

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1242

Note: Cowles Foundation Discussion Papers are preliminary materials circulated to stimulate discussion and critical comment. Requests for single copies of a Paper will be filled by the Cowles Foundation within the limits of the supply. References in publications to Discussion Papers (other than mere acknowledgment by a writer that he has access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

POLITICAL CORRECTNESS

Stephen Morris

December 1999

*Political Correctness**

Stephen Morris[†]
Yale University

September 1997
Revised November 1999

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. If the advisor is sufficiently concerned about her reputation, no information is conveyed in equilibrium. 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, Texas A&M, Warwick and Western Ontario; and from valuable conversations with David Austen-Smith, Stephen Coate, George Mailath, Andrew Postlewaite, the editor (Sherwin Rosen) and two anonymous referees. I gratefully acknowledge financial support from the Alfred P. Sloan Foundation.

[†]Cowles Foundation, Yale University, P.O.Box 208281, New Haven, CT 06520-8281. Electronic Mail: stephen.morris@yale.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 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.

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 advice 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 "advisor" 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. If reputational concerns are sufficiently important relative to the current decision problem, no information is conveyed in equilibrium. 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*. In this paper, "political correctness" refers to 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. There is an innocuous 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; this version will have few welfare consequences, as only the labelling of signals changes, not the information conveyed. 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 of this paper, 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. Loury summarizes his argument in the following syllogism (p. 437):

1. [(a)]within a given community the people who are most faithful to 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.

Loury does not present a formal model, but he notes that the theory of conformity of Bernheim (1994) could be adapted for the purpose. The explanation of this paper is narrower in scope but less “reduced form” than Loury’s. The 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. A state of the world, 0 or 1, is realized. An advisor observes a noisy signal of that state and may (costlessly) announce that signal to a decision maker. A 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 maker updates his belief about the advisor given her message and *after* observing the true state of the world. Then the game is played again, with the same advisor but a new state, signal, message and action.

Because this is a cheap talk game, there always exists a *babbling* equilibrium; that is, there is an equilibrium where 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 is whether there exist *informative* (non-babbling) equilibria where the decision maker learns something from the messages.

The game can be solved by backward induction. In the last period, the advisor will not be concerned about her reputation. So in any informative equilibrium, the good advisor will tell the truth and the bad advisor will always claim to have observed signal 1, and the decision maker will attach more significance to receiving message 1 the more confident he is that the advisor is good. This being the case, both advisors will have a strictly increasing value function for reputation entering the last period.

Now consider what happens in the first period. In any informative 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 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 (although it may alter the size of the change in reputation). Using this strong characterization of the reputational effect, it is possible to show that if reputational concerns are sufficiently important to the good advisor, no informative equilibrium exists.

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.

Reputational concerns sometimes guarantee the loss of socially valuable information. But reputational concerns themselves presumably serve some social purpose and any welfare losses associated with political correctness must be set against the benefits of reputational concerns. In this paper, the reputational concerns arise simply from a desire to transmit socially valuable information in the future. One advantage of endogenously accounting for the reputational concerns is that it is possible to carry out at least a crude welfare analysis. In particular, it is possible to distinguish three different effects of allowing the decision maker to learn about the type of the advisor 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). Any effect could dominate, depending on the parameters.

This paper belongs to the literature on cheap talk games initiated by Crawford and Sobel (1982). 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 2). 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.

Two themes of this paper are familiar from earlier work. First, Holmström (1999) and Holmström and Ricart i Costa (1986) initiated a literature on perverse reputational incentives. Scharfstein and Stein (1990) noted that if managers are concerned about their

reputation for being smart (i.e., observing accurate signals), then they will sometimes have a reputational incentive to say the expected thing, which may lead to information loss. Prendergast (1993), Prendergast and Stole (1996), Ottaviani and Sorensen (1998, 1999), Campbell (1998) and Levy (1998) further explore these issues. The preference-based reputational concerns of this paper similarly lead to information loss, although the mechanisms are rather different. 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. Examples (in wide variety of analytic settings) include Austen-Smith (1993a), Banerjee and Somanathan (1997), Brandenburger and Polak (1996), Dewatripont and Tirole (1999), Glazer and Rubinstein (1997), Krishna and Morgan (1998) and Shin (1998). That literature deals with many important issues (such as multiple informed parties and optimal mechanism design) that are ignored in this analysis. This paper focuses on one particular problem in eliciting information: the perverse reputational incentives of a “good” advisor.

2 The Two Period Advice Game

In the first period, a decision maker’s optimal decision depends on the state of the world $\omega_1 \in \{0, 1\}$. Each state occurs with equal probability. The decision maker has no information about the state but he has access to an advisor who is partially informed about the state of the world. The advisor observes a signal $s_1 \in \{0, 1\}$; with probability γ , this signal is equal to the true state; with probability $1 - \gamma$, the advisor is misinformed about the state. It is assumed the signal is informative, but not perfectly so, i.e., that $1/2 < \gamma < 1$. The decision maker is uncertain about the objectives of the advisor. Specifically, with probability λ_1 , the advisor is “good,” with utility function identical to the decision maker’s. With probability $1 - \lambda_1$, the advisor is “bad,” meaning that she is biased and always wants him to take the same decision (independent of her information). The advisor has an opportunity to announce her message m_1 (0 or 1), as a function of the signal she has observed. The decision maker will interpret the message he receives in the light of his uncertainty about the type of the advisor. Given the advisor’s message, the decision maker must choose an action $a_1 \in \mathbf{R}$. After the action is chosen, the state of the world ω_1 is publicly observed. The decision maker then rationally updates his belief about the type of the advisor, as a function of the initial reputation λ_1 , the message sent m_1 and the realized state ω_1 ; the advisor’s reputation entering the second period is written as $\lambda_2 = \Lambda(\lambda_1, m_1, \omega_1)$. The second period is identical to the first period, with a new (and independent) state ω_2 , a new noisy signal s_2 , a new message m_2 sent by the advisor and a new action a_2 chosen by the decision maker.

The decision maker’s utility in each period depends on the state of the world ω and his choice of action a . For simplicity, his utility is assumed to be given by the quadratic loss function $-(a - \omega)^2$. This implies that if the decision maker is uncertain about the state ω , his optimal action is to set a equal to the probability he assigns the possibility that $\omega = 1$. It is assumed that the decision maker may put different weights on period 1 and period 2 decisions. Thus total utility of the decision maker is given by

$$-x_1(a_1 - \omega_1)^2 - x_2(a_2 - \omega_2)^2$$

where $x_1 > 0$ and $x_2 > 0$. The good advisor is assumed to have *identical* preferences to the decision maker. The bad advisor always wants a higher action chosen, independent of the state. For simplicity, her utility in each period is taken to be simply the action a . She

too may weight the two periods differently, so her total utility in the two period game is

$$y_1 a_1 + y_2 a_2$$

where $y_1 > 0$ and $y_2 > 0$.

It is useful to keep in mind a number of interpretations of the model:

1. The decision maker is a public official maximizing a social welfare function. He is designing a policy that inevitably 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. 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. The personnel officer is advised by a supervisor who certainly has information about which 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”).
4. The decision maker is an editor of a journal who must decide on a response to a submitted paper (I am grateful to an anonymous referee for suggesting this example). The editor would like to give a more positive response, the higher the quality of the paper. He is advised by a referee, who is better able to assess the quality of the paper. The editor is uncertain whether the referee is similarly interested in rewarding quality (the “good advisor”) or if she has some ideological or other bias in favor of the paper (the “bad advisor”).

This game can be solved by backward induction.

Equilibrium in the Second Period Game (Without Reputational Concerns)

The advisor will enter the second period with a commonly known reputation λ_2 . Since the second period is the last period, the advisor (whether good or bad) will have no incentive to protect her reputation and will simply seek to achieve her current objective.

This game is an example of a *cheap talk* game (see Crawford and Sobel (1982)). The advisor’s action (her message) does not directly impact any player’s payoffs, but only indirectly influences payoffs via its impact on the beliefs of the decision maker about the state. In this sense, her action has no cost and is thus cheap talk. In any model of cheap talk, there exist equilibria where the cheap talk is ignored. If players observing cheap talk do not infer any meaning in the messages, then there is no incentive for those sending

the messages to imbue them with any meaning. Thus if advisor in the second period, independent of whether she is good or bad and independent of the signal she has observed, simply randomizes 50-50 between announcing 0 and announcing 1, the decision maker will learn nothing from the message and will continue to believe that each state is equally likely (and thus choose action $\frac{1}{2}$). Given this anticipated response by the decision maker, the advisor has no incentive to deviate from his uninformative random announcements. Such equilibria where cheap talk is ignored are known in the game theory literature as “babbling equilibria.” They exist because there is nothing in the logic of equilibrium behavior that guarantees that costless actions (cheap talk) convey meaning. The interesting question, in all cheap talk models, is when do there exist equilibria where cheap talk does convey meaning.

There will always exist a unique informative (i.e., non-babbling) equilibrium in the second period of the game. Suppose that the decision maker learns something from the message he receives and chooses a higher action after one message (say, message 1). Then bad advisor will have a strict incentive to announce 1 (independent of the signal she has observed), while the good advisor will have a strict incentive to announce her signal truthfully (since the decision maker will choose a strictly higher action if she announce 1 than if she announces 0). The advisor’s strategy may be summarized by the following table:

| m_2 | $s_2 = 0$ | $s_2 = 1$ |
|--------------|-----------|-----------|
| Good Advisor | 0 | 1 |
| Bad Advisor | 1 | 1 |

Given the advisor’s strategy, what inferences will the decision maker draw about the state of the world? If the decision maker receives message 0, he will be sure that the advisor is good and is truthfully reporting her signal. Thus he will assign probability $1 - \gamma$ to state 1 and choose action $1 - \gamma$. If he receives message 1, he will be unsure whether the advisor is bad (in which case the announcement conveys no information) or good (in which case the state is 1 with probability γ). By Bayes rule, he will assign probability

$$\frac{\frac{1}{2}(\lambda_2\gamma + (1 - \lambda_2))}{\frac{1}{2}(\lambda_2\gamma + (1 - \lambda_2)) + \frac{1}{2}((\lambda_2(1 - \gamma) + (1 - \lambda_2)))} = \frac{1 - \lambda_2 + \lambda_2\gamma}{2 - \lambda_2}$$

to state 1 and choose action $\frac{1 - \lambda_2 + \lambda_2\gamma}{2 - \lambda_2}$. Thus his action will be increasing in λ_2 , the reputation of the advisor. Now the value function for reputation for both types of advisors entering the second period can be derived:

$$v_G[\lambda_2] = -x_2 \left(\frac{1}{2}\gamma \left(\frac{1 - \lambda_2\gamma}{2 - \lambda_2} \right)^2 + \frac{1}{2}(1 - \gamma) \left(\frac{1 - \lambda_2 + \lambda_2\gamma}{2 - \lambda_2} \right)^2 \right) \quad (1)$$

$$\text{and } v_B[\lambda_2] = y_2 \left(\frac{1 - \lambda_2 + \lambda_2\gamma}{2 - \lambda_2} \right). \quad (2)$$

Both value functions are continuous and strictly increasing in λ_2 .

In the analysis that follows, it is assumed that the informative equilibrium giving rise to these value functions is played in the second period. If the babbling equilibrium were played in the second period, then there would be no reputational concerns in the first period.

Equilibrium in the First Period Game (With Reputational Concerns)

The first period game is the same as the second period game except that now the advisor has reputational concerns arising from the second stage of the game. In particular, the good advisor's payoff in the first period is given by

$$-x_1(a_1 - \omega_1)^2 + v_G[\Lambda(\lambda_1, m_1, \omega_1)]$$

while the bad advisor's payoff is given by

$$y_1 a_1 + v_B[\Lambda(\lambda_1, m_1, \omega_1)]$$

where $\Lambda(\lambda_1, m_1, \omega_1)$ is the equilibrium posterior probability assigned to the advisor being good. Once again, there will be a babbling equilibrium of the first period game: if the advisor randomized between messages independently of the signals, the decision maker would learn *neither* about the state of the world *nor* the type of the advisor, and again the advisor would have no incentive to send informative messages. The purpose of the following analysis is to characterize informative equilibria and to identify when they exist.

It is useful to focus the discussion on the nature and existence of equilibria where the good advisor always truthfully reports her signal. This case is relatively easy to analyze and provides accurate intuition concerning all possible equilibria (Appendix A provides a more formal treatment of the remaining material in this section). The argument is structured as follows. It is first assumed that there exists an equilibrium where the good advisor always tells the truth. Then it is possible to characterize how the bad advisor must be behaving in such an equilibrium. This in turn implies certain reputational incentives for the good advisor. Now it is possible to check for which parameters the strategy first proposed for the good advisor (telling the truth) is indeed optimal.

Suppose that the good advisor always told the truth. Would it be a best response for the bad advisor to also always tells the truth? In this case, there would be perfect pooling, and the decision maker would not update his beliefs about the advisor's type based on the announcement and realized state. But recall that the bad advisor would like to convince the decision maker that she has observed signal 1; if there were no reputational cost of announcing 1, she would have an incentive to always announce 1, contradicting our earlier assumption that she tells the truth. Thus the bad advisor cannot always tell the truth. By a similar logic, it is clear 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). More precisely, it can be shown that bad advisor always announces 1 if she observes signal 1, and announces 1 with some strictly positive probability ν if he observes signal 0 (we will describe how ν is determined below). This strategy can be summarized in the following table:

| | | |
|--------------|--|-----------|
| | $s_2 = 0$ | $s_2 = 1$ |
| Good Advisor | 0 | 1 |
| Bad Advisor | 0, with probability $1 - \nu$ 1, with probability ν | 1 |

Now the decision maker's inferences under such strategy can be derived by Bayes rule. Suppose, for example, that the good advisor announces message 1 and state 1 is realized.

What inference does the decision maker draw about the advisor's type? The probability that a truth-telling good advisor will announce 1 if the true state is in fact 1 is γ (the probability he observes an accurate signal). The probability that the bad advisor will announce 1 if the true state is 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 by Bayes' rule, the decision maker's posterior belief about the type of the advisor will be

$$\Lambda(\lambda_1, 1, 1) = \frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_1)(\gamma + (1 - \gamma)\nu)}$$

Observe that this is necessarily less than λ_1 (since $\nu > 0$). Thus even though the good advisor always tells the truth, and even though she turned out to be right, her reputation must go down. By similar computations,

$$\begin{aligned} \Lambda(\lambda_1, 1, 0) &= \frac{\lambda_1(1 - \gamma)}{\lambda_1(1 - \gamma) + (1 - \lambda_1)(1 - \gamma + \gamma\nu)}; \\ \Lambda(\lambda_1, 0, 1) &= \frac{\lambda_1}{\lambda_1 + (1 - \lambda_1)(1 - \nu)}; \\ \text{and } \Lambda(\lambda_1, 0, 0) &= \frac{1}{1 + (1 - \lambda_1)(1 - \nu)}. \end{aligned}$$

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

$$\Lambda(\lambda_1, 0, 1) = \Lambda(\lambda_1, 0, 0) > \lambda_1 > \Lambda(\lambda_1, 1, 1) > \Lambda(\lambda_1, 1, 0).$$

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.

We can use these equilibrium updating rules to derive ν as a function of λ_1 . If the bad advisor told the truth with probability ν on observing signal 0, then (applying Bayes' rule) the decision maker would choose action $1 - \gamma$ if he heard message 0 and action $\frac{\gamma + (1 - \lambda_1)(1 - \gamma)\nu}{1 + (1 - \lambda_1)\nu}$ if he heard message 1. Now suppose the advisor observed signal 0. Her current utility from lying (announcing 1) would be

$$y_1 \left(\frac{\gamma + (1 - \lambda_1)(1 - \gamma)\nu}{1 + (1 - \lambda_1)\nu} \right), \quad (3)$$

while her current utility from telling the truth (announcing 0) would be

$$y_1(1 - \gamma). \quad (4)$$

But since she assigns probability $1 - \gamma$ to the true state being 1, her expected value of reputation from lying (announcing 1) would be

$$(1 - \gamma)v_B \left[\frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_1)(\gamma + (1 - \gamma)\nu)} \right] + \gamma v_B \left[\frac{\lambda_1(1 - \gamma)}{\lambda_1(1 - \gamma) + (1 - \lambda_1)(1 - \gamma + \gamma\nu)} \right], \quad (5)$$

while her expected value of reputation from telling the truth (announcing 0) would be

$$v_B \left[\frac{1}{1 + (1 - \lambda_1)(1 - \nu)} \right] \quad (6)$$

In equilibrium, either $\nu = 1$ (the bad advisor always lies) and (3) plus (5) exceeds (4) plus (6); or $0 < \nu < 1$, and there is equality. There is always a unique such ν , since expressions (3 and (5) are strictly decreasing in ν and expressions (4 and (6) are weakly increasing in ν . For example, if $\gamma = \frac{3}{4}$, $y_1 = \frac{1}{10}$ and $y_2 = 1$, so that the bad advisor cares more about the second period decision than the first, then that unique value of ν is plotted (as a function of λ_1) in figure 1. Note that when her reputation is either very low or very high, she knows that her reputation will not change very much as a function of her report, so she will lie most of the time. It is for intermediate values of reputation that she invests in reputation (as in Benabou and Laroque (1992)). On the other hand, if $\gamma = \frac{3}{4}$, $y_1 = 1$ and $y_2 = 1$, so the decision problems are equally important to the bad advisor, then reputational concerns are too small induce to persuade the bad advisor to tell the truth, and the bad advisor would always announce 1 (i.e., we would have $\nu = 1$ for all λ_1).

[insert figure 1 around here]

So far, it was assumed that the good advisor told the truth. If the good advisor observes signal 0, she has an unambiguous incentive to tell the truth, since this will lead the decision maker to choose a low action *and* it will enhance her reputation. But if she observes signal 1, she will gain in terms of the current outcome if she tells the truth (announces 1), but her reputation will be enhanced if she lies (announces 0). Thus if her reputational concerns are sufficiently small, truth-telling will be consistent with equilibrium. This will be true if x_1 is sufficiently large relative to x_2 . But below some critical level of x_1 , reputational concerns will imply that there will not exist an equilibrium where the good advisor always tells the truth. The critical value of x_1 can be calculated explicitly as a function of the parameters. If $\gamma = \frac{3}{4}$ and $y_2 = x_2 = 1$, then figure 2 shows the highest value of x_1 for which a truth-telling equilibrium is possible, for two different values of y_1 . Recall that if $y_1 = 1$, then the bad advisor must always be lying in equilibrium. This makes it relatively attractive for the good advisor to establish a reputation by lying. On the other hand, if $y_1 = \frac{1}{10}$, then the bad advisor is lying often (see figure 1). This makes it harder for the good advisor to establish a reputation for lying and so reduces her incentive to lie.

[insert figure 2 around here]

In general, there will also exist equilibria that are informative but where the good advisor sometimes lies. 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* informative equilibria satisfy the three crucial properties of equilibria where the good advisor always tells the truth.

Proposition 1 *Any informative equilibrium satisfies the following three properties:*

- 1 *The good advisor always announces 0 when she observes signal 0 and announces 1 with positive probability when she observes signal 1;*
- 2 *The bad advisor announces 1 more often than the good advisor;*
- 3 *There is a strict reputational incentive for the advisor to announce 0; more specifically,*

$$\Lambda(\lambda_1, 0, 1) \geq \Lambda(\lambda_1, 0, 0) > \lambda_1 > \Lambda(\lambda_1, 1, 1) \geq \Lambda(\lambda_1, 1, 0).$$

Thus in any informative equilibrium, both types of advisor have a strict reputational incentive to announce 0, whatever signal they observe, in order to look like a good advisor. In equilibrium, such reputational incentives may lead to information being lost (i.e., the decision maker may make a less informed decision). But they cannot bias the decision. Specifically, since the (ex ante) probability of state 1 is $\frac{1}{2}$, the ex ante expected value of the decision maker's action must be $\frac{1}{2}$ in any equilibrium, since the decision maker's action equals his belief about the state and (by a standard property of probability) the expectation of his later belief is his ex ante belief.

The strict reputational incentive to announce 0 now implies that if the second period is sufficiently important, no informative equilibrium exists.

Proposition 2 *If the second period is sufficiently important (i.e., x_2 is large relative to x_1), then no information is conveyed in the first period.*

Thus uninformative play is guaranteed only when the current decision problem is relatively unimportant. In this sense, reputational concerns have an impact exactly when they are least costly. This raises the question of whether a longer relationship (beyond two periods) will make it more or less likely that informative equilibria exist. In Appendix B, there is an analysis of an infinitely repeated version of the advice game. A long horizon has a mixed effect. If the good advisor always told the truth, she would establish a high reputation. But infinite repetition and low discounting imply that the cost of speeding up the reputation acquisition (by lying) is relatively small.

3 The Key Assumptions

This paper follows Sobel (1985) and Bénabou and Laroque (1992) in analyzing reputational concerns that arise endogenously when a static cheap talk game is repeated. The advisor cares about her reputation not because others will treat her differently, but simply because she wants her advice to be accepted (i.e., believed) in the future. It was useful to focus on this explanation in order to emphasize how reputational concerns may impose constraints on communication *even among individuals whose only interaction is the communication they are engaged in.*

However, this is unlikely to be the *only* reason for reputational concerns in most environments. The economics literature typically focuses on other instrumental reasons for reputational concerns. Thus in the examples discussed in the previous section, an advisor may not wish to be perceived to favor special interests because she has political ambitions that would be thwarted if she was perceived to be a lackey of special interests; a supervisor may wish to be perceived to be a good supervisor so that she will receive salary increases in the future (Holmström (1999) and Holmström and Ricart i Costa (1986)); an investment advisor who charges a fixed fee for offering advice may wish to establish a reputation for being imparital so that she will re-hired in the future (Campbell (1998) and Chevalier and Ellison (1997, 1999)). The analysis of first period behavior summarized above in Propositions 1 and 2 was independent of why the advisor (good or bad) has reputational concerns.

However, the analysis did depend on certain key features of the advice game. The remainder of this section contains a discussion of how the main conclusions about period 1 behavior would change if the assumptions about the period 1 game were varied, holding reputational concerns fixed.

The Communication and Incentive Assumptions

The model does not allow the decision maker to commit to a contract that would allow her to be rewarded as a function of whether her advice turned out to be correct ex post. Nor is the decision maker able to commit to a decision rule (as a function of messages) before the advisor sends a message. If the decision maker had the ability to make either kind of commitment, he would in general do so. The model thus fits most clearly public debate environments when there is no relationship between the advisor and the decision maker other than the communication that they are engaged in. But there are also many contexts where there is an ongoing relationship between a decision maker and his advisor where the decision maker neither rewards the advisor directly on the basis of the accuracy of the advice, nor commits to a decision rule upfront. For example, this is typically true of the motivating examples cited above: politician / policy expert, investor / financial advisor, personnel officer / supervisor, journal editor / referee. In each case, it is reasonable to suppose that the advisor is motivated primarily by some preferences over the current decision made and a desire to improve her reputation. As noted earlier, reputational concerns may arise from many sources, including the objectives of being re-elected, being hired again, being promoted and influencing future decisions.

The model also only allows the advisor to communicate her information by cheap talk. Very different conclusions arise in costly signalling models. Recall that in equilibrium, a good advisor who observes signal 1 must trade off the current benefit of making a truthful announcement (leading to a better current decision by the decision maker) and the future cost (the lowering of her reputation). But both the cost and benefit are endogenous: they are determined by the decision maker's beliefs which in turn are a function of advisor's strategy. It is this endogeneity of signalling costs that leads to the paradoxical conclusion of Proposition 2: when the advisor is given an increased incentive to separate (i.e., increased reputational concerns), separation becomes impossible in equilibrium. To put this a different way, if the good advisor always tells the truth in equilibrium, there would be a significant current cost to announcing 0 if she in fact observed signal 1 (i.e., falsely announcing 0 for reputational reasons would lead the decision maker to choose a significantly lower action). But if the good advisor announced 0 most of the time, the current cost of announcing 0 is very small (since the decision maker does not deduce much from the announcement). So there are endogenously decreasing costs to signalling.

There is a simple way to relate this cheap talk model to a model with (exogenously) costly signalling of preference type. Suppose that the decision maker was able to delegate the decision to the advisor (who cared about her reputation for some reason). Under natural single crossing properties, a good advisor turned decision maker could always choose a sufficiently low action to separate 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.

The Biased Advisor Assumption

Our results follow from a particular and extreme assumption about the possible preferences of the advisor: the advisor's preferences over the current decision either coincide with the decision maker (the good type) or are biased in a particular, commonly known,

direction (the bad advisor). The importance of this assumption can be illustrated by briefly discussing what would happen in a number of other cases.

1. In Sobel (1985) and Bénabou and Laroque (1992), the 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, there is no reputational cost to telling the truth. Thus there is always an equilibrium where the good advisor always tells the truth. (Thus although the above two papers in fact *assumed* the good advisor always told the truth, their equilibria would remain equilibria if the good advisor also had reputational concerns).
2. Similarly, if we combined the good advisor of this paper (with the same current preferences as the decision maker) with *two* symmetrically bad advisors, with the two bad advisors biased in opposite directions, there is always an equilibrium where the good advisor truthfully announces her signal and the bad advisors always announce the signal favoring their most preferred action. However, this result is very sensitive to our two signal assumption. If we expanded the set of states and signals, this three type model would lead to clustering of messages in the middle, in the spirit of Bernheim’s (1994) model of conformity, and there would be a different kind of information loss.
3. The bad advisor of this paper (biased in a particular direction) may also be combined with a good advisor who likes to tell truth (as well as having reputational concerns). This type of good advisor would have a current incentive to tell truth *even if the decision maker does not believe her*. This would ensure that the good advisor’s cost of signalling her type was exogenous. In this case, we would lose the feature of the current model (described above) that there are endogenously decreasing costs to the good advisor of signalling her type. This would make it more likely that the good advisor would separate from the bad advisor in equilibrium.

The Noisy Information Assumption

The advisor was assumed to have noisy and unverifiable information. If the advisor’s information were perfect (i.e., $\gamma = 1$), there would always exist a sequential equilibrium where the advisor (of whatever type) would tell the truth if she cared enough about second period decisions. This behavior would be consistent with equilibrium if the decision maker inferred that any advisor whose message was not equal to the realized state were surely bad. Similarly, if the advisor were able to prove ex post what signal she had observed, truth-telling could be enforced by reputational concerns. Thus this model applies in situations where the information communicated is “soft,” i.e., reflecting the tacit knowledge of an expert assessment, and not “hard,” i.e., objectively describable.

4 Welfare Analysis

Reputational concerns lead to the loss of socially valuable information. Does that mean that it would be socially desirable to prevent learning about the advisor’s type? In particular, how do players’ utilities in the equilibria with reputational updating analyzed above compare with their utility if there were no reputational updating, i.e., if the decision

maker's belief entering the second period remained at λ_1 ? This latter scenario would arise if there were a different decision maker in the second period, with identical preferences to the first period decision maker but unable to observe first period outcomes.

In answering this question, first note that in all equilibria, the ex ante expectation of the decision maker's belief about the advisor's type is $1/2$. Thus an individual with the bad advisor's (linear) preferences is indifferent between all equilibria, since the ex ante expected action of the decision maker is always $\frac{1}{2}$ in each period. Thus the welfare analysis can be restricted to the impact on the decision maker (recall that the good advisor has identical preferences to the decision maker).

There are three welfare effects at work:

- The *Discipline Effect*. Without reputational updating, the bad advisor always announces 1 in the first period. With reputational updating, 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*. With reputational updating, 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 good for the decision maker.
- The *Political Correctness Effect*. With reputational updating, 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.

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).

The overall welfare effect is ambiguous. If truth-telling by the good advisor in both periods is consistent with equilibrium, then there is no (bad) political correctness effect with reputational updating and the (good) discipline and sorting effects must work to the decision maker's advantage. As we noted above, if the first period problem is more important to the decision maker than the second period decision problem, then this will exist such a truth-telling equilibrium. On the other hand, when informative second period behavior implies babbling in the first period, the (bad) political correctness effect arises with reputational updating, and the (good) discipline and sorting effects cannot exist, since first period behavior is completely uninformative. The overall message is that reputational updating may be valuable, but if it becomes too valuable, it can be self-defeating.

5 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 do not depend on others' beliefs directly; if people care about what other people think of them, it is for *instrumental* reasons.

In the model of this paper, 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.
- [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] Banerjee, A. and R. Somanathan (1997). "A Simple Model of Voice," M.I.T. and Emory University.
- [6] Bénabou, R. and G. Laroque (1992). "Using Privileged Information to Manipulate Markets: Insiders, Gurus and Credibility," *Quarterly Journal of Economics* **107**, 921-958.
- [7] Bernheim, D. (1994). "A Theory of Conformity," *Journal of Political Economy* **102**, 841-877.
- [8] Brandenburger, A. and B. Polak (1996). "When Managers Cover Their Posteriors: Making the Decisions the Market Wants to See," *RAND Journal of Economics* **27**, 523-541.
- [9] Campbell, C. (1998). "Learning and the Market for Information," The Ohio State University.
- [10] Chevalier, J. and G. Ellison (1997). "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, **105**, 1167-1200.
- [11] ——— (1999). "Career Concerns of Mutual Fund Managers," *Quarterly Journal of Economics* **114**, 389-432.
- [12] Crawford, V. and J. Sobel (1982). "Strategic Information Transmission," *Econometrica* **50**, 1431-1451.

- [13] Dewatripont, M. and J. Tirole (1999). “Advocates,” *Journal of Political Economy* **107**, 1-39.
- [14] Fudenberg, D. and D. Levine (1992). “Maintaining a Reputation when Strategies are Imperfectly Observed,” *Review of Economic Studies* **59**, 561-579.
- [15] Glazer, J. and A. Rubinstein (1998). “Motives and Implementation: On the Design of Mechanisms to Elicit Opinions,” *Journal of Economic Theory* **79**, 157-173.
- [16] Holmström, B. (1999). “Managerial Incentive Problems: A Dynamic Perspective,” *Review of Economic Studies* **226**, 169-182.
- [17] Holmström, B. and Ricart i Costa, J. (1986). “Managerial Incentives and Capital Management,” *Quarterly Journal of Economics* **101**, 835-860.
- [18] Krishna, V. and J. Morgan (1998). “A Model of Expertise,” Penn State and Princeton.
- [19] Levy, G. (1998). “Strategic Consultation and ‘Yes Man’ Advisors. Princeton University.
- [20] Loury, G. (1994). “Self-Censorship in Public Discourse: A Theory of ‘Political Correctness’ and Related Phenomena,” *Rationality and Society* **6**, 428-461.
- [21] 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.
- [22] Ottaviani, M. and P. Sorensen (1998). “Information Aggregation in Debate,” University College, London, and Nuffield College, Oxford.
- [23] ——— (1999). “Professional Advice,” University College, London, and Nuffield College, Oxford.
- [24] Prendergast, C. (1993). “A Theory of Yes Men,” *American Economic Review* **83**, 757-770.
- [25] ——— and L. Stole (1996). “Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning,” *Journal of Political Economy* **104**, 1105-1134.
- [26] Scharfstein, D. and J. Stein (1990). “Herd Behavior and Investment,” *American Economic Review* **80**, 465-479.
- [27] Shin, H. (1998). “Adversarial and Inquisitorial Procedures in Arbitration,” *RAND Journal of Economics* **29**, 378-405.
- [28] Sobel, J. (1985). “A Theory of Credibility,” *Review of Economic Studies* **52**, 557-573.
- [29] Spector, D. and T. Piketty (1997). “Rational Debate leads to One-Dimensional Conflict,” MIT.

Appendix A: A Static Advice Game with Exogenous Reputational Concerns

This appendix describes and analyzes a static advice game, where advisors have exogenous reputational concerns. Solving this game is equivalent to solving for first period equilibrium behavior in the two period model, given the reputational value functions (1) and (2) generated by the informative equilibrium in the second period. The model in this Appendix is more general than that in the text. In particular, the analysis will show that Propositions 1 and 2 remain true for any strictly increasing reputational value functions and for a more general class of payoffs.

A state of the world $\omega \in \{0, 1\}$ is drawn; each state is equally likely. The advisor observes a signal $s \in \{0, 1\}$, which is correct with probability γ , where $1/2 < \gamma < 1$. The advisor is good (G) with probability λ , bad (B) with probability $1 - \lambda$. The type I advisor's strategy is a function $\sigma_I : \{0, 1\} \rightarrow [0, 1]$, where $\sigma_I(s)$ is the probability of announcing message 1 when her signal is s . The decision maker's strategy is a function $\chi : \{0, 1\} \rightarrow \mathbf{R}$; $\chi(m)$ is his action if m is the message from his advisor. We will allow somewhat more general preferences than those considered in the text. The decision maker's utility given by $u_{DM}(a, \omega)$, where $u_{DM}(a, \omega)$ is differentiable and strictly concave in a and attains a maximum for each ω . Write $a^*(m) = \arg \max_{a \in \mathbf{R}} u_{DM}(a, \omega)$ and assume $a^*(1) > a^*(0)$. 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

$$\Lambda(m, \omega) = \frac{\lambda \phi_G(m | \omega)}{\lambda \phi_G(m | \omega) + (1 - \lambda) \phi_B(m | \omega)}, \quad (7)$$

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 $\phi_I(0 | \omega) = 1 - \phi_I(1 | \omega)$. Note that equation (7) for $\Lambda(m, \omega)$ 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.

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

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

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

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

where $y > 0$ and u_B is a strictly increasing and continuous on the interval $[a^*(1 - \gamma), a^*(\gamma)]$ and $v_B : [0, 1] \rightarrow \mathbf{R}$ is a strictly increasing continuous function. Note that the payoffs in the text are a special case, where $u_{DM}(a, \omega) = -(a - \omega)^2$, $u_B(a) = a$, $x = x_1/x_2$, $y = y_1/y_2$ and $v_G(\cdot)$ and $v_B(\cdot)$ are given by equations (1) and (2), respectively.

An alternative interpretation of these payoff functions is 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).

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

$$\Gamma(m) = \frac{\lambda\phi_G(m|1) + (1-\lambda)\phi_B(m|1)}{\lambda\phi_G(m|1) + (1-\lambda)\phi_B(m|1) + \lambda\phi_G(m|0) + (1-\lambda)\phi_B(m|0)}. \quad (8)$$

Again, this is well defined only if the denominator is non-zero. By convention, $\Gamma(m) = \frac{1}{2}$ 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 (7) and (8).

In the text, the value function was derived endogenously. However, we could also 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). 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).

The following notation will also be useful. Write $\hat{u}_G(q, s)$ 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 ,

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

Similarly, write $\hat{u}_B(q)$ 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:

$$\hat{u}_B(q) \equiv u_B(\tilde{a}(q)).$$

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

Fact. $\hat{u}_G(q, 1)$ is strictly increasing in q if $q \in (1-\gamma, \gamma)$; $\hat{u}_G(q, 0)$ is strictly decreasing in q if $q \in (1-\gamma, \gamma)$; $\hat{u}_B(q)$ is strictly decreasing in q if $q \in (1-\gamma, \gamma)$.

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.,

$$\begin{aligned} \Pi_G^C(s) &= x[\widehat{u}_G(\Gamma(1), s) - \widehat{u}_G(\Gamma(0), s)] \\ \text{and } \Pi_B^C(0) = \Pi_B^C(1) &= y[\widehat{u}_B(\Gamma(1)) - \widehat{u}_B(\Gamma(0))]. \end{aligned} \quad (9)$$

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.,

$$\begin{aligned} \Pi_I^R(1) &= \gamma \begin{bmatrix} v_I(\Lambda(0, 1)) \\ -v_I(\Lambda(1, 1)) \end{bmatrix} + (1 - \gamma) \begin{bmatrix} v_I(\Lambda(0, 0)) \\ -v_I(\Lambda(1, 0)) \end{bmatrix} \\ \text{and } \Pi_I^R(0) &= (1 - \gamma) \begin{bmatrix} v_I(\Lambda(0, 1)) \\ -v_I(\Lambda(1, 1)) \end{bmatrix} + \gamma \begin{bmatrix} v_I(\Lambda(0, 0)) \\ -v_I(\Lambda(1, 0)) \end{bmatrix}. \end{aligned} \quad (10)$$

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

The decision maker's optimal action depends only on how likely he thinks the two states; the assumptions 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 $(\sigma_G, \sigma_B, \chi, \Gamma, \Lambda)$,*

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

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

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

PROOF. 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

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

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

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.

Lemma 2 *Every babbling strategy profile is an equilibrium.*

PROOF. 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. ■

Thus the interesting issue is the existence and properties of informative (non-babbling) equilibria. In analyzing informative equilibria, attention is restricted 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.

Proposition 3 *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 *The bad advisor announces 1 more often than the good advisor: $\sigma_B(1) \geq \sigma_G(1)$ and $\sigma_B(0) \geq \sigma_G(0) = 0$, with one of the inequalities holding strictly;*
- 3 *There is a strict reputational incentive for the advisor to announce 0; more specifically, $\Lambda(0, 1) \geq \Lambda(0, 0) > \lambda > \Lambda(1, 1) \geq \Lambda(1, 0)$.*

PROOF. This will be proved in nine steps. Each step identifies 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 it is assumed (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. The proof shows by contradiction that no equilibrium exists if one of these conditions is violated.

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.
2. Suppose that $\Lambda(1, 1) > \Lambda(0, 1)$ and $\Lambda(1, 0) \leq \Lambda(0, 0)$. By definition of Λ (see equation 7) we have

$$\gamma\sigma_G(1) + (1 - \gamma)\sigma_G(0) = \phi_G(1 | 1) > \phi_B(1 | 1) = \gamma\sigma_B(1) + (1 - \gamma)\sigma_B(0) \quad (11)$$

$$\text{and } \gamma\sigma_G(0) + (1 - \gamma)\sigma_G(1) = \phi_G(1 | 0) \leq \phi_B(1 | 0) = \gamma\sigma_B(0) + (1 - \gamma)\sigma_B(1). \quad (12)$$

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 9 and 10). 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 (11) implies $\sigma_G(1) > \sigma_B(1)$, while (12) implies $\sigma_G(1) \leq \sigma_B(1)$, a contradiction; (ii) If $\sigma_G(0) = 0$ and $\sigma_B(1) = 1$, then (11) implies $\sigma_G(1) > 1$, a contradiction; (iii) If $\sigma_G(1) = 1$ and $\sigma_B(0) = 0$, then (12) implies $\sigma_B(1) = 1$ and $\sigma_G(0) = 0$, which implies $\phi_G(1 | 1) = \phi_B(1 | 1)$, contradicting (11); (iv) If $\sigma_G(1) = \sigma_B(1) = 1$, then (11) implies $\sigma_G(0) > \sigma_B(0)$, while (12) implies $\sigma_G(0) \leq \sigma_B(0)$, a contradiction.

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

$$\gamma\sigma_G(1) + (1 - \gamma)\sigma_G(0) = \phi_G(1 | 1) \leq \phi_B(1 | 1) = \gamma\sigma_B(1) + (1 - \gamma)\sigma_B(0) \quad (13)$$

$$\text{and } \gamma\sigma_G(0) + (1 - \gamma)\sigma_G(1) = \phi_G(1 | 0) \leq \phi_B(1 | 0) = \gamma\sigma_B(0) + (1 - \gamma)\sigma_B(1). \quad (14)$$

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 (13) and (14), 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 7) 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 7).

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 1 is proved by **P4** and **P9**. Part [2] is proved by **P2**. Part [3] is proved by **P8**. ■

The next proposition examines the existence of an informative equilibrium in the game parameterized by (λ, x, y) .

Proposition 4 *For any $\lambda \in (0,1)$ and $y \in \mathbf{R}_{++}$, there exist $0 < \underline{x}(\lambda, y) \leq \bar{x}(\lambda, y)$ 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 $x \geq \bar{x}(\lambda, y)$.*

The proof gives explicit forms for \bar{x} and \underline{x} (equations 15 and 16 respectively); these can be used to show the following limiting properties. As $\lambda \rightarrow 1$, 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 $\underline{x}(\lambda, y) \rightarrow 0$ and $\bar{x}(\lambda, y) \rightarrow 0$ as $\lambda \rightarrow 1$. As $\lambda \rightarrow 0$, 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 $\bar{x}(\lambda, y)$ tends to some positive constant

also. As $y \rightarrow 0$, the bad advisor's strategy will mimic the good advisor's strategy, so reputational concerns must become smaller; so $\underline{x}(\lambda, y) \rightarrow 0$ and $\bar{x}(\lambda, y) \rightarrow 0$ as $y \rightarrow 0$. Finally, if y is sufficiently large, the bad advisor will always announce 1 in any non-babbling equilibrium. Thus $\underline{x}(\lambda, y)$ and $\bar{x}(\lambda, y)$ become constant for all sufficiently large y .

PROOF.

[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 $\chi(\cdot) = \tilde{a}(\Gamma(\cdot))$. 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.,

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

This expression is strictly decreasing in ν , since each term is weakly decreasing in ν , and some are strictly decreasing. Also $g(0) = y[\hat{u}_B(\gamma) - \hat{u}_B(1-\gamma)] > 0$. Thus there exists exactly one value of ν where either $g(\nu) = 0$ or $\nu = 1$ and $g(\nu) > 0$. This ν parameterizes the unique equilibrium. Write $\tilde{\nu}(\lambda, y)$ 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 $\sigma_G(0) = 0$, $\sigma_G(1) = 1$, $\sigma_B(0) = \tilde{\nu}(\lambda, y)$, $\sigma_B(1) = 1$, and $\chi(\cdot) = \tilde{a}(\Gamma(\cdot))$. She will tell the truth if and only if

$$\left\{ \begin{aligned} &x \left[\hat{u}_G \left(\frac{\gamma + (1-\lambda)(1-\gamma)\tilde{\nu}(\lambda, y)}{1 + (1-\lambda)\tilde{\nu}(\lambda, y)}, 1 \right) - \hat{u}_G(1-\gamma, 1) \right] \\ &+ \gamma v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) \left(1 + \left(\frac{\gamma}{1-\gamma}\right)\tilde{\nu}(\lambda, y)\right)} \right] + (1-\gamma) v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) \left(1 + \left(\frac{1-\gamma}{\gamma}\right)\tilde{\nu}(\lambda, y)\right)} \right] \\ &- v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) (1-\tilde{\nu}(\lambda, y))} \right] \end{aligned} \right\} \geq 0,$$

i.e., $x \geq \bar{x}(\lambda, y)$, where $\bar{x}(\lambda, y)$ equals

$$\frac{v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) (1-\tilde{\nu}(\lambda, y))} \right] - \gamma v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) \left(1 + \left(\frac{\gamma}{1-\gamma}\right)\tilde{\nu}(\lambda, y)\right)} \right] - (1-\gamma) v_G \left[\frac{1}{1 + \left(\frac{1-\lambda}{\lambda}\right) \left(1 + \left(\frac{1-\gamma}{\gamma}\right)\tilde{\nu}(\lambda, y)\right)} \right]}{\left[\hat{u}_G \left(\frac{\gamma + (1-\lambda)(1-\gamma)\tilde{\nu}(\lambda, y)}{1 + (1-\lambda)\tilde{\nu}(\lambda, y)}, 1 \right) - \hat{u}_G(1-\gamma, 1) \right]} \quad (15)$$

[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 1, 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 $\omega \in \{0, 1\}$,

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

It will be shown that:

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 $\left(\frac{1-\gamma}{2\gamma} \right)$ -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 $\phi_B(1 | 1) \geq \phi_G(1 | 1) = \gamma\sigma_G(1)$; so

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

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

$$\underline{x}(\lambda, y) = \frac{f\left(\lambda, \min\left\{\frac{1-\gamma}{2\gamma}, \frac{2\gamma-1}{2(1-\gamma)}, h\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right\}\right)}{\widehat{u}_G(\gamma, 1) - \widehat{u}_G(1-\gamma, 1)}.\quad (16)$$

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

$$\begin{aligned}\Pi_G^R(1) &\leq \Pi_G^C(1) \\ &\leq x [\widehat{u}_G(\gamma, 1) - \widehat{u}_G(1-\gamma, 1)] \\ &\leq f\left(\lambda, \min\left\{\frac{1-\gamma}{2\gamma}, \frac{2\gamma-1}{2(1-\gamma)}, h\left(\lambda, \frac{1}{2}y\left(\widehat{u}_B\left(\frac{\gamma}{\gamma+\frac{1}{2}}\right) - \widehat{u}_B\left(\frac{1}{2}\right)\right)\right\}\right).\end{aligned}$$

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)**. ■

Appendix B: The Infinite Horizon Game

Now let the static game be repeated infinitely often, with a new decision problem in each period. The decision maker and bad advisor both discount the future (with perhaps different discount rates). The good advisor is assumed to the preferences of the decision maker (and no intrinsic reputational concerns). Finally, the importance of the decision problem in each period is allowed to vary through time.

Each period t 's decision problem is parameterized by (x_t, y_t) , the importance of the problem for the decision maker (and good advisor) and bad advisor respectively. It is assumed that x_t and y_t are drawn from X and Y respectively, which are discrete subsets of \mathbf{R}_{++} ; write $\phi \in \Delta(X \times Y)$ for the probability distribution on $X \times Y$. Assume that ϕ has infinite support but that

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

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 $\sum_{t=0}^{\infty} (\delta_{DM})^t x_t u_{DM}(a_t, \omega_t)$ and the bad advisor receives total payoff $\sum_{t=0}^{\infty} (\delta_B)^t y_t u_B(a_t)$. A (Markov) advisor strategy is a pair (σ_G, σ_B) , each $\sigma_I : \{0, 1\} \times (0, 1) \times X \times Y \rightarrow [0, 1]$; $\sigma_I(s; \lambda, x, y)$ 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 $\chi : \{0, 1\} \times (0, 1) \times X_G \times Y \rightarrow \mathbf{R}$, where $\chi(m; \lambda, x, y)$ is the decision maker's action if he receives message m .

Definition 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. 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 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.

Nonetheless, the analysis that follows focuses on monotonic Markov equilibria. The objective here is simply to show that the behavior described in the static model does arise in a stationary infinite horizon model. In particular, it is shown first that monotonic Markov equilibria do always exist. Then it is shown that in any such monotonic Markov equilibrium, there is always babbling in periods when the decision problem is sufficient

unimportant to the decision maker. Finally, it is shown that if there is no variation in the importance of the decision problem and the discount rate approaches one, the good advisor does not necessarily have an incentive to tell the truth.

Proposition 5 *A monotonic Markov equilibrium 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^*) .

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

$$\bar{x}_G = \left(\frac{1}{1-\varepsilon} \right) \sum_{(x,y) \neq (x^*, y^*)} x \cdot \phi(x, y) \text{ and } \bar{x}_B = \left(\frac{1}{1-\varepsilon} \right) \sum_{(x,y) \neq (x^*, y^*)} y \cdot \phi(x, y)$$

and consider the following advisor strategy

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

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

The best response for the decision maker is

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

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

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

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

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

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

$$v_G(1) - v_G(0) \leq \frac{\varepsilon x^*}{1-\delta_G} \left[\frac{1}{2}(\hat{u}_G(\gamma, 1) - \hat{u}_G(1-\gamma, 1)) + \frac{1}{2}(\hat{u}_G(1-\gamma, 0) - \hat{u}_G(\gamma, 0)) \right] \quad (17)$$

$$\text{and } v_B(1) - v_B(0) \leq \frac{\varepsilon y^*}{1-\delta_B} [\hat{u}_B(\gamma) - \hat{u}_B(1-\gamma)].$$

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.,

$$\begin{aligned}
\Pi_B^C(1) &= \Pi_B^C(0) \\
&= y^* \left(\widehat{u}_B \left(\frac{\lambda\gamma + (1-\lambda)}{\lambda + 2(1-\lambda)} \right) - \widehat{u}_B(1-\gamma) \right) \\
&\geq y^* \left(\widehat{u}_B \left(\frac{1}{2} \right) - \widehat{u}_B(1-\gamma) \right) \\
\text{and } \Pi_G^C(1) &= x^* \left\{ \begin{aligned} &\gamma \left[\widehat{u}_G \left(\frac{\lambda\gamma + (1-\lambda)}{\lambda + 2(1-\lambda)}, 1 \right) - \widehat{u}_G(1-\gamma, 1) \right] \\ &+ (1-\gamma) \left[\widehat{u}_G \left(\frac{\lambda\gamma + (1-\lambda)}{\lambda + 2(1-\lambda)}, 0 \right) - \widehat{u}_G(1-\gamma, 0) \right] \end{aligned} \right\} \\
&\geq x^* \left\{ \begin{aligned} &\gamma \left[\widehat{u}_G \left(\frac{1}{2}, 1 \right) - \widehat{u}_G(1-\gamma, 1) \right] \\ &+ (1-\gamma) \left[\widehat{u}_G \left(\frac{1}{2}, 0 \right) - \widehat{u}_G(1-\gamma, 0) \right] \end{aligned} \right\}.
\end{aligned} \tag{18}$$

Thus by choosing (x^*, y^*) with $\varepsilon = \phi(x^*, y^*)$ sufficiently small, we have (by equations 17 and 18) that $\Pi_I^C(1) > v_I(1) - v_I(0) \geq \Pi_I^R(1)$ for $I = G$ or B , and thus the proposed strategies are optimal. ■

Monotonic Markov equilibria inherit all the structure of propositions 1 and 2. In particular, fix a monotonic Markov equilibrium and any given λ and y ; there exists \underline{x} such that for all $x \leq \underline{x}$, there is babbling at every history where the advisors reputation is λ and (x, y) are the values of the current decision problem.

We conclude with a brief analysis of what happens if there is no variation in x and y and the discount rate of the decision maker goes to 1 (say without loss of generality, they were equal to 1 in every period). What can we say about monotonic Markov equilibria in this case? It will be shown by contradiction that for at least some discount rates for the bad advisor and utility functions for the decision maker, there is not a truth-telling equilibrium. In particular, suppose that the good advisor always told the truth and the bad advisor always announced 1 (this is a best response for the bad advisor if δ_B is sufficiently close to 0). Then we would have $\Gamma(0) = 1 - \gamma$, $\Lambda(0, 1) = 1$ and $\Lambda(0, 0) = 1$; and for small λ , we would also have $\Gamma(1) \approx \frac{1}{2}$, $\Lambda(1, 1) \approx 0$ and $\Lambda(1, 0) \approx 0$. So suppose the good advisor has reputation close to 0 and has just observed signal 1. What is the expected net gain from lying and announcing 0? There is a current *loss* of $\widehat{u}_{DM}(\frac{1}{2}, 1) - \widehat{u}_{DM}(1-\gamma, 1)$. The benefit is that at future histories where signal 1 is realized and signal 0 has never occurred, the decision maker will receive $\widehat{u}_{DM}(\gamma, 1) - \widehat{u}_{DM}(\frac{1}{2}, 1)$. There is a probability $\frac{1}{2}$ that such a history occurs next period, probability $\frac{1}{4}$ occurs next period, and so on. Thus if δ_{DM} is sufficiently close to 1 and λ is sufficiently close to 0, there is an incentive for the good advisor to lie if

$$2\widehat{u}_{DM} \left(\frac{1}{2}, 1 \right) < \widehat{u}_{DM}(1-\gamma, 1) + \widehat{u}_{DM}(\gamma, 1)$$

This condition holds for some strictly concave utility functions, and in this case there cannot be any truth-telling equilibrium.

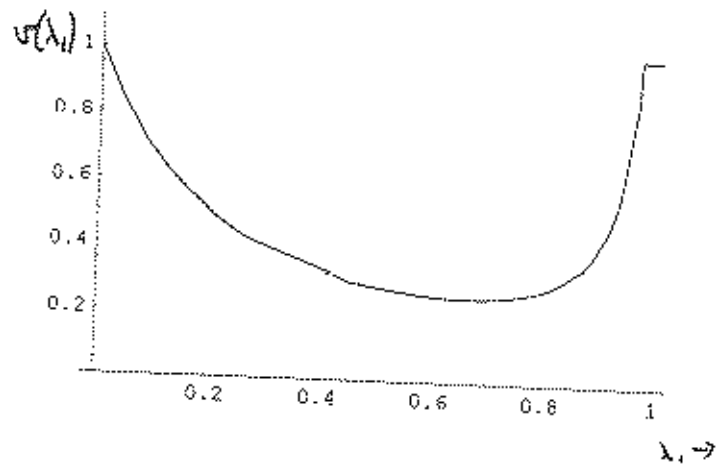


FIGURE 1

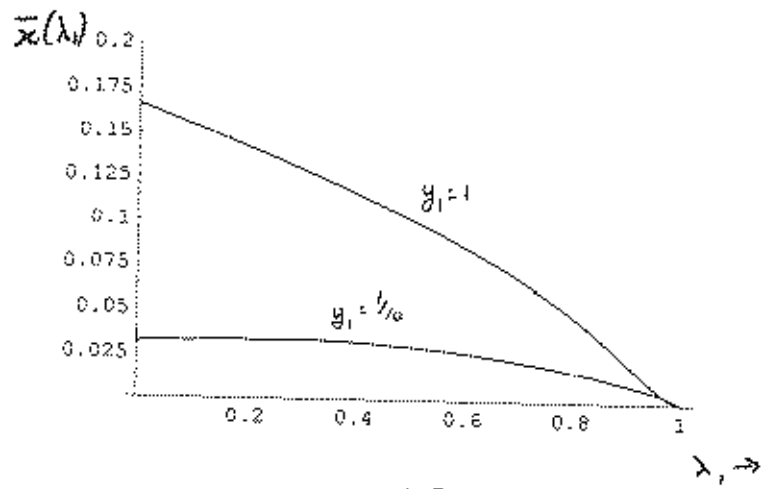


FIGURE 2