write $T\left(\lambda, \pi\right) \subseteq [0,1]^2$ for the set of such truth-telling inducing strategies.

Lemma 2.

$$T\left(\lambda,\pi\right) = \left\{ (\beta_0,\beta_1) : \begin{array}{l} \textit{[1]}\ \beta_0 \geq 1-\gamma_G \\ \textit{[2]}\ \beta_1 \leq \gamma_G \\ \textit{[3]}\ \frac{1-\gamma_G}{\gamma_G} \leq \left(\frac{\pi}{1-\pi}\right)\left(\frac{v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)\beta_1}\right)-v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)(1-\beta_1)}\right)}{v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)(1-\beta_0)}\right)-v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)\beta_0}\right)}\right) \leq \frac{\gamma_G}{1-\gamma_G} \end{array} \right\}$$

The following example can be used to illustrate the shape of $T\left(\lambda,\pi\right)$. Let $v_G\left(\lambda\right) = -\left(\frac{1-\lambda}{\lambda}\right)$. In this case, condition [3] becomes

$$\frac{1-\gamma_G}{\gamma_G} \leq \left(\frac{\pi}{1-\pi}\right)\left(\frac{\gamma_G-\beta_1}{\beta_0-(1-\gamma_G)}\right) \leq \frac{\gamma_G}{1-\gamma_G}.$$

Figure 1 illustrates the shape of this region when $\gamma_G = \frac{3}{4}$ and $\pi = \frac{1}{2}$; figure 2 illustrates the shape of this region when $\gamma_G = \frac{3}{4}$ and $\pi = \frac{9}{10}$. In this particular case, the region is independent of λ. But the qualitative properties of this example generalize. Thus as π tends to 1, $T\left(\lambda,\pi\right)$ becomes very narrow: more generally, the characterization implies that if $(\beta_0^n, \beta_1^n) \in T\left(\lambda, \pi^n\right)$. As $\pi^n \to 0$, $\beta_1^n \to \gamma_G$; as $\pi^n \to 1$, $\beta_0^n \to 1-\gamma_G$.

3.2. Separating from a Smart Zealot (Revisited)

- Consider again the case from the last section, with $\gamma_G = \gamma_B = \gamma$, but allowing π to take any value (not just $\frac{1}{2}$).

Generalizing the analysis of the previous section, one can show that if the good advisor tells the truth after history (λ, π, x, y), we must have $\sigma_B\left(1\left|\lambda,\pi,x,y\right.\right) =$

1 and $\sigma_B(1|\lambda, \pi, x, y) > 0$. Thus the corresponding (β_0, β_1) satisfy $\beta_1 = \gamma + (1-\gamma)\nu$ and $\beta_0 = (1-\gamma)(1-\nu)$. Since $\beta_1 > \gamma$ and $\beta_0 < 1-\gamma$, $(\beta_0, \beta_1) \notin T(\lambda, \pi)$. Truth-telling is never optimal for small x. See figures 3 and 4 for the example with $\gamma = \frac{3}{4}$ and π equal to $\frac{1}{2}$ and $\frac{9}{10}$ respectively.

3.3. Separating from an Honest Fool

- Suppose $\gamma_B < \gamma_G$ and the bad advisor always tells the truth.

Then $\beta_0 = 1 - \gamma_B$ and $\beta_1 = \gamma_B$. In this case, conditions [1] and [2] hold automatically and condition [3] reduces to $\pi \in [1-\gamma_G, \gamma_G]$. Thus the good advisor has a reputational incentive to tell the truth only if the decision problem is sufficiently symmetric. See figures 5 and 6 for the example with $\gamma = \frac{3}{4}$ and π equal to $\frac{1}{2}$ and $\frac{9}{10}$ respectively.

3.4. Separating from a Foolish Zealot

- Suppose $\gamma_B = \frac{1}{2}$ and $\gamma_G > \frac{1}{2}$, and the bad advisor has the biased preferences of the previous section.

In this case, the bad advisor will announce 1 more often that she will announce 0 but her stupidly introduces some slack in the definition of $T(\lambda, \pi)$.

Proposition 5. *Consider the reduced form game where $v_G(\lambda) = cv_B(\lambda)$ for some $c > 0$. Assume $\gamma_B = \frac{1}{2}$ and $\gamma_G > \frac{1}{2}$. Fix λ and π. There exists y^* such that if $y \leq y^*$, the (λ, x, y) game has a truth-telling equilibrium (for all x).*

See figures 7 and 8 for the example with $\gamma = \frac{3}{4}$ and π equal to $\frac{1}{2}$ and $\frac{9}{10}$ respectively.

3.5. Separating from a Smart Enemy

- In the earlier analysis, the zealot was assumed to have a systematic bias in his preferences. Consider the alternative case where the decision problem is symmetric and the bad advisor always wants the exact opposite of the decision maker. We focus again on the case where $\pi = \frac{1}{2}$, with $\gamma_G = \gamma_B = \gamma > \frac{1}{2}$.

Thus suppose now that $u_G(a,\omega) = w_G(|a-\omega|)$, where w_G is some strictly concave function, while the bad advisor's utility from action a in state ω is $u_B(a,\omega) = w_B(|a-(1-\omega)|)$. Thus the decision maker (and good advisor) want action 1 in state 1 and action 0 in state 0, while the bad advisor wants action 0 in state 1 and action 1 in state 0.

This is essentially the case studied by Benabou and Laroque (1982), building on the perfect signals analysis of Sobel (1985). They consider what happens in this setting assuming the good advisor always tells the truth ($\sigma_G(1) = 1$ and $\sigma_G(0) = 0$), and focussing on equilibria where the bad advisor always behaves symmetrically ($\sigma_G(1\,|\lambda) = 1 - \sigma_G(0\,|\lambda) = \xi(\lambda)$). They show that there is a unique monotonic Markov equilibrium within this class with $\xi(\lambda) = 0$ for all $\lambda \geq \lambda^*$ and $\xi(\lambda) \in (0,1)$ for all $\lambda < \lambda^*$.

We can verify if it would in fact be optimal for the good advisor to tell the truth if he cared only about the decision maker's utility. In the symmetric equilibria studied by Benabou and Laroque (with $\pi = \frac{1}{2}$), the answer is yes. The posterior reputation in such symmetric equilibria would be

$$\begin{aligned}\Lambda(1,1) &= \Lambda(0,0) = \lambda^+ = \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\xi(\lambda)+\left(\frac{1-\gamma}{\gamma}\right)(1-\xi(\lambda))\right)} > \lambda \\ \Lambda(1,0) &= \Lambda(0,1) = \lambda^- = \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\xi(\lambda)+\left(\frac{\gamma}{1-\gamma}\right)(1-\xi(\lambda))\right)} < \lambda\end{aligned}$$

Thus whatever signal is observed, the reputational value of telling the truth is $\gamma v_G(\lambda^+) + (1-\lambda)\, v_G(\lambda^-)$, while the reputational value of lying is $\gamma v_G(\lambda^-) + (1-\lambda)\, v_G(\lambda^+)$. So there is always a reputational benefit of telling the truth.

References

[1] Benabou, R. and G. Laroque (1992). "Using Privileged Information to Manipulate Markets: Insiders, Gurus and Credibility," *Quarterly Journal of Economics* 107, 921-958.

[2] Bernheim, D. (1994). "A Theory of Conformity," *Journal of Political Economy* 102, 841-877.

[3] Crawford, V. and J. Sobel (1982). "Strategic Information Transmission," *Econometrica* 50, 1431-1451.

[4] Dewatripont, M. and J. Tirole (1995). “Advocates.”

[5] Geanakoplos, J., D. Pearce and E. Stacchetti (1989). “Psychological Games and Sequential Rationality,” *Games and Economics Behavior* 1, 60-79.

[6] Loury, G. (1994). “Self-Censorship in Public Discourse: A Theory of ‘Political Correctness’ and Related Phenomena,” *Rationality and Society* 6, 428-461.

[7] Milgrom, P. and J. Roberts (1986). “Relying on the Information of Informed Parties,” *Rand Journal of Economics* 17, 18-32.

[8] Prendergast, C. and L. Stole (1996). “Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning,” *Journal of Political Economy* 104, 1105-1134.

[9] Shin, H. (1996). “Adversarial and Inquisitorial Procedures in Arbitration.”

[10] Sobel, J. (1985). “A Theory of Credibility,” *Review of Economic Studies* 52, 557-573.

Appendix A: A Formal Definition of Equilibrium

To characterize equilibrium, write $\widehat{u}_I\left(m\,|s,(\sigma_G,\sigma_B,\chi)\right)$ for the expected utility of an advisor of type I if she sends message m on observing signal s, given strategy profile (σ_G,σ_B,χ); and $\widehat{u}_{DM}\left(x\,|m,(\sigma_G,\sigma_B,\chi)\right)$ for the expected utility of an advisor who observes message m.

$$\begin{aligned}
\widehat{u}_G\left(m\,|1,(\sigma_G,\sigma_B,\chi)\right) &= \left\{\begin{array}{l} \gamma\left(x.u_{DM}\left(\chi(m),1\right)+v_G\left[\Lambda\left[\sigma_G,\sigma_B\right](m,1)\right]\right) \\ +(1-\gamma)\left(x.u_{DM}\left(\chi(m),0\right)+v_G\left[\Lambda\left[\sigma_G,\sigma_B\right](m,0)\right]\right) \end{array}\right\}; \\
\widehat{u}_G\left(m\,|0,(\sigma_G,\sigma_B,\chi)\right) &= \left\{\begin{array}{l} (1-\gamma)\left(x.u_{DM}\left(\chi(m),1\right)+v_G\left[\Lambda\left[\sigma_G,\sigma_B\right](m,1)\right]\right) \\ +\gamma\left(x.u_{DM}\left(\chi(m),0\right)+v_G\left[\Lambda\left[\sigma_G,\sigma_B\right](m,0)\right]\right) \end{array}\right\}; \\
\widehat{u}_B\left(m\,|1,(\sigma_G,\sigma_B,\chi)\right) &= \left\{\begin{array}{l} \gamma\left(y.u_B\left(\chi(m),1\right)+v_B\left[\Lambda\left[\sigma_G,\sigma_B\right](m,1)\right]\right) \\ +(1-\gamma)\left(y.u_B\left(\chi(m),0\right)+v_B\left[\Lambda\left[\sigma_G,\sigma_B\right](m,0)\right]\right) \end{array}\right\}; \\
\widehat{u}_B\left(m\,|0,(\sigma_G,\sigma_B,\chi)\right) &= \left\{\begin{array}{l} (1-\gamma)\left(y.u_B\left(\chi(m),1\right)+v_B\left[\Lambda\left[\sigma_G,\sigma_B\right](m,1)\right]\right) \\ +\gamma\left(y.u_B\left(\chi(m),0\right)+v_B\left[\Lambda\left[\sigma_G,\sigma_B\right](m,0)\right]\right) \end{array}\right\}; \\
\text{and } \widehat{u}_{DM}\left(a\,|m,(\sigma_G,\sigma_B,\chi)\right) &= \Gamma\left[\sigma_G,\sigma_B\right](m).u_{DM}(a,1)+\left(1-\Gamma\left[\sigma_G,\sigma_B\right](m)\right)u_{DM}(a,0).
\end{aligned}$$

Definition 3. *$(\sigma_G, \sigma_B, \chi)$ is an equilibrium if for each $I = G, B$,*

$$\begin{aligned} \sigma_I(s) > 0 \Rightarrow 1 \in & \underset{m \in \{0,1\}}{\arg\max}\, \widehat{u}_I(m\,|s, (\sigma_G, \sigma_B, \chi)) \\ \textit{and } \sigma_I(s) < 1 \Rightarrow 0 \in & \underset{m \in \{0,1\}}{\arg\max}\, \widehat{u}_I(m\,|s, (\sigma_G, \sigma_B, \chi))\,; \\ \textit{and } \chi(m) \in & \underset{a \in [0,1]}{\arg\max}\, \widehat{u}_{DM}(a\,|m, (\sigma_G, \sigma_B, \chi))\,. \end{aligned}$$

Formally, this reduced form game is equivalent to a psychological game in the sense of Geanakoplos, Pearce and Stacchetti (1989).

Appendix B: Proofs

Some preliminary notation and results will be useful. Write $\widehat{u}_G(q, s)$ for expected value of u_{DM} for the good advisor if he has observed signal s and the decision maker believes the true state is 1 with probability q,

$$\begin{aligned} \widehat{u}_G(q, 1) &= \gamma . u_{DM}(\widetilde{a}(q), 1) + (1-\gamma)\, u_{DM}(\widetilde{a}(q), 0) \\ \text{and } \widehat{u}_G(q, 0) &= (1-\gamma)\, u_{DM}(\widetilde{a}(q), 1) + \gamma . u_{DM}(\widetilde{a}(q), 0)\,. \end{aligned}$$

Similarly, write $\widehat{u}_B(q)$ for expected value of u_B for the bad advisor if the decision maker believes the true state is 1 with probability q; note that this is independent of the signal observed by the bad advisor:

$$\widehat{u}_B(q) = u_B(\widetilde{a}(q))\,.$$

We will use repeatedly the following properties of $\widehat{u}_G$ and $\widehat{u}_B$.

Lemma 3. *$\widehat{u}_G(q, 1)$ is strictly increasing in q if $q \in (1-\gamma, \gamma)$; $\widehat{u}_G(q, 0)$ is strictly decreasing in q if $q \in (1-\gamma, \gamma)$; $\widehat{u}_B(q)$ is strictly decreasing in q if $q \in (1-\gamma, \gamma)$.*

The following notation will also be useful. Given $(\sigma_B, \sigma_G, \chi)$, write $\Pi_I^C(s)$ for the net current expected gain to the type I advisor choosing message 1, rather than message 0, when she observes signal s, assuming the decision maker follows his optimal strategy, i.e., $\Pi_G^C(s) = x\left[\widehat{u}_G(\Gamma(1), s) - \widehat{u}_G(\Gamma(0), s)\right]$ and $\Pi_B^C(s) = y\left[\widehat{u}_G(\Gamma(1)) - \widehat{u}_G(\Gamma(0))\right]$. Write $\Pi_I^R(s)$ for the net expected reputational gain to the type I advisor of choosing message 0 rather than 1 when she observes signal s, i.e.,

$$\begin{aligned}\Pi_I^R(1) &= \gamma\begin{bmatrix} v_I(\Lambda(0,1)) \\ -v_I(\Lambda(1,1))\end{bmatrix} + (1-\gamma)\begin{bmatrix} v_I(\Lambda(0,0)) \\ -v_I(\Lambda(1,0))\end{bmatrix} \\ \Pi_I^R(0) &= (1-\gamma)\begin{bmatrix} v_I(\Lambda(0,1)) \\ -v_I(\Lambda(1,1))\end{bmatrix} + \gamma\begin{bmatrix} v_I(\Lambda(0,0)) \\ -v_I(\Lambda(1,0))\end{bmatrix}\end{aligned}$$

Thus an advisor of type I has a strict incentive to announce 1 when observing signal s exactly if $\Pi_I^C(s) > \Pi_I^R(s)$.

PROOF OF LEMMA 1

If the decision maker believes that the probability of state 1 is q, his expected utility from action a is

$$q.u_{DM}(a,1) + (1-q)\,u_{DM}(a,0).$$

This maximand is strictly convex in a and uniquely achieves a maximum when

$$q.u'_{DM}(a,1) + (1-q)\,u'_{DM}(a,0) = 0. \quad \blacksquare$$

PROOF OF PROPOSITION 2.

This will be proved via a pair of lemmas.

Lemma 4. *In any equilibrium* $(\sigma_G, \sigma_B, \chi)$, *we must have* $\Lambda(0,1) \geq \Lambda(1,1)$ *and* $\Lambda(0,0) \geq \Lambda(1,0)$.

Proof. We will show by contradiction that no other equilibria exist. Recall that if $(\sigma_G, \sigma_B, \chi)$ is an equilibrium, $\chi = \widetilde{\chi}[\sigma_G, \sigma_B]$, and that we are assuming (without loss of generality) that $\chi(1) \geq \chi(0)$.

Case 1. Suppose that $\Lambda(1,1) > \Lambda(0,1)$ and $\Lambda(1,0) > \Lambda(0,0)$. Now $\Pi_B^R(s) < 0$ and $\Pi_B^C(s) \geq 0$ for each $s = 0, 1$, we must have $\sigma_B(0) = \sigma_B(1) = 1$. But now if $\sigma_G(0) = \sigma_G(1) = 1$, $\Lambda(1,1) = \Lambda(0,1) = \Lambda(1,0) = \Lambda(0,0) = \lambda$, a contradiction. But if $\sigma_G(0) \neq 1$ or $\sigma_G(1) \neq 1$, then $\Lambda(0,1) = \Lambda(0,0) = 1$, another contradiction. Thus there is no such equilibrium.

Case 2. Suppose that $\Lambda(1,1) > \Lambda(0,1)$ and $\Lambda(1,0) \leq \Lambda(0,0)$. By definition of Λ, we have

$$\phi_G(1|1) > \phi_B(1|1) \tag{3.1}$$

$$\phi_G(1|0) \leq \phi_B(1|0) \tag{3.2}$$

Observe first that $\Pi_I^R(1) < \Pi_I^R(0)$ and $\Pi_I^C(1) \geq \Pi_I^C(0)$ for $I = B, G$. Thus for each I, $\sigma_I(0) = 0$ or $\sigma_I(1) = 1$. This implies four subcases:

(i) If $\sigma_G(0) = \sigma_B(0) = 0$, then (3.1) implies $\sigma_G(1) > \sigma_B(1)$, while (3.2) implies $\sigma_G(1) \leq \sigma_B(1)$, a contradiction.

(ii) If $\sigma_G(0) = 0$ and $\sigma_B(1) = 1$, then (3.1) implies $\sigma_G(1) > 1$, a contradiction.

(iii) If $\sigma_G(1) = 1$ and $\sigma_B(0) = 0$, then (3.2) implies $\sigma_B(1) = 1$ and $\sigma_G(0) = 0$, which implies $\phi_G(1|1) > \phi_B(1|1)$, contradicting (3.1).

(iv) If $\sigma_G(1) = \sigma_B(1) = 1$, then (3.1) implies $\sigma_G(0) > \sigma_B(0)$, while (3.2) implies $\sigma_G(0) \leq \sigma_B(0)$, a contradiction.

Case 3. Suppose that $\Lambda(1,1) \leq \Lambda(0,1)$ and $\Lambda(1,0) > \Lambda(0,0)$. By definition of Λ, we have

$$\phi_G(1|1) > \phi_B(1|1) \tag{3.3}$$

$$\phi_G(1|0) \leq \phi_B(1|0) \tag{3.4}$$

In this case, $\Pi_B^R(1) > \Pi_B^R(0)$ and $\Pi_B^C(1) = \Pi_B^C(0)$, so either $\sigma_B(1) = 0$ or $\sigma_B(0) = 1$. Thus $\phi_B(1|1) \leq \phi_B(1|0)$. By (3.3) and (3.4), this implies $\phi_G(1|1) < \phi_G(1|0)$. But now $\widetilde{\chi}[\sigma_G, \sigma_B][1] > \frac{1}{2} > \widetilde{\chi}[\sigma_G, \sigma_B][0]$, a contradiction. ■

Lemma 5. *In any non-babbling equilibrium,* $\chi(1) > \chi(0)$, $\Lambda(0,1) \geq \Lambda(1,1)$ *and* $\Lambda(0,0) \geq \Lambda(1,0)$, *with one of the latter two inequalities holding as a strict inequality.*

Proof. Lemma 4 implies the two weak inequalities. Suppose both held with equality. Recall that we have $\chi(1) \geq \chi(0)$ by assumption. If $\chi(1) > \chi(0)$, the bad advisor would have a strict incentive to choose 1 (whatever his signal), leading to a contradiction. But if $\chi(1) = \chi(0)$, we have a babbling equilibrium.

So assume at least one of the weak inequalities is strict. If $\chi(1) = \chi(0)$, the bad advisor would have a strict incentive to choose 0 (whatever his signal), leading again to a contradiction. ■

Lemma 5 proves part [2] of proposition 2. The rest of the proposition is proved as follows. By lemma 5, $\Pi_I^R(s) > 0$ for $I = G, B$ and $s = 0, 1$. Since $\Pi_G^C(0) < 0$, we have $\sigma_G(0) = 0$, proving part [1] of proposition 2. Now observe that

$$\begin{aligned}
\Lambda(1,1) &= \frac{\lambda\gamma\sigma_G(1)}{\lambda\gamma\sigma_G(1) + (1-\lambda)(\gamma\sigma_B(1) + (1-\gamma)\sigma_B(0))} \\
&= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\frac{1}{\sigma_G(1)}\left(\sigma_B(1) + \left(\frac{1-\gamma}{\gamma}\right)\sigma_B(0)\right)} \\
&\geq \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\frac{1}{\sigma_G(1)}\left(\sigma_B(1) + \left(\frac{\gamma}{1-\gamma}\right)\sigma_B(0)\right)} \\
&= \frac{\lambda(1-\gamma)\sigma_G(1)}{\lambda(1-\gamma)\sigma_G(1) + (1-\lambda)((1-\gamma)\sigma_B(1) + \gamma\sigma_B(0))} \\
&= \Lambda(1,0)
\end{aligned}$$

Now consider three cases.

Case 1: $\Lambda(0,0) \geq \Lambda(0,1)$; together with our earlier assumptions and results, this implies $\Lambda(0,0) \geq \Lambda(0,1) \geq \Lambda(1,1) \geq \Lambda(1,0)$ with at least one strict inequality. Now $\Pi_B^R(0) > 0 \Rightarrow \Pi_B^R(1) > 0$, so either $\sigma_B(0) = 0$ or $\sigma_B(1) = 1$. But $\Lambda(0,0) \geq \Lambda(0,1)$ implies that $\frac{\phi_B(0|0)}{\phi_G(0|0)} \leq \frac{\phi_B(0|1)}{\phi_G(0|1)}$, i.e., $\frac{\phi_B(0|0)}{\phi_B(0|1)} \leq \frac{\phi_G(0|0)}{\phi_G(0|1)}$. But

$$\frac{\phi_G(0|0)}{\phi_G(0|1)} = \frac{(1-\gamma)(1-\sigma_G(1)) + \gamma}{\gamma(1-\sigma_G(1)) + 1 - \gamma} \leq \frac{\gamma}{1-\gamma}$$

Now if $\sigma_B(0) = 0$, then

$$\frac{\phi_B(0|0)}{\phi_B(0|1)} = \frac{(1-\gamma)(1-\sigma_B(1)) + \gamma}{\gamma(1-\sigma_B(1)) + 1 - \gamma}$$

which is less than or equal to $\frac{\phi_G(0|0)}{\phi_G(0|1)}$ only if $\sigma_B(1) \leq \sigma_G(1)$. But this implies $\phi_B(1|\omega) \leq \phi_G(1|\omega)$ for $\omega = 1, 2$, a contradiction.

But if $\sigma_B(1) = 1$, then

$$\frac{\phi_B(0|0)}{\phi_B(0|1)} = \frac{\gamma(1-\sigma_B(0))}{(1-\gamma)(1-\sigma_B(0))} = \frac{\gamma}{1-\gamma}$$

which is less than or equal to $\frac{\phi_G(0|0)}{\phi_G(0|1)}$ only if $\sigma_G(1) = 1$. This gives our first class of "truth-telling equilibria" with $\sigma_G(0) = 0$, $\sigma_G(1) = 1$ and $\sigma_B(1) = 1$. If

$\sigma_B(0) = 0$, we have $\phi_B(1|\omega) \leq \phi_G(1|\omega)$ for $\omega = 1, 2$, a contradiction. Now

$$\begin{aligned}
\Lambda(1,1) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(1 + \left(\frac{1-\gamma}{\gamma}\right)\sigma_B(0)\right)} \\
\Lambda(1,0) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(1 + \left(\frac{\gamma}{1-\gamma}\right)\sigma_B(0)\right)} \\
\Lambda(0,1) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)(1-\nu)} \\
\Lambda(0,0) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)(1-\nu)}
\end{aligned}$$

and $\Lambda(0,1) = \Lambda(0,0) > \lambda > \Lambda(1,1) > \Lambda(1,0)$.

Case 2: $\Lambda(0,1) > \Lambda(0,0)$ and $\sigma_B(0) = 0$. If $\sigma_G(1) \geq \sigma_B(1)$, we have $\phi_B(1|\omega) \leq \phi_G(1|\omega)$ for $\omega = 1, 2$, a contradiction. Thus $\sigma_G(1) < \sigma_B(1)$ and

$$\begin{aligned}
\Lambda(1,1) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\sigma_B(1)}{\sigma_G(1)}\right)} \\
\Lambda(1,0) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\sigma_B(1)}{\sigma_G(1)}\right)} \\
\Lambda(0,1) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-\gamma\sigma_B(1)}{1-\gamma\sigma_G(1)}\right)} \\
\Lambda(0,0) &= \frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-(1-\gamma)\sigma_B(1)}{1-(1-\gamma)\sigma_G(1)}\right)}
\end{aligned}$$

and $\Lambda(0,1) > \Lambda(0,0) > \lambda > \Lambda(1,1) = \Lambda(1,0)$.

Case 3: $\sigma_B(0) > 0$ and $\Lambda(0,1) > \Lambda(0,0)$. The latter requires $\frac{\phi_B(0|0)}{\phi_G(0|0)} > \frac{\phi_B(0|1)}{\phi_G(0|1)}$, i.e.,

$$\frac{(1-\gamma)(1-\sigma_B(1)) + \gamma(1-\sigma_B(0))}{\gamma(1-\sigma_B(1)) + (1-\gamma)(1-\sigma_B(0))} = \frac{\phi_B(0|0)}{\phi_B(0|1)} > \frac{\phi_G(0|0)}{\phi_G(0|1)} = \frac{(1-\gamma)(1-\sigma_G(1)) + \gamma}{\gamma(1-\sigma_G(1)) + 1 - \gamma}$$

This holds if and only if

$$\frac{(1-\gamma)\left(\frac{1-\sigma_B(1)}{1-\sigma_B(0)}\right) + \gamma}{\gamma\left(\frac{1-\sigma_B(1)}{1-\sigma_B(0)}\right) + 1 - \gamma} > \frac{(1-\gamma)(1-\sigma_G(1)) + \gamma}{\gamma(1-\sigma_G(1)) + 1 - \gamma}$$

$$
\begin{aligned}
&\Leftrightarrow \quad \left(\frac{1-\sigma_B(1)}{1-\sigma_B(0)}\right) < 1-\sigma_G(1) \\
&\Leftrightarrow \quad \sigma_B(1) > 1-(1-\sigma_G(1))(1-\sigma_B(0)). \quad \blacksquare
\end{aligned}
$$

PROOF OF PROPOSITION 3.

[1] *Truth-Telling.* By proposition 2 and the definition of a truth telling strategy, we must have $\sigma_G(0)=0$, $\sigma_G(1)=1$, $\sigma_B(0)=\nu$ for some $\nu>0$, $\sigma_B(1)=1$, and $\chi=\widetilde{\chi}(\sigma_G,\sigma_B)$.

Under these strategies,

$$
\begin{aligned}
\Gamma(1) &= \frac{\gamma+(1-\lambda)(1-\gamma)\nu}{1+(1-\lambda)\nu}; \\
\Gamma(0) &= 1-\gamma; \\
\Lambda(1,1) &= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\nu\right)}; \\
\Lambda(1,0) &= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\nu\right)}; \\
\Lambda(0,1) &= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1-\nu)}; \\
\text{and } \Lambda(0,0) &= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1-\nu)}.
\end{aligned}
$$

Write $g(\nu)$ for the utility gain to the bad advisor of announcing 1 (rather than 0) when his signal is 0, i.e.,

$$
g(\nu)=\left\{\begin{array}{l}
y\left(\widehat{u}_B\left(\frac{\gamma+(1-\lambda)(1-\gamma)\nu}{1+(1-\lambda)\nu}\right)-\widehat{u}_B(1-\gamma)\right)+\gamma v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\nu\right)}\right] \\
+(1-\gamma)v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\nu\right)}\right]-v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1-\nu)}\right]
\end{array}\right\}.
$$

This expression is decreasing in ν, since each term is weakly decreasing in ν. Also $g(0)=y(\widehat{u}_B(\gamma)-\widehat{u}_B(1-\gamma))>0$. Thus there exists exactly one value of ν where either $g(\nu)=0$ or $\nu=1$ and $g(\nu)>0$. This ν thus parameterizes the unique equilibrium.

Write $\widetilde{\nu}(\lambda,y)$ for that unique value of ν (for given λ and y). Observe that [1] $\widetilde{\nu}(\lambda,y)\to 0$ as $y\to 0$ and [2] $\widetilde{\nu}(\lambda,y)=1$ for all sufficiently large y. Also [3]

$\widetilde{\nu}(\lambda, y) = 1$ for all λ sufficiently close to 1. But what happens for small λ? There are two cases to consider. Either

$$y\left(\widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma)\right) > v_B[1] - v_B[0],$$

in which case $\widetilde{\nu}(\lambda, y) = 1$ for all sufficiently small λ; or

$$y\left(\widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma)\right) \leq v_B[1] - v_B[0],$$

and

$$\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1-\widetilde{\nu}(\lambda,y))} \longrightarrow v_B^{-1}\left[v_B[0] + y\left(\widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma)\right)\right]$$

as $\lambda \longrightarrow 0$. Write

$$h_B(y) = v_B^{-1}\left(\min\left\{v_B[1], v_B[0] + y\left(\widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma)\right)\right\}\right)$$

Now $\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1-\widetilde{\nu}(\lambda,y)\right)} \longrightarrow h_B(y)$ and $\widetilde{\nu}(\lambda, y) \longrightarrow 1$ as $\lambda \longrightarrow 0$.

Now consider the good advisor's incentive to tell the truth when she observes signal 1 under strategy profile $\sigma_G(0) = 0$, $\sigma_G(1) = 1$, $\sigma_B(0) = \widetilde{\nu}(\lambda, y)$, $\sigma_B(1) = 1$, and $\chi = \widetilde{\chi}(\sigma_G, \sigma_B)$. She will tell the truth if and only if

$$\left\{\begin{array}{c} x\left[\widehat{u}_G\left(\frac{\gamma+(1-\lambda)(1-\gamma)\widetilde{\nu}(\lambda,y)}{1+(1-\lambda)\widetilde{\nu}(\lambda,y)}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right] \\ +\gamma v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right] + (1-\gamma) v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right] \\ -v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1-\widetilde{\nu}(\lambda,y)\right)}\right] \end{array}\right\} \geq 0,$$

i.e.,

$$\begin{array}{rcl} x & \geq & \widetilde{x}_2(\lambda, y) \\ & = & \left\{\frac{v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1-\widetilde{\nu}(\lambda,y)\right)}\right] - \gamma v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right] - (1-\gamma) v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right]}{\left[\widehat{u}_G\left(\frac{\gamma+(1-\lambda)(1-\gamma)\widetilde{\nu}(\lambda,y)}{1+(1-\lambda)\widetilde{\nu}(\lambda,y)}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right]}\right. \end{array}$$

Now $\widetilde{\nu}(\lambda, y) = 1$ for all sufficiently large λ, so $\widetilde{x}_2(\lambda, y) \rightarrow 0$ as $\lambda \rightarrow 1$. As $\lambda \rightarrow 0$, $\widetilde{\nu}(\lambda, y) \rightarrow 1$ and $\left(\frac{1-\lambda}{\lambda}\right)(1 - \widetilde{\nu}(\lambda, y)) \rightarrow h_B(y)$, so

$$\begin{aligned} \widetilde{x}_2(\lambda, y) &\rightarrow \frac{v_G[h_B(y)] - v_G[0]}{\left[\widehat{u}_G\left(\frac{1}{2}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right]} \\ &= \frac{v_G\left[v_B^{-1}\left(\min\left\{v_B[1], v_B[0] + y\left(\widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma)\right)\right\}\right)\right] - v_G[0]}{\left[\widehat{u}_G\left(\frac{1}{2}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right]}. \end{aligned}$$

As $y \rightarrow 0$, $\widetilde{\nu}(\lambda, y) \rightarrow 0$ and thus $\widetilde{x}_2(\lambda, y) \rightarrow 0$; as $y \rightarrow \infty$, $\widetilde{\nu}(\lambda, y) \rightarrow 1$ and thus

$$\widetilde{x}_2(\lambda, y) \rightarrow \left\{\frac{v_G[1] - \gamma v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\right)}\right] - (1-\gamma) v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\right)}\right]}{\left[\widehat{u}_G\left(\frac{\gamma+(1-\lambda)(1-\gamma)}{1+(1-\lambda)}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right]}\right\}.$$

[2] *Babbling.* This will be proved via a series of lemmas.

Lemma 6. *In any non-babbling equilibrium,* $\Pi_G^R(1) \leq \Pi_G^C(1) \leq x[\widehat{u}_G(\gamma, 1) - \widehat{u}_G(1-\gamma, 1)]$.

Definition 4. *ϕ_G and ϕ_B are δ-close if for each $\omega \in \{0, 1\}$,*

$$\frac{1}{1+\delta} \leq \frac{\phi_B(0|\omega)}{\phi_G(0|\omega)} \leq 1 \leq \frac{\phi_B(1|\omega)}{\phi_G(1|\omega)} \leq 1 + \delta.$$

Now let $f(\lambda, \delta) = (1-\gamma)\min\left\{v_G(\lambda) - v_G\left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1+\delta)}\right), v_G\left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1}{1+\delta}\right)}\right) - v_G(\lambda)\right\}$.

Lemma 7. *If $\Pi_G^R(1) < f(\lambda, \delta)$ in some equilibrium, then ϕ_G and ϕ_B are δ-close.*

Proof. Suppose ϕ_G and ϕ_B are not δ-close. Then $\frac{\phi_B(1|\omega)}{\phi_G(1|\omega)} > 1 + \delta$ or $\frac{\phi_B(0|\omega)}{\phi_G(0|\omega)} < \frac{1}{1+\delta}$ for some ω. So

$$\Pi_G^R(1) = \gamma\left[\begin{array}{c} v_G(\Lambda(0,1)) \\ -v_G(\Lambda(1,1)) \end{array}\right] + (1-\gamma)\left[\begin{array}{c} v_G(\Lambda(0,0)) \\ -v_G(\Lambda(1,0)) \end{array}\right] > f(\lambda, \delta)$$

If $\kappa \geq v_B(1) - v_B(0)$, let $g(\lambda, \kappa) = \infty$; if $\kappa < v_B(1) - v_B(0)$, let $g(\lambda, \kappa)$ be the unique value of δ solving

$$\kappa = v_B\left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1}{1+\delta}\right)}\right) - v_B\left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)(1+\delta)}\right). \quad \blacksquare$$

Lemma 8. *Fix $\kappa > 0$. If ϕ_B and ϕ_G are $g(\lambda, \kappa)$-close, then*

$$\Pi_B^R(1) = \gamma \left[\begin{array}{c} v_B(\Lambda(0,1)) \\ -v_B(\Lambda(1,1)) \end{array} \right] + (1-\gamma) \left[\begin{array}{c} v_B(\Lambda(0,0)) \\ -v_B(\Lambda(1,0)) \end{array} \right] \leq \kappa.$$

Lemma 9. *If ϕ_B and ϕ_G are δ-close, then $\Gamma(1) \geq \frac{\gamma}{\gamma+(1-\gamma)(1+\delta)}$.*

Proof. Write $\zeta = \sigma_G(1)$. Now:

$$\begin{aligned} \phi_G(1|1) &= \gamma\zeta \\ \phi_G(1|0) &= (1-\gamma)\zeta \end{aligned}$$

$$\begin{aligned} \gamma\zeta &\leq \phi_B(1|1) \leq \gamma\zeta(1+\delta) \\ (1-\gamma)\zeta &\leq \phi_B(1|0) \leq (1-\gamma)\zeta(1+\delta) \end{aligned}$$

$$\begin{aligned} \Gamma(1) &= \frac{\lambda\phi_G(1|1) + (1-\lambda)\phi_B(1|1)}{\lambda\phi_G(1|1) + (1-\lambda)\phi_B(1|1) + \lambda\phi_G(1|0) + (1-\lambda)\phi_B(1|0)} \\ &\geq \frac{\gamma\zeta}{\gamma\zeta + (1-\gamma)\zeta(\lambda + (1-\lambda)(1+\delta))} \\ &= \frac{\gamma}{\gamma + (1-\gamma)(\lambda + (1-\lambda)(1+\delta))} \\ &\geq \frac{\gamma}{\gamma + (1-\gamma)(1+\delta)} \end{aligned}$$

Now define:

$$\tilde{x}_1(\lambda, y) = \frac{f\left(\lambda, \min\left\{\frac{2\gamma-1}{2(1-\gamma)}, g\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right)\right\}\right)}{\widehat{u}_G(\gamma, 1) - \widehat{u}_G(1-\gamma, 1)}$$

Suppose $x \leq \tilde{x}_1(\lambda, y)$. The following claims must hold true of any non-babbling equilibrium. By lemma 6,

$$\Pi_G^R(1) \leq f\left(\lambda, \min\left\{\frac{2\gamma-1}{2(1-\gamma)}, g\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right)\right\}\right)$$

By lemma 7, ϕ_G and ϕ_B are

$$\min\left\{\frac{2\gamma-1}{2(1-\gamma)}, g\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right)-\widehat{u}_B\left(\frac{1}{2}\right)\right)\right)\right\}\text{-close}$$

By lemma 8, $\Pi_B^R(1) \leq \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right)-\widehat{u}_B\left(\frac{1}{2}\right)\right)$. But $\Gamma(0) \leq \frac{1}{2}$ and, setting $\delta = \frac{2\gamma-1}{2(1-\gamma)}$ in lemma 9, $\Gamma(1) \geq \widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right)$. Thus $\Pi_B^R(1) \geq y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right)-\widehat{u}_B\left(\frac{1}{2}\right)\right)$, a contradiction. ■

Proof of Proposition 4

Fix (x^*, y^*) and consider the following advisor strategy

$$\begin{aligned}
\sigma_G(s\,|\lambda, x, y) &= \begin{cases} \frac{1}{2}, \text{ if } (x,y) \neq (x^*,y^*) \\ s, \text{ if } (x,y) = (x^*,y^*) \end{cases} \\
\text{and } \sigma_B(s\,|\lambda, x, y) &= \begin{cases} \frac{1}{2}, \text{ if } (x,y) \neq (x^*,y^*) \\ 1, \text{ if } (x,y) = (x^*,y^*) \end{cases}.
\end{aligned}$$

The best response for the decision maker is

$$\chi(m\,|\lambda, x, y) = \begin{cases} \widetilde{a}\left(\frac{1}{2}\right), \text{ if } (x,y) \neq (x^*,y^*) \\ \widetilde{a}\left(\frac{\lambda\gamma+(1-\lambda)}{\lambda+2(1-\lambda)}\right), \text{ if } (x,y) = (x^*,y^*) \text{ and } m = 1 \\ \widetilde{a}(1-\gamma), \text{ if } (x,y) = (x^*,y^*) \text{ and } m = 0 \end{cases}.$$

The value function for the good advisor must satisfy $v_G = T_G[v_G]$ where

$$T_G[v_G](\lambda) = \left\{ \begin{array}{l} (1-\varepsilon)\left[\frac{1}{2}\widehat{u}_G\left(\frac{1}{2},1\right)+\frac{1}{2}\widehat{u}_G\left(\frac{1}{2},1\right)+\delta_G v_G(\lambda)\right] \\ +\varepsilon\left[\begin{array}{l} \frac{1}{2}\widehat{u}_G\left(\frac{\lambda\gamma+(1-\lambda)}{\lambda+2(1-\lambda)},1\right)+\frac{1}{2}\widehat{u}_G(1-\gamma,0) \\ +\delta_G\left[\frac{1}{2}\gamma v_G\left(\frac{\lambda\gamma}{\lambda\gamma+1-\lambda}\right)+\frac{1}{2}(1-\gamma)v_G\left(\frac{\lambda(1-\gamma)}{\lambda(1-\gamma)+1-\lambda}\right)+\frac{1}{2}v_G(1)\right] \end{array}\right] \end{array} \right\}.$$

The value function for the bad advisor must satisfy $v_B = T_B[v_B]$ where

$$T_B[v_B](\lambda) = \left\{ \begin{array}{l} (1-\varepsilon)\left[\widehat{u}_B\left(\frac{1}{2}\right)+\delta_B v_B(\lambda)\right] \\ +\varepsilon\left[\widehat{u}_B\left(\frac{\lambda\gamma+(1-\lambda)}{\lambda+2(1-\lambda)}\right)+\delta_B\left[\frac{1}{2}v_B\left(\frac{\lambda\gamma}{\lambda\gamma+1-\lambda}\right)+\frac{1}{2}v_B\left(\frac{\lambda(1-\gamma)}{\lambda(1-\gamma)+1-\lambda}\right)\right]\right] \end{array} \right\}.$$

Each T_I maps the set of strictly non-decreasing continuous functions on $[0,1]$ continuously onto itself. By construction, $T_I(v+c) = T_I(v) + \delta c$. So by Blackwell's contraction mapping theorem, each equation has a unique strictly increasing continuous fixed point.

Now we must verify optimality. Observe that

$$v_G(1) - v_G(0) \le \frac{\varepsilon}{1-\delta_G}\left[\begin{array}{c} \frac{1}{2}\left(\widehat{u}_G(\gamma,1) - \widehat{u}_G(1-\gamma,1)\right) \\ +\frac{1}{2}\left(\widehat{u}_G(1-\gamma,0) - \widehat{u}_G(\gamma,0)\right) \end{array}\right]$$

$$\text{and } v_B(1) - v_B(0) \le \frac{\varepsilon}{1-\delta_B}\left[\widehat{u}_B(\gamma) - \widehat{u}_B(1-\gamma)\right].$$

Now suppose that each player follows the candidate strategies. Any strategy is always a best response to babbling. We must check that it is optimal to follow the proposed strategies when $(x,y) = (x^*,y^*)$. Observe that the current expected gains (to both types) from following the proposed strategies are bounded below (independently of λ), i.e.,

$$\begin{aligned} \Pi_B^C(1) &= \Pi_B^C(0) = \widehat{u}_B\left(\frac{\lambda\gamma + (1-\lambda)}{\lambda + 2(1-\lambda)}\right) - \widehat{u}_B(1-\gamma) \\ &\ge \widehat{u}_B\left(\frac{1}{2}\right) - \widehat{u}_B(1-\gamma) \end{aligned}$$

$$\begin{aligned} \text{and } \Pi_G^C(1) &= \left\{\begin{array}{l} \gamma\left[\widehat{u}_G\left(\frac{\lambda\gamma+(1-\lambda)}{\lambda+2(1-\lambda)},1\right) - \widehat{u}_G(1-\gamma,1)\right] \\ +(1-\gamma)\left[\widehat{u}_G\left(\frac{\lambda\gamma+(1-\lambda)}{\lambda+2(1-\lambda)},0\right) - \widehat{u}_G(1-\gamma,0)\right] \end{array}\right\} \\ &\ge \left\{\begin{array}{l} \gamma\left[\widehat{u}_G\left(\frac{1}{2},1\right) - \widehat{u}_G(1-\gamma,1)\right] \\ +(1-\gamma)\left[\widehat{u}_G\left(\frac{1}{2},0\right) - \widehat{u}_G(1-\gamma,0)\right] \end{array}\right\}. \end{aligned}$$

Thus we can choose ε sufficiently small to ensure optimality. ■

Proof of Lemma 2

The net reputational gain to the good advisor from announcing 0 (rather than 1) if her signal were 1 is

$$\Pi_G^R(m\,|1) = \left\{ \begin{array}{c} \frac{\pi\gamma_G}{\pi\gamma_G+(1-\pi)(1-\gamma_G)} \left(\begin{array}{c} v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)(1-\beta_1)}\right) \\ -v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)\beta_1}\right) \end{array} \right) \\ +\frac{(1-\pi)(1-\gamma_G)}{\pi\gamma_G+(1-\pi)(1-\gamma_G)} \left(\begin{array}{c} v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)(1-\beta_0)}\right) \\ -v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)\beta_0}\right) \end{array} \right) \end{array} \right\}.$$

The net reputational gain to the good advisor from announcing 0 (rather than 1) if her signal were 0 is:

$$\Pi_G^R(m\,|0) = \left\{ \begin{array}{c} \frac{\pi(1-\gamma_G)}{\pi(1-\gamma_G)+(1-\pi)\gamma_G} \left(\begin{array}{c} v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)(1-\beta_1)}\right) \\ -v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)\beta_1}\right) \end{array} \right) \\ +\frac{(1-\pi)\gamma_G}{\pi(1-\gamma_G)+(1-\pi)\gamma_G} \left(\begin{array}{c} v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)(1-\beta_0)}\right) \\ -v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)\beta_0}\right) \end{array} \right) \end{array} \right\}.$$

Now

$$T(\lambda,\pi) = \left\{(\beta_0,\beta_1) : \Pi_G^R(m\,|0) \geq 0 \text{ and } \Pi_G^R(m\,|1) \leq 0\right\}.$$

This requires first that

$$\begin{array}{rcl} v_G\left(\dfrac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)(1-\beta_0)}\right) & \geq & v_G\left(\dfrac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)\beta_0}\right) \\ \text{and } v_G\left(\dfrac{\lambda\gamma_G}{\lambda\gamma_G+(1-\lambda)\beta_1}\right) & \geq & v_G\left(\dfrac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)(1-\beta_1)}\right). \end{array}$$

Thus $\beta_0 \geq 1-\gamma_G$ and $\beta_1 \leq \gamma_G$. ■

Proof of Proposition 5

Assume the good advisor tells the truth and the bad advisor announces 1 with probability μ (i.e., $\mu = \frac{1}{2}\sigma_B(1) + \frac{1}{2}\sigma_B(0)$). The reputational gain to the bad advisor of announcing 0 is

$$h(\pi,\lambda,\mu) = \left\{ \begin{array}{c} \pi\left(v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-\mu}{1-\gamma}\right)}\right] - v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\mu}{\gamma}\right)}\right]\right) \\ (1-\pi)\left(v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-\mu}{\gamma}\right)}\right] - v_B\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\mu}{1-\gamma}\right)}\right]\right) \end{array} \right\}.$$

Observe that $h(\pi, \lambda, \mu)$ is strictly increasing in μ with

$$\begin{aligned} h(\pi, \lambda, 1-\gamma) &= \pi \left(v_B \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\gamma}{1-\gamma}\right)} \right] - v_B \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-\gamma}{\gamma}\right)} \right] \right) < 0 \\ \text{and } h(\pi, \lambda, \gamma) &= (1-\pi) \left(v_B \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1-\gamma}{\gamma}\right)} \right] - v_B \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right)\left(\frac{\gamma}{1-\gamma}\right)} \right] \right) > 0. \end{aligned}$$

If

$$h(\pi, \lambda, 1) = \left\{ \begin{array}{c} v_B(1) - \pi v_B \left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1}{\gamma}\right)} \right) \\ -(1-\pi) v_B \left(\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(\frac{1}{1-\gamma}\right)} \right) \end{array} \right\} \leq y \left[\widehat{u}_B(\gamma) - \widehat{u}_B(1-\gamma) \right],$$

set $\widetilde{\mu}(\pi, \lambda, y) = 1$. Otherwise, let $\widetilde{\mu}(\pi, \lambda, y)$ be the unique solution to

$$h(\pi, \lambda, \mu) = y \left[\widehat{u}_B(\gamma) - \widehat{u}_B(1-\gamma) \right].$$

For the good advisor to always have incentive to tell the truth, we must have

$$\begin{aligned} \frac{1-\gamma_G}{\gamma_G} &\leq f_G(\mu) \leq \frac{\gamma_G}{1-\gamma_G} \\ \text{where } f_G(\mu) &= \left(\frac{\pi}{1-\pi} \right) \left(\frac{ v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G + (1-\lambda)\mu} \right) - v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)(1-\mu)} \right) }{ v_G\left(\frac{\lambda\gamma_G}{\lambda\gamma_G + (1-\lambda)(1-\mu)} \right) - v_G\left(\frac{\lambda(1-\gamma_G)}{\lambda(1-\gamma_G)+(1-\lambda)\mu} \right) } \right) \end{aligned}$$

Now $f_G(\widetilde{\mu}(\pi, \lambda, 0)) = 1$ and $f_G(\widetilde{\mu}(\pi, \lambda, y))$ is continuously decreasing in y. So for any given π and λ, we can choose y^* such that $f_G(\widetilde{\mu}(\pi, \lambda, y)) \in \left(\frac{1-\gamma_G}{\gamma_G}, 1 \right]$ for all $y \in [0, y^*]$. ■