Discussion Paper No. 1209

An Instrumental Theory of Political Correctness

by

Stephen Morris

February 1998

*An Instrumental Theory of Political Correctness**

Stephen Morris†
University of Pennsylvania and
Northwestern University

September 1997
revised February 1998

Abstract

An informed advisor wishes to convey her valuable information to an uninformed decision maker with identical preferences. Thus she has a current incentive to truthfully reveal her information. But if the decision maker thinks the advisor might be biased in favor of one decision, and the advisor does not wish to be thought to be biased, the advisor has a reputational incentive to lie. I show that if the advisor is sufficiently concerned about her reputation, no information is conveyed in equilibrium. I also show that in a repeated version of this game, the advisor will care (instrumentally) about her reputation simply because she wants her valuable and unbiased advice to have an impact on future decisions.

*I have benefited from the comments of seminar participants at Georgetown, Michigan, Northwestern, Warwick and Western Ontario; and from valuable conversations with David Austen-Smith, Stephen Coate, George Mailath and Andrew Postlewaite. I gratefully acknowledge financial support from the Alfred P. Sloan Foundation.

†Center for Mathematical Studies in Economics and Management Science, Northwestern University, Leverone Hall 371, 2001 Sheridan Road, Evanston IL 60208. Electronic Mail: smorris@econ.sas.upenn.edu.

1. Introduction

Consider the plight of an informed social scientist advising an uninformed policy maker on the merits of affirmative action by race. If the social scientist were racist, she would oppose affirmative action. In fact, she is not racist; but suppose she has come to the conclusion that affirmative action is an ill-conceived policy to address racism. The policy maker is not racist, but since he attaches a high probability to the social scientist not being racist, he would take an anti-affirmative action recommendation seriously and adjust government policy accordingly. But an anti-affirmative action recommendation would increase the probability that the policy maker attaches to the social scientist being racist. If the social scientist is *sufficiently* concerned about being perceived to be racist, she will have an incentive to lie and recommend affirmative action. But this being the case, she would not be believed even if she sincerely believed in affirmative action, and recommended it. Either way, the social scientist's socially valuable information is lost.[1]

Should we expect the social scientist to be *that* concerned about her reputation? While there are many reasons why the social scientist would not wish to be perceived to be racist, would not a social scientist sufficiently concerned about social welfare tell the truth? The answer is no, if the social scientist expects to be a regular participant in public policy debate (and cares enough about the outcomes of that debate). Suppose that (1) the social scientist cares only about the policy maker's policy decisions now and in the future; (2) the social scientist will have valuable information about many of those future decisions; and (3) the social scientist has *identical* preferences to the policy maker and in particular has no *intrinsic* reputational concerns. If the social scientist recommended affirmative action today, her reputation would decline. If she is believed to be racist, her advise on other policy issues will be discounted. Thus even though she has no intrinsic reputational concerns, she may have *instrumental* reputational concerns arising exclusively from her desire to have her unbiased and valuable advice listened to in the future.

This paper proposes a theory that captures this account. An informed "advi-

[1] A similar logic applies in many contexts. Consider, for example, a public figure who favored the Clinton health plan but was not in general (and did not wish to be perceived to be) in favor of government intervention in the economy; or a foreign policy analyst during the cold war who favored improved relations with Cuba but was not (and did not wish to be perceived to be) soft on communism.

sor" wishes to convey her valuable information to an uninformed "decision maker" with identical preferences. If talk is cheap, she has a *current* incentive to truthfully reveal her information. But suppose that in addition, the advisor is concerned about her reputation with the decision maker. In particular, the decision maker attaches positive probability to the advisor being "bad," i.e., having different preferences biased in favor of a particular decision. In this case, reputational concerns will give a "good" advisor an incentive to make (true or false) announcements that separate her from the bad advisor. I show that if reputational concerns are sufficiently important relative to the current decision problem, no information is conveyed in equilibrium; and I show that in a repeated version of this cheap talk game, the reputational concerns leading to this phenomenon arise for purely instrumental reasons.

The theory explains at least one aspect of so-called "political correctness." By political correctness, I mean the following phenomenon: because certain statements will lead listeners to make adverse inferences about the type of the speaker, speakers have an incentive to alter what they say to avoid that inference.[2] There is a harmless version of this phenomenon, when speakers use different signals (words) to convey their meaning (to avoid the adverse inferences) but listeners are nonetheless able to invert the signals and deduce the true meaning.[3] This paper is concerned with the potentially more important version, where speakers' attempts to avoid the adverse inference lead to real information being lost. In the model that I present, the information may be socially valuable: that is, all parties may lose from the suppression of information due to political correctness.

This paper follows Loury (1994) in developing a reputational explanation for political correctness.[4] Loury summarizes his argument in the following syllogism[5] (p. 437):

> (a) within a given community the people who are most faithful to

[2]The expression "political correctness" is sometimes also associated with a particular set of political views. As used in this paper, it is not.

[3]This harmless version is unlikely to be amenable to equilibrium analysis without endowing listeners with exogenous preferences over the choice of words (since the labelling of signals, i.e. the choice of words, is irrelevant in equilibrium analysis).

[4]Reputational concerns in social interaction more generally, and their implications, are the subject of Goffman (1959) and Kuran (1995).

[5]Loury does not present a formal model, but he notes that the theory of conformity of Bernheim (1994) could be adapted for the purpose.

communal values are by-and-large also those who want most to remain in good standing with their fellows and;

(b) the practice is well established in this community that those speaking in ways that offend community values are excluded from good standing. Then,

(c) when a speaker is observed to express himself offensively the odds that the speaker is not in fact faithful to communal values, as estimated by a listener otherwise uninformed about his views, are increased.

The explanation of this paper is narrower in scope but less "reduced form" than Loury's. My model is driven by specific assumptions about who is communicating with whom and why. But by making these specific assumptions, and by including valuable information in the model, it is possible to (1) explain *which* speech is "offensive" in equilibrium (i.e., lowers the reputation of the speaker); (2) identify the social costs of political correctness; and (3) endogenously account for the reputational concerns.

Formally, the analysis of this paper concerns a repeated cheap talk game, extending the framework of Sobel (1985) and Bénabou and Laroque (1992).[6] A 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. If the advisor is "good," she has the same preferences as the decision maker. If she is "bad," she always wants as high an action as possible. The state is realized (and publicly observed) after the decision maker's action is chosen. The decision

[6]Sobel (1985) introduced the tractable repeated cheap talk game with reputation studied in this paper. Bénabou and Laroque (1992) analyzed a version of Sobel's game where advisors have noisy signals. Both *assumed* that a good advisor tells the truth; they showed that a bad advisor (with opposing interests to the decision maker) will sometimes tell the truth (investing in reputation) and sometimes lie (exploiting that reputation). This paper endogenizes the behavior of the good advisor in Bénabou and Laroque's noisy advisor model. [There is also an important difference in the modelling of the bad advisor; see the discussion of the biased advisor assumption following proposition 3]. Just as the bad advisor sometimes 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.

maker updates his belief about the advisor given her message and *after* observing the true state of the world.

I first analyze what happens taking the advisors' reputational value functions as given. Because this is a cheap talk game, there always exist equilibria where there is babbling; that is, the advisor sends messages that are uncorrelated with her type and signal, and thus the decision maker learns nothing. Since the decision maker ignores the advisor's message in this case, the advisor has no incentive to change her strategy. The interesting question, then, is whether there exist equilibria where at least the good advisor truthfully reveals her information. In any such equilibrium, the bad advisor must be sending message 1 more often than the good advisor (if she sent message 1 less, she would have *both* a reputational and a current incentive to announce 1). Thus it turns out that in any non-babbling equilibrium, announcing 0 always increases the reputation of the advisor while announcing 1 always lowers it, *independent of the realized state.* In this environment, sending a message that turns out to be correct does not alter the direction of the inference. Using this strong characterization of the reputational effect, it is possible to show that if reputational concerns are sufficiently important to the good advisor, only babbling equilibria exist.

This result has a paradoxical element. By increasing the reputational concerns of the decision maker, we increase the incentive of the good advisor to separate from the bad advisor (holding fixed the incentive of the bad advisor to pool). In a standard costly signalling model, this increased incentive to separate would tend to favor the existence of *separating* equilibria. In this cheap talk model, it ensures the most complete form of *pooling* (i.e., babbling equilibrium). What happens is that increased reputational concerns provide an incentive for the good advisor to be more politically correct (i.e., announce 0 more often); this lowers the incentive of the bad advisor to say the politically incorrect thing (i.e., announce 1) since, given the good advisor's politically correct strategy, the reputational cost of announcing 1 has increased and she will not be believed anyway. When the good advisor's reputational concerns are big enough, the bad advisor loses all incentive to separate. Babbling equilibrium is the result. Incentives to separate by being politically correct are thus self-defeating.

The static model demonstrates how reputational concerns lead to the loss of socially valuable information. But in order to evaluate the welfare consequences of reputational concerns, it is necessary to have a model of why reputational concerns arise in the first place. The very existence of reputational concerns suggests that

someone must have a positive value for information about advisors' types. This positive value must be weighed against any loss of socially valuable information.

The simplest case with *instrumental* reputational concerns arises when advisors have no intrinsic reputational concerns and the cheap talk game is repeated twice. Advisors care about their reputation at the end of the first period only because they want to influence the decision maker in the second period. In this setting, it is possible to isolate three different welfare effects of allowing the decision maker to learn about the advisor's type in the first period. First, reputational concerns lead the bad advisor to offer less biased advice (the discipline effect). Second, the decision maker may learn about the type of the advisor from the first period game (the sorting effect). Both these effects suggest that the decision maker has an incentive to try and deduce the advisor's type from her first period advice. But, third, the good advisor may be deterred from offering sincere advice (the political correctness effect). This effect gives the decision maker an incentive not to use first period information in the second period (*if* he could so commit). It is shown that any effect could dominate, depending on the parameters.

I also consider an infinite horizon model. Focussing attention on Markov equilibria, it is possible to demonstrate how comparative statics results translate to the infinite horizon; and it is noted that even as the good advisor's discount rate approaches 1, truth-telling is always inconsistent with equilibrium after at least some histories.

This paper belongs to the literature on cheap talk games initiated by Crawford and Sobel (1982). As discussed above, it follows Sobel (1985) and Bénabou and Laroque (1992) in incorporating reputational concerns into that setting. The static game (taking the reputational value functions as given) can be understood as a cheap talk game with two dimensional types (the advisor's preference type and signal). Cheap talk games with multidimensional types are the subject of Austen-Smith (1993b) and Spector and Piketty (1997). In Austen-Smith (1992) and Austen-Smith (1995), as in this paper, two dimensional types consist of a preference type and a signal about policy (these types are partially revealed in equilibrium by a combination of cheap talk and costly actions).

Two themes of this paper are familiar from earlier work. First, Holmström and Ricart i Costa (1986) initiated a literature on perverse reputational incentives. Some of the more closely related papers from that literature are discussed later in this paper. Second, the problem of eliciting information from interested parties is the subject of a large literature, both under the cheap talk assumption and in

more general settings.[7] This literature deals with many important issues (such as multiple informed parties and optimal mechanism design) that are avoided in this analysis. This paper focuses on one particular problem in eliciting information: the perverse reputational incentives of a "good" advisor.

2. A Static Model of Political Correctness

In this section, I analyze the equilibrium behavior of the advisor, taking as given her reputational concerns. In the next section, one instrumental explanation for her reputational concerns is provided and analyzed.

2.1. The Static Game

I first provide an abstract description of the model. The reader may want to keep in mind the affirmative action example discussed in the introduction. After the description, three alternative interpretations are offered.

A decision maker's optimal decision depends on the state of the world $\omega \in \{0,1\}$. Each state occurs with equal probability.[8] 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 $\gamma \in \left(\frac{1}{2}, 1\right)$.

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 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

[7] Examples (in wide variety of analytic settings) include Austen-Smith (1993a), Banerjee and Somanathan (1997), Dewatripont and Tirole (1995), Glazer and Rubinstein (1997), Krishna and Morgan (1998), Ottaviani and Sorensen (1998) and Shin (1996).

[8] This assumption is made for notational convenience only: all results hold qualitatively with an asymmetric prior probability distribution on states.

$a^*(m) = \arg\max_{a \in \Re} u_{DM}(a, \omega)$ 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 $\Lambda(m, \omega)$ for the posterior probability that the advisor is good if she sends message m and state ω is realized. Then

$$\Lambda(m, \omega) = \frac{\lambda \phi_G(m|\omega)}{\lambda \phi_G(m|\omega) + (1-\lambda)\phi_B(m|\omega)}, \tag{2.1}$$

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 $\phi_I(0|\omega) = 1 - \phi_I(1|\omega)$.[9]

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

$$x u_{DM}(a, \omega) + v_G[\Lambda(m, \omega)],$$

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

$$y u_B(a) + v_B[\Lambda(m, \omega)],$$

where $y > 0$, u_B is a strictly increasing and continuous on the interval $[a^*(1-\gamma), a^*(\gamma)]$ and $v_B : [0,1] \to \Re$ is a strictly increasing continuous function.[10]

Write $\Gamma(m)$ for the DM's posterior belief that the actual state is 1 if message 1 is announced. By Bayes' rule,[11]

$$\Gamma(m) = \frac{\lambda \phi_G(m|1) + (1-\lambda)\phi_B(m|1)}{\lambda \phi_G(m|1) + (1-\lambda)\phi_B(m|1) + \lambda \phi_G(m|0) + (1-\lambda)\phi_B(m|0)}. \tag{2.2}$$

[9] $\Lambda(m, \omega)$ is well defined only if the denominator is non-zero. I adopt the convention that $\Lambda(m, \omega) = \lambda$ if $\sigma_G(m|1) = \sigma_G(m|0) = \sigma_B(m|1) = \sigma_B(m|0) = 0$. Allowing for other out-of-equilibrium beliefs does not lead to any different equilibrium behavior.

[10] An alternative interpretation would be that the bad advisor had the same preferences as the good advisor, but had an extreme prior where she assigned prior probability 1 (instead of $\frac{1}{2}$) to state 1. In this case, we would have $u_B(a) = u_{DM}(a, 1)$; this automatically satisfies the assumptions above. Banerjee and Somanathan (1997) examine the equilibrium credibility of advisors with such differences in priors (but without reputational concerns).

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

Now $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$ is an *equilibrium* if (1) the advisor's message given her signal maximizes her utility given the decision maker's strategy χ and the type inference function Λ; (2) the decision maker's action is optimal given the state inference function Γ; and (3) the type and state inference functions, Λ and Γ, are derived from the advisor's strategy according to inference rules (2.1) and (2.2).[12]

Three examples may help motivate the model:

1. The decision maker is a public official maximizing a social welfare function. Action a corresponds to a policy that creates transfers to a special interest. The socially optimal level of the policy depends on the state of the world. The public official is advised by an expert who certainly has some information about the state and cares about her reputation; her current objective may be to maximize social welfare (the "good advisor"); but she may be trying to maximize transfers to the special interest by maximizing the level of the policy (the "bad advisor").

2. The decision maker is a risk averse investor deciding how much to invest in a risky asset (a high value of a corresponds to a large investment). His financial advisor certainly has information about the likely performance of the asset and cares about her reputation; her current objective may be to maximize the expected utility of the investor (the "good advisor"); but she may be trying to off-load surplus stock of the asset (the "bad advisor").

3. The decision maker is a personnel officer allocating a salary budget between a male employee and a female employee. The personnel officer wants to allocate a larger share to the more productive employee. A high action a corresponds to a higher allocation to the male employee. The personnel officer is advised by a supervisor who certainly has information about which

[12] The value function is (for now) being taken as given, so this is not a standard game. However, we could think of the decision maker taking the action a before observing ω, and then taking a second action $\lambda \in [0,1]$ after observing ω, where the decision maker's optimal action is to set λ equal to her posterior probability that the advisor is good [this will be optimal if the decision maker's payoff is $-\lambda^2$ if the advisor is bad, and $-(1-\lambda)^2$ if the advisor is good]. The static game is thus a cheap talk game with two dimensional types: the preference type G or B; and the signal type, 0 or 1. Type $(G,0)$ would like to be perceived to be type $(G,0)$; type $(G,1)$ would like to be perceived to be type $(G,1)$; types $(B,0)$ and $(B,1)$ would both also like to be perceived to be type $(G,1)$. Notice that allowing the advisor to announce her preference type would not matter (she would always claim to be good).

employee is more productive and cares about his reputation; his current objective may be to reward the more productive employee (the "good advisor"); but he may be a sexist who wants to see the male employee rewarded independently of productivity (the "bad advisor").

2.2. Characterization of Equilibria

The decision maker's optimal action depends only on how likely he thinks the two states; indeed, the assumptions I made on the decision maker's preferences ensure that his optimal action is an increasing function of the probability he assigns to state 1.

Lemma 1. *In any equilibrium* $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$,

$$\chi(m) = \widetilde{a}(\Gamma(m))$$

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

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

The proof of the lemma, and all the propositions in the paper, are in the appendix. This lemma can be illustrated by a simple example. Suppose that the decision maker's preferred action 1 in state 1, 0 in state 0 and he has a quadratic loss function depending on the distance between the actual action and his preferred action, i.e., $u_{DM}(a, \omega) = -(a-\omega)^2$. In this case, $\widetilde{a}(q) = q$.

As in any model of cheap talk, there will always exist equilibria where the costless announcements are uninformative: as long as announcements are uninformative, the decision maker will ignore them; and as long as the decision maker ignores announcements, the advisor does not have any incentive to make them informative; see Crawford and Sobel (1982).

Definition 1. $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$ *is a babbling strategy profile if, for some* $c \in [0,1]$, $\sigma_G(0) = \sigma_B(0) = \sigma_G(1) = \sigma_B(1) = c$; $\chi(0) = \chi(1) = \widetilde{a}\left(\frac{1}{2}\right)$; $\Gamma(0) = \Gamma(1) = \frac{1}{2}$; $\Lambda(1,1) = \Lambda(0,1) = \Lambda(1,0) = \Lambda(0,0) = \lambda$.

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

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

Thus the interesting question is the existence and properties of non-babbling equilibria. In analyzing non-babbling equilibria, I will restrict attention to equilibria $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$ where message 1 is (weakly) correlated with state 1, i.e., $\Gamma(1) \geq \Gamma(0)$. This assumption is without loss of generality.

I start by analyzing the properties of "truth-telling" equilibria; that is, equilibria where the good advisor always tells the truth, i.e., $\sigma_G(0) = 0$ and $\sigma_G(1) = 1$. It will turn out that the intuition from this case translates to all non-babbling equilibria.

Assume then that the good advisor always told the truth (leaving aside for now the question of whether it is optimal to do so). What would the bad advisor's best response be? Note that the bad advisor must announce 1 strictly more (on average) than the good advisor. If not, announcing 1 would (in equilibrium) reduce (or at least not increase) the likelihood the advisor was good. But since announcing 1 maximizes the action of the decision maker, it would therefore be strictly optimal for the bad advisor to announce 1 (contradicting our premise that the bad advisor announced 1 no more than the good advisor). In fact, it can be shown that the bad advisor has a unique best response to the good advisor telling the truth, where the bad advisor always announces 1 if he observes signal 1, and announces 1 with some strictly positive probability if he observes signal 0, i.e., $\sigma_B(0) = \nu \in (0, 1]$ and $\sigma_B(1) = 1$.

We can write down explicitly what inferences will be drawn under these strategies, using equation (2.1). For example, the probability that the good advisor announces 1 if the true state is 1, $\phi_G(1|1)$, is γ, since she announces 1 only if she observes signal 1 and she observes signal 1 with probability γ if the true state is 1. The probability that the bad advisor announces 1 if the true state is 1, $\phi_B(1|1)$, is $\gamma + (1-\gamma)\nu$, since with probability γ she observes 1 and announces 1 for sure, and with probability $1-\gamma$ she observes 0 and announces 1 with probability ν. Now

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

$$= \frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\nu\right)}.$$

By similar computations,

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

Since $\nu > 0$, this implies in particular that

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

Thus each advisor has a strict reputational incentive to announce 0, and this is true independent of what state they expect to be realized. Even if an advisor somehow knew for sure that the true state would turn out to be 1, she would have a reputational incentive to announce 0.

Truth-telling equilibria are relatively simple to characterize. But there exist equilibria that are neither babbling nor truth-telling. The good advisor, on observing signal 1, may randomize between telling the truth (despite the reputational consequences) and lying (to enhance her reputation at the expense of her current utility). However, *all* non-babbling equilibria inherit three crucial properties of truth-telling equilibria.

Proposition 2. *Any non-babbling equilibrium* $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$ *satisfies the following three properties:*

1 *The good advisor always announces* 0 *when she observes signal* 0 ($\sigma_G(0) = 0$) *and announces* 1 *with positive probability when she observes signal* 1 ($\sigma_G(1) > 0$);

2 *Messages are strictly informative:* $\Gamma(1) > \Gamma(0)$ *and thus* $\chi(1) > \chi(0)$;

3 *There is a strict reputational incentive for the advisor to announce* 0; *more specifically,* $\Lambda(0,1) \geq \Lambda(0,0) > \lambda > \Lambda(1,1) \geq \Lambda(1,0)$.

Proposition 2 tells us that in *every* non-babbling equilibrium, both types of advisor have a strict reputational incentive to announce 0, whatever signal they observe. The bad advisor always has a strict current incentive to announce 1. The good advisor has a strict current incentive to tell the truth. Note that proposition 2 says nothing about the existence of non-babbling equilibria, and I will return to this question in the next proposition. Two other issues are discussed first. Does the bad advisor prefer babbling equilibria or non-babbling equilibria? And in what sense do these equilibria exhibit "political correctness"?

Since the (ex ante) probability of state 1 is $\frac{1}{2}$, the ex ante expected value of $\Gamma(m)$ must be $\frac{1}{2}$ *in any equilibrium* (this is a standard property of conditional probability). That is, the existence of the bad advisor can never bias the decision maker's beliefs systematically towards state 1. The possibility that the advisor is bad merely introduces noise in the decision making, and there is no reason to expect this to be to the bad advisor's advantage. Consider the case where $u_{DM}(a,\omega) = -(a-\omega)^2$, and thus (as I noted above) $\widetilde{a}(q) = q$. In this case, we have $\chi(0) = \Gamma(0)$ and $\chi(1) = \Gamma(1)$ in any equilibrium, and thus the expected action equal to $\frac{1}{2}$ in any equilibrium. Now suppose that $u_B(a) = a$: then the ex ante expected value of $u_B(\chi(m))$ is $\frac{1}{2}$ in any equilibrium. This means that someone with the bad advisor's current preferences (but not knowing the type of the advisor) is indifferent between all equilibria (her current utility is $\frac{1}{2}$ in any such equilibrium). This is true in the special case where $\widetilde{a}(\cdot)$ and $u_B(\cdot)$ are both linear functions. More generally, someone with the bad advisor's preferences prefers more uncertainty in the outcome (i.e., non-babbling to babbling equilibria) if $u_B(\widetilde{a}(\cdot))$ is concave and less uncertainty (i.e. babbling equilibria to non-babbling equilibria) if $u_B(\widetilde{a}(\cdot))$ is convex. Each of these properties is consistent with my assumptions, which imply only that $\widetilde{a}(\cdot)$ and $u_B(\cdot)$ are continuous strictly increasing functions.

Propositions 1 and 2 together characterize all possible equilibria. In what sense do they exhibit "political correctness"? All non-babbling equilibria have a strict reputational incentive to say the politically correct thing, i.e., announce 0. In truth-telling equilibria, this reputational incentive is not sufficient to elicit politically correct *behavior* from the good advisor. But in non-truth-telling, non-babbling equilibria, the good advisor sometimes says something insincere in order to enhance her reputation. She *behaves* in a politically correct way. In babbling equilibria, there are no reputational incentives and the advisor is indifferent between all actions. But sometimes there is no equilibrium other than babbling

equilibria, since any non-babbling strategy profile would lead to sufficiently large reputational inferences to make that strategy profile non-viable as an equilibrium. Thus there is a sense in which political correctness out of equilibrium implies the loss of information. We now turn to identifying when this is the case.

In order to analyze which kinds of equilibria exist, it is useful to think through what happens as x, the value of the current decision problem to the good advisor, is varied. If x is sufficiently large, the current gain to the good advisor of telling the truth will exceed the reputational cost, and a truth-telling equilibrium will exist. As x becomes smaller (holding fixed the good advisor's reputational value function), truth-telling equilibria will cease to exist although there may still exist non-babbling equilibria. As x declines even further, the overwhelming importance of reputational concerns will guarantee that only babbling equilibria exist. The following proposition formalizes this discussion by considering when the game parameterized by (λ, x, y) has different kinds of equilibria.

Proposition 3. *For any $\lambda \in (0,1)$ and $y \in \Re_{++}$, there exist $0 < \underline{x}(\lambda, y) \leq \overline{x}(\lambda, y)$ such that [1] if $x \leq \underline{x}(\lambda, y)$, all equilibria of the (λ, x, y) game are babbling; and [2] there exists a truth-telling equilibrium in the (λ, x, y) game if and only if $x \geq \overline{x}(\lambda, y)$.*[13]

Thus I have shown that politically correct inferences will occur in any non-babbling equilibrium (proposition 2); and that if the current decision problem is of sufficiently small importance relative to reputational concerns, the possibility of such politically correct inferences prevents the existence of any non-babbling equilibrium (proposition 3). I will conclude this section by discussing three key ingredients of the model.

CHEAP TALK. The good advisor would like to signal that she is good. There is (sometimes) a "costly" action that she can take to signal that she is good:

[13]The proof (in the appendix) gives explicit forms for $\overline{x}$ and $\underline{x}$ (equations 4.7 and 4.8 respectively); these can be used to show the following limiting properties. As $\lambda \to 1$, the reputational cost of any action goes to zero (with noisy signals, it is impossible to lose much reputation for λ close to 1); thus $\underline{x}(\lambda, y) \to 0$ and $\overline{x}(\lambda, y) \to 0$ as $\lambda \to 1$. As $\lambda \to 0$, and if the good advisor follows a truth-telling strategy, the reputational gain to lying and the current gain to telling the truth both tend to a constant, so $\overline{x}(\lambda, y)$ tends to some positive constant also. As $y \to 0$, the bad advisor's strategy will mimic the good advisor's strategy, so reputational concerns must become smaller; so $\underline{x}(\lambda, y) \to 0$ and $\overline{x}(\lambda, y) \to 0$ as $y \to 0$. Finally, if y is sufficiently large, the bad advisor will always announce 1 in any non-babbling equilibrium. Thus $\underline{x}(\lambda, y)$ and $\overline{x}(\lambda, y)$ become constant for all sufficiently large y.

announcing 0 when she has observed signal 1. The cost associated with this announcement is endogenous: it is costly to announce 0 because (in equilibrium) the decision maker is induced to choose a worse action. It is precisely because the costs of signalling preference type (good or bad) are endogenous that we get the paradoxical result that an increased incentive to separate makes separation impossible.

It is useful to compare the cheap talk model with (exogenously) costly signalling of preference type. Suppose that the advisor was able to directly choose an action (instead of advising the decision maker on an action). Under natural single crossing properties, choosing a sufficiently low action would separate the good advisor from the bad advisor. Thus if the good advisor were sufficiently concerned about her reputation, there would be equilibria where she separated out from the bad advisor by choosing sufficiently low ("politically correct") actions. But the advisor's information about the state would still not be revealed in those equilibria.[14]

THE NATURE OF THE BAD ADVISOR. The conflict of interest between the decision maker and the bad advisor in this paper takes an extreme form: the "bad" advisor is always biased in a particular direction. With more general conflicts of interest between the decision maker and the bad advisor, we would expect the reputational costs of truth-telling to be mitigated. It turns out that Sobel (1985) and Bénabou and Laroque (1992) analyzed another extreme case where the reputational cost of truth-telling may disappear altogether. Their bad advisor's preferences were the *opposite* of the decision maker's. That is, while the decision maker wanted to take action 1 in state 1 and action 0 in state 0, the bad advisor wanted him to take action 0 in state 1 and action 1 in state 0. In this case, if the good advisor always tells the truth (as they *assumed*), there is no reputational cost for the good advisor of telling the truth (at least with symmetric strategies). In other words, if the good advisor were given the preferences of the decision maker in their models and her behavior were endogenized, there would always exist a (symmetric) equilibrium where the good advisor always tells truth.

Another characteristic that advisors may want to signal is the quality of their observations. This can be modelled using the framework described above by having the bad advisor observe a signal that is less informative than the good advisor's. The conclusions are rather sensitive to the exact assumptions about the

[14]See Austen-Smith and Banks (1997) for more on the relation between cheap talk and costly signalling.

bad advisor's preferences. If the bad advisor is sufficiently concerned about her reputation, she will attempt to mimic the good advisor's strategy, e.g., announcing 1 about proportion γ of the time and announcing 0 about proportion $1-\gamma$ of the time (even if her information were worthless). In this circumstance, truth-telling *will* be a best response for the good advisor (however much she cares about her reputation). On the other hand, if, for some reason, the bad advisor announced her relatively uninformative signal truthfully (and in particular did not take into account the decision maker's prior belief on states), the good advisor would have a reputational incentive to lie in favor of ex ante likely signals, in order to sound more informed.[15] Scharfstein and Stein's (1990) model of reputational herding and Prendergast's (1993) model of "yes men" developed the idea that saying the expected thing is sometimes the best way of sounding smart.[16]

NOISY SIGNALS. If signals were perfect (i.e., $\gamma = 1$) and the good and bad advisors both cared a lot about their reputations (i.e., x and y were both small), then there would always exist a sequential equilibrium of the game where both advisors always tell the truth. This would be supported by the (out-of-equilibrium) belief that anyone whose announcement was not equal to the observed state is certainly the bad advisor. This equilibrium is the limit of equilibria for γ close to 1. Specifically, as long as y is not too large, one can verify (using expression (4.7) for $\overline{x}$ in the appendix) that as $\gamma \to 1$, $\overline{x} \to 0$, i.e., there is a truth-telling equilibrium for smaller and smaller values of x.

3. Instrumental Reputation in a Dynamic Game

The analysis of the previous section was independent of why the advisor values reputation. For example, she might directly care how she is perceived. But in

[15] This would happen if the assumption that each state were equally likely was relaxed.

[16] A number of other papers develop important insights about developing a reputation for being informed. Prendergast and Stole (1996) show how (with costly actions) competent individuals may signal their type by taking extreme actions (since competent individuals are more likely to have extreme posteriors). [Note that with cheap talk, competent individuals would not be able to separate in this way: see the above discussion of cheap talk]. Ottaviani and Sorensen (1998) explore how reputational concerns influence communication by a group of individuals, under alternative mechanisms (e.g., sequential versus simultaneous). Campbell (1997) considers a consultant whose equilibrium fee for advice depends only on her reputation; he shows that there are always some situations where the consultant's reputation for competence is not maximized by telling the truth; so some information is always lost.

economic models, such reputational concerns typically arise because reputation is instrumentally valuable in the future. For examples, employees may receive higher wages if they are perceived as having higher ability (e.g., Holmström and Ricart i Costa (1986)). Politicians may be more likely to be re-elected if they are perceived to have the interests of their constituents at heart. One could readily construct models combining the cheap talk of the previous section with such standard models of instrumental reputational concerns.

But in this section, I follow Sobel (1985) and Bénabou and Laroque (1992) in working with a simpler model of how reputational concerns arise. The advisor cares about her reputation not because others will treat her differently (e.g., pay her more, re-elect her), but simply because she wants her advice to be accepted in the future. In most environments, this is unlikely to be the *only* reason for reputational concerns. I nonetheless focus on this explanation because I want to emphasize how reputational concerns may impose constraints on communication *even among individuals whose only interaction is the communication they are engaged in.*

I perform two exercises in this section. First, I consider what happens if the cheap talk game of the previous section is repeated twice. This twice repeated game allows a simple demonstration of how reputational concerns arise endogenously; and it can be used to give a primative welfare analysis of the loss of information associated with political correctness. Second, I describe a stationary environment where the cheap talk game is repeated infinitely often. Strong restrictions on strategies are required to say anything about what happens in this environment. But under these restrictions, it is possible to demonstrate how the comparative statics described in proposition 3 arise in a stationary environment.

3.1. Welfare Analysis in the Twice Repeated Game

The game of the previous section is repeated twice (without discounting). Thus, as before, in period 1, an advisor is good with initial probability λ; a state ω_1 is drawn but not observed; the advisor receives a signal s_1 and sends a message m_1 to the decision maker; the decision maker chooses an action a_1; and the state ω_1 is publicly observed. The decision maker updates his beliefs about the advisor (given equilibrium strategies) to $\Lambda(m_1, \omega_1)$. The same scenario is repeated in period 2, with the advisor starting with this new reputation. Thus a new state ω_2 is (independently) drawn but not observed; the advisor receives a new signal s_2

and sends a new message m_2 to the decision maker; the decision maker chooses a new action a_2; and finally the state ω_2 is publicly observed.

I now assume that the advisors no longer have (intrinsic) reputational concerns. The good advisor's payoff in the two period game is *exactly* the decision maker's payoff, $x_1 u_{DM}(a_1, \omega_1) + x_2 u_{DM}(a_2, \omega_2)$, where x_1 and x_2 are parameters measuring the importance of each period's decision problem. Similarly, the bad advisor's utility is $y_1 u_B(a_1) + y_2 u_B(a_2)$. Thus the two period game is parameterized by the advisor's initial reputation λ, and the payoff parameters (x_1, y_1, x_2, y_2).

I solve by backward induction. Write λ' for the reputation of the advisor entering the second period; this is the only payoff relevant variable from the first period play. Because there are no reputational concerns in the second period, there is always a truth-telling equilibrium in the second period, which is in fact the unique non-babbling equilibrium. The good advisor must be telling the truth while the bad advisor must always announce 1. Given these strategies, if the decision maker receives message 0, he will assign probability $1-\gamma$ to state 1 and choose action $\widetilde{a}(1-\gamma)$; if he receives message 1, he will assign probability $\frac{\lambda'\gamma+(1-\lambda')}{\lambda'+(1-\lambda')} = \frac{1-\lambda'+\lambda'\gamma}{2-\lambda'}$ to state 1 and choose action $\widetilde{a}\left(\frac{1-\lambda'+\lambda'\gamma}{2-\lambda'}\right)$. Thus assuming non-babbling behavior in the second period,[17] we can derive the value function of reputation for both types of advisors entering the second period:

$$v_G(\lambda') = x_2 \left\{ \begin{array}{l} \frac{1}{2}\gamma u_{DM}\left(\widetilde{a}\left(\frac{1-\lambda'+\lambda'\gamma}{2-\lambda'}\right), 1\right) + \frac{1}{2}(1-\gamma) u_{DM}\left(\widetilde{a}\left(\frac{1-\lambda'+\lambda'\gamma}{2-\lambda'}\right), 0\right) \\ +\frac{1}{2}(1-\gamma) u_{DM}(\widetilde{a}(1-\gamma), 1) + \frac{1}{2}\gamma u_{DM}(\widetilde{a}(1-\gamma), 0) \end{array} \right\}$$

$$\text{and } v_B(\lambda') = y_2 u_B\left(\widetilde{a}\left(\frac{1-\lambda'+\lambda'\gamma}{2-\lambda'}\right)\right).$$

Both functions are continuous and strictly increasing in λ'. Now the structure of equilibrium strategies in period 1 follows exactly the analysis of the previous section. In particular, there exist $0 < \underline{x}_1(\lambda, y_1, x_2, y_2) \leq \overline{x}_1(\lambda, y_1, x_2, y_2)$ such that there exists a truth-telling equilibrium if $x_1 \geq \overline{x}_1(\lambda, y_1, x_2, y_2)$ and all equilibria are babbling if $x_1 \leq \underline{x}_1(\lambda, y_1, x_2, y_2)$.

Thus, for some parameters, if there is non-babbling behavior in the second period, there must be babbling in the first period, i.e., all the first period information

[17] If there was always babbling in the second period, there would be zero value for reputation in the first period, and thus the first period analysis would be identical to the second period analysis I just described. But it could also be the case that there was babbling after some histories and truth-telling after other histories. Such strategies could support a wide variety of first period behavior in equilibrium.

must be lost. This sounds undesirable. Formal welfare analysis is hard because of the multiplicity of equilibria, but I will describe one limited comparison. I would like to compare the equilibria just described with a situation where the decision maker is not able to update his beliefs in the light of the first period play. For example, suppose that in the second period, the decision maker was replaced with another decision maker with identical preferences who had not observed the first period message and state, and thus assigned the original probability λ to the advisor being good. Thus I will compare:

- *The No Updating Scenario.* The advisor has reputation λ entering *each* period. The good advisor tells the truth, the bad advisor always says 1 and the decision maker acts accordingly, in each period.

- *The Updating Scenario.* Whatever the advisor's reputation entering the second period, the good advisor tells the truth, the bad advisor always says 1 and the decision maker acts accordingly. There is equilibrium play in the first period.

I will focus on the utility of the decision maker under the two scenarios.[18] Allowing the decision maker to update his beliefs about the advisor following the first period play has three effects.

- The *Discipline Effect.* In the No Updating scenario, the bad advisor always announces 1 in the first period. Under the Updating scenario, the bad advisor may sometimes announce 0, in order to enhance her reputation, revealing valuable information. This is good for the decision maker.

- The *Sorting Effect.* In the Updating scenario, the decision maker learns about the bad advisor's type from first period play. Since the second period strategies are independent of the advisor's reputation entering that period, this must be valuable for the decision maker.

- The *Political Correctness Effect.* The decision maker's concern about the type of the advisor may provide incentives to the good advisor to lie in the first period; this is bad for the decision maker.

[18] Recall that the good advisor's utility is identical to the decision maker's and the bad advisor's utility could go either way depending on the shape of the function $u_B(\widetilde{a}(\cdot))$.

To take a more concrete example, suppose that the bad advisor was a racist. If the racist advisor offers less racist advice in order to appear less racist (the discipline effect), this is good for the decision maker; and if the decision maker receives more information about whether his advisor is racist (the sorting effect), this must be good for the decision maker too. But an unintended consequence of the decision maker's concern about his advisor's possible racism might be that the decision maker learns *neither* whether the advisor is in fact racist *nor* the valuable information that a non-racist advisor might otherwise have conveyed (the political correctness effect).

I can illustrate all three effects in my model. Consider first the case where, under the Updating scenario, there is a truth-telling equilibrium in the first period where the bad advisor sometimes announces 0; such an equilibrium exists if y_1 is sufficiently small (for any given λ, x_1, x_2 and y_2). In this case, in the first period, the good advisor tells the truth under both the No Updating and Updating scenarios, but the bad advisor reveals strictly more information under the Updating scenario. Thus there is no political correctness effect and there is a discipline effect in favor of the Updating scenario. In the second period, the decision maker has better information about the advisor's type (the sorting effect). Thus the Updating scenario is unambiguously better.

Now consider the case where $x_1 \leq \underline{x}_1(\lambda, y_1, x_2, y_2)$ and first period play, in the Updating scenario, *must* consist of a babbling equilibrium. Thus there will be no updating (in equilibrium) of the advisor's type (no sorting), and thus the second period will be identical under the Updating and No Updating scenarios. But in the first period, the decision maker received valuable information about the state under the No Updating scenario, but no information under the Updating scenario (the political correctness effect). Thus the No Updating scenario is unambiguously better.

3.2. Comparative Statics in a Stationary Infinite Horizon Model

There is some apparent tension in the conclusion of proposition 3: there must be babbling if reputational concerns are sufficiently high; but if there were always babbling, there could be no reputational concerns. In this section, I show how the tension can be resolved in a stationary infinite horizon version of the cheap talk game, as long as sufficient heterogeneity is permitted in decision problems across periods. In that case, we may have a constant value of reputation. Babbling then

must occur for sufficiently unimportant decision problems. But truth-telling may occur if the decision problem is relatively important.

Let the game of section 2 be repeated infinitely often, with 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 $(0,1)$. Thus the good advisor and the decision maker both receive total payoff $\sum_{t=0}^{\infty}(\delta_{DM})^t u_{DM}(a_t, \omega_t)$ and the bad advisor receives total payoff $\sum_{t=0}^{\infty}(\delta_B)^t u_B(a_t)$. 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 equilibrium if the value functions are continuous and strictly increasing.*

There will exist Markov equilibria with value functions that are continuous but *not* monotonic.[19] Nonetheless, I will focus on monotonic Markov equilibria. The idea here is just to verify that well-behaved reputational concerns are consistent with the infinite horizon model.

[19]Consider the following construction. Suppose the good advisor always told the truth. By a variation on an argument of Bénabou 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 close to 1, this best response will have the bad advisor's probability of lying

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 $x \leq x^*$,

$$\sigma_G(1; \lambda, x, y) = \sigma_G(0; \lambda, x, y) = \sigma_B(1; \lambda, x, y) = \sigma_B(0; \lambda, x, y).$$

The proposition and characterization hold independent of the discount rates, δ_{DM} and δ_B. Thus, in particular, even as $\delta_{DM} \to 1$, we continue to get babbling for sufficiently low values of x in any monotonic Markov equilibrium.[20] This observation tells us something curious about acquiring reputation in this setting. If it were common knowledge that the advisor were good, there would exist an equilibrium where the good advisor always told the truth. If it were not common knowledge that the advisor were good, but nonetheless the advisor always told the truth, then the advisor would (in expectation) have a reputation approaching 1.[21] Nonetheless, these results together do not imply that there is an equilibrium where the good advisor tells the truth, even if her discount rate is close to 1. The

increasing in her reputation (for some values of reputation). 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. If the bad advisor's probability of lying is increasing in his reputation sufficiently fast, the good advisor will prefer to have a lower reputation.

[20] But what would happen if there was no variation in x and y? Would there be an equilibrium with the good advisor telling truth if δ_{DM} were sufficiently close to 1? Notice that for δ_{DM} close to 1, the good advisor is more patient about achieving a reputation (reducing the incentive to lie). But the current cost of lying is also reduced. It can be shown that if x and y are constant, and for at least some discount rates of the bad advisor and decision maker utility functions, the latter effect is more important and there is no monotonic Markov equilibrium with truth-telling for δ_{DM} close to 1.

[21] Bénabou and Laroque (1992) showed this in their model, and the argument extends.

problem is that she would have an incentive to acquire a reputation for being good even faster than she could by always telling the truth.[22]

4. Conclusion

People care very much about what other people think of them; it is possible to explain much of their behavior by such concerns. In particular, anytime a speaker offers an opinion on any subject, the listener learns something about *both* that subject *and* the speaker. The possibility of such inferences influences what speakers say. The theory of this paper builds on such a view, but maintains the traditional economists' assumption that utility functions that do not depend on others' beliefs directly; if people care about what other people think of them, it is for *instrumental* reasons.[23]

I discussed a model where a speaker (advisor) communicates with the objective of conveying information, but the listener (decision maker) is initially unsure if the speaker is biased. There were three main insights from that model. First, in any informative equilibrium, certain statements will lower the reputation of the speaker *independent of whether they turn out to be true.* Second, if reputational concerns are sufficiently important, no information is conveyed in equilibrium. Third, while instrumental reputational concerns might arise for many reasons, a sufficient reason is that speakers wish to be listened to.

References

[1] Austen-Smith, D. (1992). "Explaining the Vote: Constituency Constraints on Sophisticated Voting," *American Journal of Political Science* **36**, 68-95.

[22]This may seem to contradict the lessons from the theoretical literature on reputation acquisition (e.g., Fudenberg and Levine (1992)). The difference is that, in this paper, the good advisor is trying to *separate* from the bad advisor, i.e., in the language of Mailath and Samuelson (1997), to demonstrate (correctly) who she is not. By contrast, results in that theoretical literature are driven by the option of *pooling* with a dominant strategy type.

[23]One advantage of such an approach is that it helps identify when reputational concerns matter most. For example, in a society where racial attitudes are well predicted by observable characteristics such as class, occupation and own race, communication on subjects to do with race will be relatively unfettered by reputational concerns. Sincere discussion of such subjects will be less likely in a society where attitudes are in flux and there is considerable uncertainty about individuals' racial attitudes based on observable characteristics.

[2] ——— (1993a). "Interested Experts and Policy Advice: Multiple Referrals under Open Rule," *Games and Economic Behavior* **5**, 3-43.

[3] ——— (1993b). "Information Acquisition and the Orthogonal Argument," in A. Barnett., M. Hinich and N. Schofield, *Political Economy: Institutions, Competition and Representation*. Cambridge University Press.

[4] ——— (1995). "Campaign Contributions and Access," *American Political Science Review* **89**, 566-580.

[5] ——— and J. Banks (1997). "Cheap Talk and Burned Money," Northwestern University and University of Rochester.

[6] Banerjee, A. and R. Somanathan (1997). "A Simple Model of Voice," M.I.T. and Emory University.

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

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

[9] Campbell, C. (1997). "Learning and the Market for Information," The Ohio State University.

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

[11] Dewatripont, M. and J. Tirole (1995). "Advocates," ECARE, ULB.

[12] Fudenberg, D. and D. Levine (1992). "Maintaining a Reputation when Strategies are Imperfectly Observed," *Review of Economic Studies* **59**, 561-579.

[13] Glazer, J. and A. Rubinstein (1996). "What Motives Should Guide Referees? On the Design of Mechanisms to Elicit Opinions," Tel Aviv University.

[14] Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Anchor Books, Doubleday.

[15] Holmström, B. and Ricart i Costa, J. (1986). "Managerial Incentives and Capital Management," *Quarterly Journal of Economics* **101**, 835-860.

[16] Krishna, V. and J. Morgan (1998). "A Model of Expertise," Penn State and Princeton.

[17] Kuran, T. (1995). *Private Lies, Public Truths.* Harvard University Press.

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

[19] Mailath, G. and L. Samuelson (1997). "Your Reputation is Who you're not, not Who you would like to be," Universities of Pennsylvania and Wisconsin.

[20] Ottaviani, M. and P. Sorensen (1998). "Information Aggregation in Debate," University College, London, and Nuffield College, Oxford.

[21] Prendergast, C. (1993). "A Theory of Yes Men," *American Economic Review* **83**, 757-770.

[22] ——— and L. Stole (1996). "Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning," *Journal of Political Economy* **104**, 1105-1134.

[23] Scharfstein, D. and J. Stein (1990). "Herd Behavior and Investment," *American Economic Review* **80**, 465-479.

[24] Shin, H. (1996). "Adversarial and Inquisitional Procedures in Arbitration," Nuffield College, Oxford; forthcoming in the *Rand Journal of Economics.*

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

[26] Spector, D. and T. Piketty (1997). "Rational Debate leads to One-Dimensional Conflict," MIT.

Appendix

Some preliminary notation and results will be useful. Write $\widehat{u}_G(q,s)$ for the expected value of u_{DM} for the good advisor if she 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:

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

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

Fact. $\widehat{u}_G(q,1)$ 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 $(\sigma_B,\sigma_G,\chi,\Gamma,\Lambda)$, 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.,

$$\begin{aligned} \Pi_G^C(s) &= x[\widehat{u}_G(\Gamma(1),s) - \widehat{u}_G(\Gamma(0),s)] \\ \text{and } \Pi_B^C(0) &= \Pi_B^C(1) = y[\widehat{u}_B(\Gamma(1)) - \widehat{u}_B(\Gamma(0))]. \end{aligned} \tag{4.1}$$

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.,

$$\begin{aligned} \Pi_I^R(1) &= \gamma \begin{bmatrix} v_I(\Lambda(0,1)) \\ -v_I(\Lambda(1,1)) \end{bmatrix} + (1-\gamma) \begin{bmatrix} v_I(\Lambda(0,0)) \\ -v_I(\Lambda(1,0)) \end{bmatrix} \\ \text{and } \Pi_I^R(0) &= (1-\gamma) \begin{bmatrix} v_I(\Lambda(0,1)) \\ -v_I(\Lambda(1,1)) \end{bmatrix} + \gamma \begin{bmatrix} v_I(\Lambda(0,0)) \\ -v_I(\Lambda(1,0)) \end{bmatrix}. \end{aligned} \tag{4.2}$$

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

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

This maximand is differentiable and strictly concave in a and thus uniquely achieves a maximum when

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

PROOF OF PROPOSITION 1. This is an immediate consequence of the definition of a babbling strategy profile. The message m sent by the advisor does not influence the decision maker's action ($\chi(m)$) or the decision maker's belief ($\Lambda(m,\omega)$). Thus the advisor is indifferent between all strategies including the uninformative one she uses in equilibrium. The advisor's strategy conveys no information, uniquely determining the decision maker's beliefs and optimal action. $\blacksquare$

PROOF OF PROPOSITION 2. This will be proved in nine steps. At each step, I demonstrate a property that must hold in any non-babbling equilibrium $(\sigma_G,\sigma_B,\chi,\Gamma,\Lambda)$. Recall that if $(\sigma_G,\sigma_B,\chi,\Gamma,\Lambda)$ is an equilibrium, $\chi(m)=\tilde{a}(\Gamma(m))$, and that we are assuming (without loss of generality) that $\Gamma(1)\geq\Gamma(0)$ and thus $\chi(1)\geq\chi(0)$.

P1. $\Lambda(0,1)\geq\Lambda(1,1)$ and $\Lambda(0,0)\geq\Lambda(1,0)$.

P1 asserts that there must always be a weak reputational incentive to announce 0. I show by contradiction that no equilibria exist if one of these conditions is violated.

- 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.

- Suppose that $\Lambda(1,1)>\Lambda(0,1)$ and $\Lambda(1,0)\leq\Lambda(0,0)$. By definition of Λ (see equation 2.1) we have

$$\gamma\sigma_G(1)+(1-\gamma)\sigma_G(0)=\phi_G(1\,|1)>\phi_B(1\,|1)=\gamma\sigma_B(1)+(1-\gamma)\sigma_B(0) \tag{4.3}$$

and $\gamma\sigma_G(0)+(1-\gamma)\sigma_G(1) = \phi_G(1|0) \leq \phi_B(1|0) = \gamma\sigma_B(0)+(1-\gamma)\sigma_B(1)$. (4.4)

Observe first that $\Pi_I^R(1) < \Pi_I^R(0)$ and $\Pi_I^C(1) \geq \Pi_I^C(0)$ for $I = B, G$ (by equations 4.1 and 4.2). Thus for both I, $\sigma_I(0) = 0$ or $\sigma_I(1) = 1$. This implies four subcases: (i) If $\sigma_G(0) = \sigma_B(0) = 0$, then (4.3) implies $\sigma_G(1) > \sigma_B(1)$, while (4.4) implies $\sigma_G(1) \leq \sigma_B(1)$, a contradiction; (ii) If $\sigma_G(0) = 0$ and $\sigma_B(1) = 1$, then (4.3) implies $\sigma_G(1) > 1$, a contradiction; (iii) If $\sigma_G(1) = 1$ and $\sigma_B(0) = 0$, then (4.4) implies $\sigma_B(1) = 1$ and $\sigma_G(0) = 0$, which implies $\phi_G(1|1) = \phi_B(1|1)$, contradicting (4.3); (iv) If $\sigma_G(1) = \sigma_B(1) = 1$, then (4.3) implies $\sigma_G(0) > \sigma_B(0)$, while (4.4) implies $\sigma_G(0) \leq \sigma_B(0)$, a contradiction.

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

$$\gamma\sigma_G(1) + (1-\gamma)\sigma_G(0) = \phi_G(1|1) \leq \phi_B(1|1) = \gamma\sigma_B(1) + (1-\gamma)\sigma_B(0) \tag{4.5}$$

and $\gamma\sigma_G(0)+(1-\gamma)\sigma_G(1) = \phi_G(1|0) \leq \phi_B(1|0) = \gamma\sigma_B(0)+(1-\gamma)\sigma_B(1)$. (4.6)

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 (4.5) and (4.6), this implies $\phi_G(1|1) < \phi_G(1|0)$. But now $\Gamma(1) < \frac{1}{2} < \Gamma(0)$, a contradiction.

P2. $\Lambda(0,1) \geq \Lambda(1,1)$ and $\Lambda(0,0) \geq \Lambda(1,0)$; and at least one these inequalities is strict.

P2 asserts that there must always be a *strict* reputational incentive to announce 0. The inequalities hold by **P1**. Suppose both held with equality. Recall that $\chi(1) \geq \chi(0)$ by assumption. If $\chi(1) > \chi(0)$, the bad advisor would have a strict incentive to choose 1 (whatever her signal), leading to a contradiction. But if $\chi(1) = \chi(0)$, we have a babbling equilibrium.

P3. $\chi(1) > \chi(0)$.

If $\chi(1) = \chi(0)$, then (by **P2**) the bad advisor would have a strict incentive to choose 0 (whatever his signal), leading again to a contradiction.

P4. $\sigma_G(0) = 0$.

By **P2**, $\Pi_G^R(0) > 0$; by **P3**, $\Pi_G^C(0) < 0$; so $\sigma_G(0) = 0$.

P5. $\Lambda(1,1) \geq \Lambda(1,0)$.

By the definition of Λ (equation 2.1) and **P4**,

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

P6. $\Lambda(0,1) \geq \Lambda(0,0)$.

Suppose not, i.e., $\Lambda(0,0) > \Lambda(0,1)$. Then we would have $\Lambda(0,0) > \Lambda(0,1) \geq \Lambda(1,1) \geq \Lambda(1,0)$. Now $\Pi_B^R(0) > \Pi_B^R(1)$, so $\Pi_B^R(1) > 0 \Rightarrow \Pi_B^R(0) > 0$; so either $\sigma_B(0) = 0$ or $\sigma_B(1) = 1$. But $\Lambda(0,0) > \Lambda(0,1)$ implies that $\frac{\phi_B(0|0)}{\phi_G(0|0)} < \frac{\phi_B(0|1)}{\phi_G(0|1)}$, i.e., $\frac{\phi_B(0|0)}{\phi_B(0|1)} < \frac{\phi_G(0|0)}{\phi_G(0|1)}$. 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 $\frac{\phi_G(0|0)}{\phi_G(0|1)}$ only if $\sigma_B(1) < \sigma_G(1)$. But this implies $\phi_B(1|0) < \phi_G(1|0)$, contradicting $\Lambda(0,0) > \Lambda(1,0)$. 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 cannot be less than $\frac{\phi_G(0|0)}{\phi_G(0|1)}$.

P7. For each $\omega \in \{0,1\}$, *either* $\Lambda(0,\omega) > \lambda > \Lambda(1,\omega)$ *or* $\Lambda(0,\omega) = \lambda = \Lambda(1,\omega)$.

We have $\Lambda(0,\omega) \geq \Lambda(1,\omega)$ from **P1**. Then **P7** follows from the definition of Λ (equation 2.1).

P8. $\Lambda(0,1) \geq \Lambda(0,0) > \lambda > \Lambda(1,1) \geq \Lambda(1,0)$.

We have established that, by **P1** and **P6**, (a) $\Lambda(0,1) \geq \Lambda(0,0) \geq \Lambda(1,0)$; by **P1** and **P5**, (b) $\Lambda(0,1) \geq \Lambda(1,1) \geq \Lambda(1,0)$. Now if $\Lambda(0,0) = \Lambda(1,0)$, then (by **P7**) $\Lambda(0,0) = \Lambda(1,0) = \lambda$; so by (b) and **P7**, $\Lambda(1,1) = \lambda = \Lambda(0,1)$, contradicting **P2**. But if $\Lambda(0,1) = \Lambda(1,1)$, then (by **P7**) $\Lambda(0,1) = \Lambda(1,1) = \lambda$; so by (a) and **P7**, $\Lambda(0,0) = \lambda = \Lambda(1,0)$, again contradicting **P2**. Thus $\Lambda(0,0) > \lambda > \Lambda(1,0)$ and $\Lambda(0,1) > \lambda > \Lambda(1,1)$. These two inequalities, with (a) and (b), show **P8**.

P9. $\sigma_G(1) > 0$.

Suppose $\sigma_G(1) = 0$. To have $\Gamma(1) > \Gamma(0)$, we must have $\sigma_B(1) > \sigma_B(0)$. These properties imply $\Lambda(0,1) > \Lambda(0,0) > \lambda$; $\Lambda(1,1) = \Lambda(1,0) = 0$. Thus $\Pi_B^R(1) > \Pi_B^R(0)$ and so $\sigma_B(1) \leq \sigma_B(0)$, a contradiction.

Now Part [1] of proposition 2 is proved by **P4** and **P9**. Part [2] is proved by **P3**. Part [3] is proved by **P8**. ■

PROOF OF PROPOSITION 3.

[1] *TRUTH-TELLING.* Suppose $\sigma_G(0) = 0$ and $\sigma_G(1) = 1$; to have $\Lambda(0,1) \geq \Lambda(0,0)$, must have $\sigma_B(1) = 1$; but $\sigma_B(0) = 0$ gives a contradiction. So we must have $\sigma_G(0) = 0$, $\sigma_G(1) = 1$, $\sigma_B(0) = \nu$ for some $\nu > 0$, $\sigma_B(1) = 1$ and $\chi(\cdot) = \widetilde{a}(\Gamma(\cdot))$. 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 net 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 strictly decreasing in ν, since each term is weakly decreasing in ν, and some are strictly decreasing. Also $g(0) = y\left(\widehat{u}_B(\gamma) - \widehat{u}_B(1-\gamma)\right) > 0$. Thus there exists exactly one value of ν where either $g(\nu) = 0$ or $\nu = 1$ and $g(\nu) > 0$. This ν parameterizes the unique equilibrium. Write $\widetilde{\nu}(\lambda, y)$ for that unique value of ν (for given λ and y).

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 $\chi(\cdot) = \widetilde{a}(\Gamma(\cdot))$. 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., $x \geq \overline{x}(\lambda, y)$, where $\overline{x}(\lambda, y)$ equals

$$\frac{v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1-\widetilde{\nu}(\lambda,y)\right)}\right] - \gamma v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{\gamma}{1-\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right] - (1-\gamma)\, v_G\left[\frac{1}{1+\left(\frac{1-\lambda}{\lambda}\right)\left(1+\left(\frac{1-\gamma}{\gamma}\right)\widetilde{\nu}(\lambda,y)\right)}\right]}{\left[\widehat{u}_G\left(\frac{\gamma+(1-\lambda)(1-\gamma)\widetilde{\nu}(\lambda,y)}{1+(1-\lambda)\widetilde{\nu}(\lambda,y)}, 1\right) - \widehat{u}_G(1-\gamma, 1)\right]}. \tag{4.7}$$

[2] *BABBLING.* The idea of the proof is to show that if x is very small and the equilibrium is non-babbling, the reputational gain (for the good advisor) to announcing 0 must be very small. This implies that the good advisor and bad advisor must be following similar strategies. This in turn implies (i) that the bad advisor does not always announce 1; (ii) $\Gamma(1)$ is much bigger than $\frac{1}{2}$ while $\Gamma(0)$ is no more than $\frac{1}{2}$; and (iii) the reputational gain (to the bad advisor) to announcing 0 must be small. Now (ii) and (iii) imply that the bad advisor always has a strict incentive to announce 1, contradicting (i).

Much notation is needed to make this argument formally. Let

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

and let $h(\lambda,\kappa)$ be the unique value of δ solving

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

if $\kappa < v_B(1) - v_B(0)$; if $\kappa \geq v_B(1) - v_B(0)$, let $h(\lambda,\kappa) = \infty$. Recall that by proposition 2, we have $\frac{\phi_B(0|\omega)}{\phi_G(0|\omega)} \leq 1 \leq \frac{\phi_B(1|\omega)}{\phi_G(1|\omega)}$ in any equilibrium; say that ϕ_G and ϕ_B are δ-close if for each $\omega \in \{0,1\}$,

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

I will show:

1. If $\Pi_G^R(1) < f(\lambda,\delta)$, then ϕ_G and ϕ_B are δ-close.
2. If ϕ_B and ϕ_G are $\left(\frac{1-\gamma}{2\gamma}\right)$-close, then $\sigma_B(0) < 1$ or $\sigma_B(1) < 1$.
3. If ϕ_B and ϕ_G are $\left(\frac{2\gamma-1}{2(1-\gamma)}\right)$-close, then $\Gamma(1) \geq \frac{\gamma}{\gamma+\frac{1}{2}}$ and $\Gamma(0) \leq \frac{1}{2}$.
4. If ϕ_B and ϕ_G are $h(\lambda,\kappa)$-close, then $\Pi_B^R(s) \leq \kappa$ for $s = 0,1$.

To prove (1), 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

$$\Pi_G^R(1) = \gamma\begin{bmatrix} v_G(\Lambda(0,1)) \\ -v_G(\Lambda(1,1)) \end{bmatrix} + (1-\gamma)\begin{bmatrix} v_G(\Lambda(0,0)) \\ -v_G(\Lambda(1,0)) \end{bmatrix} > f(\lambda,\delta).$$

To prove (2), recall that $\sigma_G(0) = 0$, so $\phi_G(1|1) \leq \gamma$, so if ϕ_B and ϕ_G are $\left(\frac{1-\gamma}{\gamma}\right)$-close, then $\phi_B(1|1) \leq \gamma\left(1+\frac{1-\gamma}{2\gamma}\right) < 1$.

To prove (3), note that if ϕ_G and ϕ_B are $\left(\frac{2\gamma-1}{2(1-\gamma)}\right)$-close, then

$$\begin{aligned}
\phi_B(1|0) &\leq \left(1+\frac{2\gamma-1}{2(1-\gamma)}\right)\phi_G(1|0) \\
&= \frac{1}{2(1-\gamma)}\phi_G(1|0) \\
&= \frac{1}{2(1-\gamma)}(1-\gamma)\sigma_G(1) \\
&= \frac{\sigma_G(1)}{2}
\end{aligned}$$

and $\varphi_B(1|1) \geq \phi_G(1|1) = \gamma\sigma_G(1)$; so

$$\begin{aligned}\Gamma(1) &= \frac{\lambda\phi_G(1|1) + (1-\lambda)\phi_B(1|1)}{\lambda\phi_G(1|1) + (1-\lambda)\phi_B(1|1) + \lambda\phi_G(1|0) + (1-\lambda)\phi_B(1|0)} \\ &\geq \frac{\gamma\sigma_G(1)}{\gamma\sigma_G(1) + \frac{\sigma_G(1)}{2}} \\ &= \frac{\gamma}{\gamma + \frac{1}{2}}.\end{aligned}$$

Now $\Gamma(1) > \frac{1}{2} \Rightarrow \Gamma(0) < \frac{1}{2}$.

To prove (4), observe that if ϕ_B and ϕ_G are $h(\lambda,\kappa)$ close, then (by construction of h) $v_B(\Lambda(0,1)) - v_B(\Lambda(1,1)) \leq \kappa$ and $v_B(\Lambda(0,0)) - v_B(\Lambda(1,0)) \leq \kappa$. Thus

$$\begin{aligned}\Pi_B^R(1) &= \gamma\begin{bmatrix} v_B(\Lambda(0,1)) \\ -v_B(\Lambda(1,1)) \end{bmatrix} + (1-\gamma)\begin{bmatrix} v_B(\Lambda(0,0)) \\ -v_B(\Lambda(1,0)) \end{bmatrix} \leq \kappa \\ \text{and } \Pi_B^R(0) &= (1-\gamma)\begin{bmatrix} v_B(\Lambda(0,1)) \\ -v_B(\Lambda(1,1)) \end{bmatrix} + \gamma\begin{bmatrix} v_B(\Lambda(0,0)) \\ -v_B(\Lambda(1,0)) \end{bmatrix} \leq \kappa.\end{aligned}$$

Now let

$$\underline{x}(\lambda,y) = \frac{f\left(\lambda, \min\left\{\frac{1-\gamma}{2\gamma}, \frac{2\gamma-1}{2(1-\gamma)}, h\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right)\right\}\right)}{\widehat{u}_G(\gamma,1) - \widehat{u}_G(1-\gamma,1)}. \tag{4.8}$$

Suppose that $x \leq \underline{x}(\lambda,y)$; in any non-babbling equilibrium,

$$\Pi_G^R(1) \leq \Pi_G^C(1) \leq x\left[\widehat{u}_G(\gamma,1) - \widehat{u}_G(1-\gamma,1)\right].$$

So

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

By (1), φ_G and ϕ_B are δ-close, where

$$\delta = \min\left\{\frac{1-\gamma}{2\gamma}, \frac{2\gamma-1}{2(1-\gamma)}, h\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right)\right\}.$$

Since $\delta \leq \frac{1-\gamma}{2\gamma}$, (2) implies **(A)** either $\sigma_B(0) < 1$ or $\sigma_B(1) < 1$. Since $\delta \leq \frac{2\gamma-1}{2(1-\gamma)}$, (3) implies **(B)** $\Gamma(1) \geq \frac{\gamma}{\gamma+\frac{1}{2}}$ and $\Gamma(0) \leq \frac{1}{2}$. Since $\delta \leq h\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right)$.

(4) implies **(C)** $\Pi_B^R(s) \leq \frac{1}{2} y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)$ for each $s = 0, 1$. But **(B)** and **(C)** imply that for each $s \in \{0.1\}$,

$$\begin{aligned}\Pi_B^C(s) &\geq y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right) \\ &> \frac{1}{2} y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right) \\ &\geq \Pi_B^R(s).\end{aligned}$$

Thus the bad advisor has a strict incentive to announce 1 whatever signal she observes. But this contradicts **(A)**. ■

PROOF OF PROPOSITION 4. Fix (x^*, y^*), let $\varepsilon = \phi(x^*, y^*)$, write

$$\overline{x} = \left(\frac{1}{1-\varepsilon}\right) \sum_{(x,y)\neq(x^*,y^*)} x.\phi(x,y) \text{ and } \overline{y} = \left(\frac{1}{1-\varepsilon}\right) \sum_{(x,y)\neq(x^*,y^*)} y.\phi(x,y)$$

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

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