Note on Morris (2001): "Political Correctness"

Pei-yu Lo
December 4, 2013

1 Gist

reputational concern may decrease information transmitted in equilibrium, even if the sender and the receiver have identical preference.

An action may be desirable in some situation, but is associated with some "stigma". Therefore, an informed expert may hesitate to recommend such an action for fear of compromised reputation even if the expert believes that such an action is preferred given her information.

When reputation concern is very strong, in equilibrium, no information is transmitted.

2 Model

2.1 Elicit information from experts.

A decision maker’s optimal decision depends on state of the world $\omega \in \{0, 1\}$. DM believes that $\omega = 1$ with probability $\frac{1}{2}$.

An advisor observes a signal $s_1 \in \{0, 1\}$. $p(s|\omega) = \gamma$ if $s = \omega$ and $(1 - \gamma)$ if $s \neq \omega$. $\gamma > \frac{1}{2}$.

Whether an informative equilibrium exists in the static advice game depends on DM and advisor’s preferences.

$u_{DM}(a, \omega) = -(a - \omega)^2$.

Good advisor’s preference is identical to DM’s.

Bad advisor prefers $a = 1$ independent of $\omega$. Bad advisor’s preference can be represented by utility function $a$.

Cheap talk game: after observing a signal, advisor sends $m \in \{0, 1\}$ to DM. Then DM makes a decision.

2.1.1 Common knowledge of advisor’s preference

There is always a babbling equilibrium.

If it is common knowledge that advisor is good, then there exists an informative equilibrium. Advisor sends $m = 1$ if she observes $s_1 = 1$ and $m = 0$ if
\[ s_1 = 0. \]

\[
E_\omega [u_{DM} (a, \omega)] = - \Pr \{\omega = 1\} (a - 1)^2 - \Pr \{\omega = 0\} a^2 \\
= - (a - \Pr \{\omega = 1\})^2 - \Pr \{\omega = 1\} \Pr \{\omega = 0\}.
\]

\[
0 = \frac{\partial E_\omega [u_{DM} (a^*, \omega)]}{\partial a}
\]

implies that \( a^* = E_\omega [\omega] = \Pr \{\omega = 1\} \). So in the informative equilibrium, DM chooses \( a = \gamma \) when receiving \( m = 1 \) and \( a = 1 - \gamma \) when receiving \( m = 0 \).

### 2.1.2 Incomplete information about the advisor’s preference, static advice game

Suppose DM believes that advisor is good with probability \( \lambda \). \( \lambda \) is advisor’s "reputation".

There’s always a babbling equilibrium.

We focus on the informative equilibrium.

**Proposition 1** In the static advice game, there is a unique informative equilibrium in which DM takes a higher action after one message (say message 1) than another (say message 0).

**Proof.** The bad advisor strictly prefers to send the message that induce the highest action from DM regardless of the signal she receives.

Given any belief DM holds about the advisor’s type and the signal the advisor receives, DM’s belief that \( \omega = 1 \) must be in \([1 - \gamma, \gamma]\). This is because DM cannot have more information than is contained in the advisor’s signal. But then because the good advisor’s preference is identical to DM’s, and

\[
E_\omega [u_{DM} (a, \omega)] = - (a - \Pr \{\omega = 1\})^2 - \Pr \{\omega = 1\} \Pr \{\omega = 0\},
\]

when \( s_1 = 0, \Pr \{\omega = 1|s_1 = 0\} = (1 - \gamma) \), so the good advisor’s payoff is maximized at \( a = 1 - \gamma \) and decreases as \( a \) increases; when \( s_1 = 1, \Pr \{\omega = 1|s_1 = 1\} = \gamma \), so the good advisor’s payoff is maximized at \( a = \gamma \) and decreases as \( a \) decreases from \( \gamma \). Therefore, when \( s_1 = 1 \), good advisor prefers to send the message that will induce the highest action from DM, and when \( s_1 = 0 \), good advisor will send the message that will induce the lowest action from DM. Therefore, it is w.l.o.g. to assume that there are only two possible messages, \( m \in \{0, 1\} \). Good advisor has strict incentive to tell the truth. Because a bad advisor prefers a higher action, bad advisor sends \( m = 1 \) with probability 1 regardless of the signal he receives.

After receiving message 0, DM knows that the advisor has received signal 0. So he believes that the state is 1 with probability \( 1 - \gamma \) and takes action \( 1 - \gamma \).
After receiving message 1, DM believes that the state is 1 with probability

\[ \frac{1}{2} \left( \lambda \gamma + (1 - \lambda) \right) \]

\[ = \frac{1 - \lambda + \lambda \gamma}{2 - \lambda} \]

and optimally chooses \( a = \frac{1 - \lambda + \lambda \gamma}{2 - \lambda} \). It is equal to \( \frac{1}{2} \), the ex ante expected value of \( \omega \), if \( \lambda = 0 \) (common knowledge that advisor is bad) and \( \gamma \) if \( \lambda = 1 \) (common knowledge that advisor is good). This action is increasing in \( \lambda \) because \( \frac{1 - \gamma + \frac{1 - \lambda}{1 + \frac{1 - \lambda}{2}}} {1 + \frac{1 - \lambda}{1 + \frac{1 - \lambda}{2}}} \) is increasing in \( \frac{1 - \lambda}{\lambda} \). Thus, the higher the advisor’s reputation is, the more influence he will have, i.e. the higher DM’s action is conditional on message 1.

### 2.2 Two Period advice game

Endogenize reputational concern through desire to influence DM’s future decision.

The game in the first and second period are identical to the static game, with independent states of the world \( \omega_1, \omega_2 \in \{0, 1\} \), signals \( s_1, s_2 \) the distribution of which depends only on the state of the world in that period.

In the beginning of the two-period game, DM believes that the advisor is good with probability \( \lambda_1 \). In the first period, DM receives message \( m_1 \), takes action \( a_1 \), and afterwards, observes the true state of the world \( \omega_1 \). DM then updates his belief about the advisor to \( \lambda_2 = \Lambda(\lambda_1, m_1, \omega_1) \) using equilibrium strategy and Bayes rule. There is a new state of the world \( \omega_2 \) in period 2. The advisor observes a signal \( s_2 \) and sends message \( m_2 \). The advisor then takes action \( a_2 \).

The total utility of the DM and the good advisor is given by

\[ -x(a_1 - \omega_1)^2 - (a_2 - \omega_2)^2 \]

and the total utility of the bad advisor is given by

\[ ya_1 + a_2. \]

\( x > 0 \) and \( y > 0 \) captures the relative weight of the first period payoff for DM/good advisor and for the bad advisor.

#### 2.2.1 Second Period

After any history \((m_1, a_1, \omega_1)\), DM updates his belief about the advisor to \( \lambda_2 = \Lambda(m_1, \omega_1) \). The subgame after the history \((m_1, a_1, \omega_1)\) is the same as
the static game with prior about the advisor equal to $\lambda_2$. Assume that the unique informative equilibrium is played at every second period history. We can characterize the second period payoff as a function of $\lambda_2$, the reputation at the beginning of period 2.

Recall that

$$E_\omega [u_G (a, \omega)] = E_\omega [u_{DM} (a, \omega)]$$

$$= -(a - Pr \{\omega = 1\})^2 - Pr \{\omega = 1\} Pr \{\omega = 0\},$$

then a good advisor’s second period payoff when second period reputation is $\lambda_2$

$$v_G (\lambda_2) = - Pr \{s = 1\} \left( (a_2 (m = 1) - Pr \{\omega = 1|s = 1\})^2 + Pr \{\omega = 1|s = 1\} Pr \{\omega = 0|s = 1\} \right)$$

$$- Pr \{s = 0\} \left( (a_2 (m = 0) - Pr \{\omega = 1|s = 0\})^2 + Pr \{\omega = 1|s = 0\} Pr \{\omega = 0|s = 0\} \right)$$

$$= - \left[ Pr \{s = 1|\omega = 1\} Pr \{s = 1, \omega = 1\} + Pr \{s = 0, \omega = 1\} Pr \{s = 0, \omega = 0\} \right]$$

$$+ Pr \{s = 1\} \left( \frac{1 - \lambda_2 + 2\lambda_2 \gamma}{2 - \lambda_2} - \gamma \right)^2$$

$$+ Pr \{s = 0\} \left( 1 - \gamma - (1 - \gamma) \right)^2$$

$$= - \left[ \frac{1}{4} \gamma (1 - \gamma) + \frac{1}{4} \gamma (1 - \gamma) + \frac{1}{2} \left( \frac{1 - \lambda_2 + 2\lambda_2 \gamma}{2 - \lambda_2} - \gamma \right)^2 \right]$$

and the value of reputation to a bad advisor is

$$v_B (\lambda_2) = a_2 (m = 1) = \frac{1 - \lambda_2 + 2\lambda_2 \gamma}{2 - \lambda_2}.$$

Because $\frac{1 - \lambda_2 + 2\lambda_2 \gamma}{2 - \lambda_2} < \gamma$, both value functions are increasing in $\lambda_2$. There is value to reputation from the desire to influence DM’s action in the second period.

Note that if the babbling equilibrium is played at every second period history, there is no value to reputation.

### 2.2.2 First Period

An assessment of the whole game is $(\sigma_G (\cdot), \sigma_{G2} (\cdot), \sigma_B (\cdot), \sigma_{B2} (\cdot), \alpha (\cdot), \alpha_2 (\cdot), \Gamma (\cdot), \Lambda (\cdot), \Gamma_2 (\cdot))$ where $\Lambda (m_1, \omega_1, a_1) = \Lambda (m_1, \omega_1)$ is DM’s posterior belief about the advisor after history $(m_1, \omega_1, a_1)$ and $\Gamma_2 (m_1, \omega_1, a_1, m_2)$ is DM’s belief that $\omega_2 = 1$ after receiving message $m_2$ in period 2 after history $(m_1, \omega_1, a_1)$. $\sigma_I (\cdot)$ is the probability type $I$ advisor sends message 1 in period 1, $\alpha (m_1)$ is action DM takes after receiving message $m_1$ in period 1. We have found $\sigma_{G2}, \sigma_{B2}, \alpha_2$ and $\Gamma_2$ as a function of the the reputation at the beginning of period 2, $\Lambda (m_1, \omega_1)$. (That is, in all subgames after histories $h_1 = (m_1, \omega_1, a_1)$ such that the DM’s updated belief about the advisor is $\lambda_2$, players use the same strategy profile.)

The first period strategy in the equilibrium of the two period game in which informative strategy is used in every second period strategy is identical to equilibrium strategy in a static advice game in which DM’s payoff is $-(a_1 - \omega_1)^2$.
and the good advisor’s payoff is

\[-x \left(a_1 - \omega_1\right)^2 + v_G \left(\Lambda \left(m_1, \omega_1\right)\right)\]

and the bad advisor’s payoff is

\[ya_1 + v_B \left(\Lambda \left(m_1, \omega_1\right)\right)\].

If DM believes that \(\omega_1 = 1\) with probability \(\Gamma \left(m_1\right)\), then his best response is to take \(a_1 = \Gamma \left(m_1\right)\). W.l.o.g. assume that \(\Gamma \left(1\right) \geq \Gamma \left(0\right)\). (It’s just relabeling of the two messages.) \(\Gamma \left(m_1\right) = \frac{\frac{1}{2} (\lambda \phi_G (m_1|1) + (1-\lambda) \phi_B (m_1|1))}{\frac{1}{2} (\lambda \phi_G (m_1|1) + (1-\lambda) \phi_B (m_1|1)) + \frac{1}{2} (\lambda \phi_G (m_1|0) + (1-\lambda) \phi_B (m_1|0))} = \frac{1}{1 + \frac{\lambda \phi_G (m_1|1) + (1-\lambda) \phi_B (m_1|1)}{\lambda \phi_G (m_1|0) + (1-\lambda) \phi_B (m_1|0)}}\)

\(\Lambda \left(m_1, \omega_1\right)\) is derived from equilibrium strategy using Bayes update.

Let \(\phi_I \left(m_1|\omega_1\right)\) denote the probability that type \(I\) advisor sends message \(m_1\) conditional on state \(\omega_1\). Then

\[\phi_I \left(1|\omega_1\right) = \gamma \sigma_G \left(s_1 = \omega_1\right) + (1-\gamma) \sigma_G \left(s_1 = 1 - \omega_1\right)\]

and \(\phi_I \left(0|\omega_1\right) = 1 - \phi_I \left(1|\omega_1\right)\). Then

\[\Lambda \left(m_1, \omega_1\right) = \frac{\lambda_1 \phi_G \left(m_1|\omega_1\right)}{\lambda_1 \phi_G \left(m_1|\omega_1\right) + (1-\lambda_1) \phi_B \left(m_1|\omega_1\right)} = \frac{1}{1 + \frac{1-\lambda_1}{\lambda_1} \frac{\phi_B \left(m_1|\omega_1\right)}{\phi_G \left(m_1|\omega_1\right)}}\]

The higher \(\frac{\phi_B \left(m_1|\omega_1\right)}{\phi_G \left(m_1|\omega_1\right)}\) is, the lower the posterior reputation \(\Lambda \left(m_1, \omega_1\right)\) is. Given \(\omega_1\), the message that the bad type is more likely to send relatively to the good type conditional on \(\omega_1\), the lower the posterior reputation is after \((m_1, \omega_1)\). Given \(m_1\), the state at which the bad type is relatively more likely to send this message \(m_1\) than the good type, the lower the posterior reputation is after the realized state.

### 2.2.3 Characterization of an eq in which the good advisor always tells the truth

Want to find conditions under which there exists a PBE in which the good advisor sends message 1 when \(s_1 = 1\) and message 0 when \(s_1 = 0\), i.e. \(\sigma_G \left(1\right) = 1\) and \(\sigma_G \left(0\right) = 0\).

Suppose such an equilibrium exists.

1. \(\sigma_B \left(1\right) = 1\) and \(\sigma_B \left(0\right) > 0\).

First, it cannot be the case that the bad advisor also tells the truth. If so, then it is a pooling equilibrium and posterior \(\Lambda \left(m_1, \omega_1\right) = \lambda\) for all \((m_1, \omega_1)\). But then there is no reputation cost in announcing \(m_1 = 1\). On the other hand, \(\Gamma \left(1\right) = \gamma > \Gamma \left(0\right) = (1-\gamma)\), so DM will take a higher action after message 1. Because bad advisor prefers higher action, there
is strict current incentive to send message 1 and bad advisor will strictly prefer \( m = 1 \) regardless of her signal. Contradiction.

Second, posterior reputation must be higher after announcing 0 regardless of \( \omega_1 \), i.e. \( \Lambda (m_1 = 0, \omega_1 = 1) \geq \Lambda (m_1 = 1, \omega_1 = 1) \) and \( \Lambda (m_1 = 0, \omega_1 = 0) \geq \Lambda (m_1 = 1, \omega_1 = 0) \) and strictly inequality must hold for at least one. Otherwise, the bad advisor strictly prefers announcing 1 and we get a contradiction.

Therefore, \( \phi_B (1 | 1) \geq \phi_B (0 | 1) = \frac{1 - \phi_B (1 | 1)}{\phi_C (1 | 1)} \) and \( \phi_B (1 | 0) \geq \phi_B (0 | 0) = \frac{1 - \phi_B (1 | 0)}{\phi_C (0 | 0)} \), and one holds strictly. These hold iff \( \phi_B (1 | 1) \geq \phi_C (1 | 1) \) and \( \phi_B (1 | 0) \geq \phi_C (1 | 0) \) and one holds strictly. It follows that \( \sigma_B (1) \geq \sigma_G (1) \) and \( \sigma_B (0) \geq \sigma_G (0) \) and one holds strictly.

Since \( \sigma_G (1) = 1, \sigma_B (1) = 1 \). Thus \( \sigma_B (0) > 0 \).

2. \[
\Lambda (m_1 = 1, \omega_1 = 1) = \frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_1) (\gamma + (1 - \gamma) \sigma_B (0))}
\]

This is lower than \( \lambda_1 \).

\[
\begin{align*}
\Lambda (m_1 = 1, \omega_1 = 0) & = \frac{\lambda_1 (1 - \gamma)}{\lambda_1 (1 - \gamma) + (1 - \lambda_1) (1 - \gamma + \gamma \sigma_B (0))} \\
\Lambda (m_1 = 0, \omega_1 = 1) & = \frac{\lambda_1 (1 - \gamma)}{\lambda_1 (1 - \gamma) + (1 - \lambda_1) (1 - \gamma) (1 - \sigma_B (0))} \\
\Lambda (m_1 = 0, \omega_1 = 0) & = \frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_1) \gamma (1 - \sigma_B (0))}.
\end{align*}
\]

One can see that

\[
\Lambda (m_1 = 0, \omega_1 = 1) = \Lambda (m_1 = 0, \omega_1 = 0) > \lambda_1 > \Lambda (m_1 = 1, \omega_1 = 1) > \Lambda (m_1 = 1, \omega_1 = 0).
\]

So even though the good advisor always tells the truth, after sending message 1, the advisor’s reputation declines, even if the advice was proven correct, i.e. \( \omega_1 = 1 = m_1 \). On the other hand, the advisor’s reputation is higher if he sends message 0, regardless of the state. The reputational cost of sending message 1, i.e. \( v_I (\Lambda (m_1 = 0, \omega_1)) - v_I (\Lambda (m_1 = 1, \omega_1)) \), is bigger when \( \omega_1 = 0 \). Given signal \( s_1 \), probability that \( \omega_1 = 0 \) is higher when \( s_1 = 0 \). So reputational costs of sending message 1 are higher when the advisor receives signal 0.

3. For this to be an equilibrium, it has to be optimal for the bad advisor to choose \( \sigma_B (0) \). When the bad advisor gets signal 0, current benefits from
lying are

\[ y (\alpha (1) - \alpha (0)) \]
\[ = y \left( \frac{\gamma + (1 - \lambda_1) (1 - \gamma) \sigma_B (0)}{1 + (1 - \lambda_1) \sigma_B (0)} - (1 - \gamma) \right) \]

which is strictly decreasing in \( \sigma_B (0) \), and her reputation costs from lying are

\[ (1 - \gamma) v_B (\Lambda (m_1 = 1, \omega_1 = 1)) + \gamma v_B (\Lambda (m_1 = 1, \omega_1 = 0)) \]
\[ - (1 - \gamma) v_B (\Lambda (m_1 = 0, \omega_1 = 1)) - \gamma v_B (\Lambda (m_1 = 0, \omega_1 = 0)) \]
\[ = (1 - \gamma) v_B \left( \frac{\lambda_1 \gamma}{\lambda_1 \gamma + (1 - \lambda_1) [\gamma + (1 - \gamma) \sigma_B (0)]} \right) \]
\[ + \gamma v_B \left( \frac{\lambda_1 (1 - \gamma)}{\lambda_1 (1 - \gamma) + (1 - \lambda_1) (1 - \gamma + \gamma \sigma_B (0))} \right) \]
\[ - v_B \left( \frac{\lambda_1}{\lambda_1 + (1 - \lambda_1) (1 - \sigma_B (0))} \right) \]

which are strictly decreasing in \( \sigma_B (0) \). Thus either current benefits ≥ future reputation costs for all \( \sigma_B (0) \in [0, 1] \) and thus \( \sigma_B^* (0) = 1 \), or there exists a unique \( \sigma_B^* (0) \in (0, 1) \) such that equality holds. The unique value of \( \sigma_B^* (0) \) is higher for \( \lambda_1 \) close to 0 or 1 because in that case, reputation doesn’t change much with message, so reputation costs are smaller. It is lower (bad advisor lies less often) for intermediate \( \lambda_1 \).

4. It has to be optimal for the good advisor to tell the truth. When \( s_1 = 0 \), current benefits for sending \( m_1 = 1 \) is negative, and there is future costs of reputation from sending \( m_1 = 1 \). So she strictly prefers \( m_1 = 0 \). When \( s_1 = 1 \), current benefits from telling the truth are positive, but there are reputation costs from telling the truth because announcing 1 results in lower reputation regardless of \( \omega_1 \). It is necessary and sufficient that

\[ x \text{ (current gain from sending } m_1 = 1) \]
\[ \geq \text{ future reputation costs from sending } m_1 = 1. \]

For any parameter, we can find a threshold \( x^* \) such that for \( x > x^* \), the inequality holds and there exists a PBE in which the good advisor always tells the truth, and for \( x < x^* \), it does not hold.

2.2.4 When does there exist an eq in which the good advisor always tells the truth?

It exists if and only if \( x \), the relative weight of the good advisor’s first period payoff, is below some threshold.

So if future is very important, no information can be transmitted in the first period.
2.2.5 The Role of the biased advisor assumption

The assumption that the biased advisor is biased in a commonly known direction is important. If the biased advisor has opposing preference from DM, then no message is "tainted". If the advisor's biases are uncertain, then messages will not be "tainted" in the same way.

2.2.6 What if DM cannot update her belief about the advisor

For example, if the decision maker in the second period does not have access to the first period history.

There are three effects

1. discipline effect: without reputation concern, the bad advisor always announces 1 in the first period. With reputation concern, $\sigma_B(0)$ may be smaller than 0. Shutting down updating is bad.

2. sorting effect: DM learns about the expert after the first period and is influenced less by the advisor if she now believes that the expert is bad with higher probability. Both types of the advisor's second period strategy is independent of DM's belief, so the more informed DM about the advisor's type, the better it is for DM.

3. political correctness: the good advisor may lie in the first period due to reputation concern. bad for DM. When the concern is so big that all equilibria involve babbling in the first period, discipline effect and sorting effect are both lost and shutting down updating would clearly benefit DM.

2.2.7 Babbling always exists

Definition 2 An equilibrium involves babbling in the first period if $\sigma_G(s_1) = \sigma_B(s_1)$ for $s_1 \in \{0, 1\}$, $\sigma(0) = \sigma(1) = \frac{1}{2}$, $\Gamma(0) = \Gamma(1) = \frac{1}{2}$, $\lambda(m_1, \omega_1) = \lambda$ for all $m_1 \in \{0, 1\}$ and $\omega_1 \in \{0, 1\}$.

Lemma 3 There always exist an equilibrium that involves babbling in the first period.

2.2.8 General properties of an equilibrium that does not involve babbling in the first period

Proposition 4 Any equilibrium that does not involve babbling in the first period must be such that

1. $\sigma_G(0) = 0$ and $\sigma_G(1) > 0$.

2. The bad advisor uses message 1 more often than the good advisor; $\sigma_B(1) \geq \sigma_G(1)$ and $\sigma_B(0) \geq \sigma_G(0)$. 

8
3. **Strict reputational incentive to use message 0, i.e.**

\[\Lambda(0, 1) \geq \Lambda(0, 0) > \lambda > \Lambda(1, 1) \geq \Lambda(1, 0)\,.

**Proof.**

**Property** \(\Lambda(0, 1) \geq \Lambda(1, 1)\) and \(\Lambda(0, 0) \geq \Lambda(1, 0)\).

1. Suppose \(\Lambda(0, 1) < \Lambda(1, 1)\) and \(\Lambda(0, 0) < \Lambda(1, 0)\). Then the expected value of reputation is higher after announcing 1 regardless of \(s_1\). Because \(DM\) takes a weakly higher action after \(m_1 = 1\), bad advisor strictly prefers \(m_1 = 1\) and always announces 1. But then an announcement of 0 implies that the advisor is good with probability 1. Contradiction.

2. Suppose \(\Lambda(0, 1) < \Lambda(1, 1)\) and \(\Lambda(0, 0) \geq \Lambda(1, 0)\). When \(\omega_1 = 1\), reputation costs from announcing 1 is negative. When \(\omega_1 = 0\), reputation costs from announcing 1 is positive. Therefore, reputation costs from announcing 1 is lower when \(s_1 = 1\) than when \(s_1 = 0\). But both types of advisor has a weakly higher current benefit from announcing 1 when \(s_1 = 1\) than when \(s_1 = 0\). So when \(s_1 = 1\), current benefits higher, future costs lower than when \(s_1 = 0\). Thus either \(\sigma_I(1) = 1\) or \(\sigma_I(0) = 0\) for both \(I = G, B\). Recall that

\[\Lambda(m_1, \omega_1) = \frac{1}{1 + \frac{1-\lambda_I}{\lambda_I} \frac{\phi_B(m_1|\omega_1)}{\phi_G(m_1|\omega_1)}}\]

Then \(\frac{\phi_B(1|1)}{\phi_G(1|1)} < \frac{\phi_B(0|1)}{\phi_G(0|1)} = \frac{1-\phi_B(1|0)}{1-\phi_G(1|0)}\) and \(\frac{\phi_B(1|0)}{\phi_G(1|0)} \geq \frac{\phi_B(0|0)}{\phi_G(0|0)} = \frac{1-\phi_B(0|0)}{1-\phi_G(0|0)}\). So \(\phi_B(1|1) < \phi_G(1|1)\) and \(\phi_B(1|0) \geq \phi_G(1|0)\). That is,

\[
(1 - \gamma) \sigma_B(0) + \gamma \sigma_B(1) < (1 - \gamma) \sigma_G(0) + \gamma \sigma_G(1) \quad (1)
\]

\[
\gamma \sigma_B(0) + (1 - \gamma) \sigma_B(1) \geq \gamma \sigma_G(0) + (1 - \gamma) \sigma_G(1) \quad (2)
\]

If \(\sigma_G(0) = \sigma_B(0) = 0\), the two inequalities cannot hold at the same time. Likewise for \(\sigma_G(1) = \sigma_B(1) = 1\). If \(\sigma_G(0) = 0\) and \(\sigma_B(1) = 1\), then for inequality 1 to hold, we need \(\sigma_G(1) > 1\) impossible. If \(\sigma_G(1) = 1\) and \(\sigma_B(0) = 0\), then for inequality 2 to hold, we need \(\sigma_B(1) = 1\) and \(\sigma_G(0) = 0\). But then inequality 1 holds with equality, contradiction.

3. Suppose \(\Lambda(0, 1) \geq \Lambda(1, 1)\) and \(\Lambda(0, 0) < \Lambda(1, 0)\). Then expected costs of reputation from announcing 1 are higher when \(s_1 = 1\) than when \(s_1 = 0\). Current benefits from announcing 1 is independent of signal for the bad advisor. Thus either \(\sigma_B(1) = 0\) or \(\sigma_B(0) = 1\). So \(\sigma_B(1) < \sigma_B(0)\) and thus \(\phi_B(1|1) < \phi_B(1|0)\). Similarly, for \(\Lambda(0, 1) \geq \Lambda(1, 1)\) and \(\Lambda(0, 0) < \Lambda(1, 0)\), we need \(\phi_B(1|1) \geq \phi_G(1|1)\) and \(\phi_B(1|0) < \phi_G(1|0)\). So \(\phi_G(1|1) < \phi_G(1|0)\). Thus \(\Gamma(1) < \Gamma(0)\), contradiction.

**Property** \(\Lambda(0, 1) \geq \Lambda(1, 1)\) and \(\Lambda(0, 0) \geq \Lambda(1, 0)\) and at least one inequality holds strictly.
Suppose not. The there are zero reputation costs in announcing 1. If \( \Gamma(1) = \Gamma(0) \), then it is a babbling equilibrium, contradiction. So \( \Gamma(1) > \Gamma(0) \). But then bad advisor strictly prefers to announce 1. And announcing 0 should result in reputation of 1, contradiction.

This property implies that \( \sigma_G(1) \leq \sigma_B(1) \) and \( \sigma_G(0) \leq \sigma_B(0) \) and one inequality holds strictly.

**Property** \( \Gamma(1) > \Gamma(0) \).

Suppose not. Then current benefits from announcing 1 are zero. Future reputation costs from announcing 1 are positive. Thus the bad advisor strictly prefers to announce 0. But then announcing 1 would result in reputation of 1, a higher reputation. Contradiction.

**Property** \( \sigma_G(0) = 0 \).

\( \Gamma(1) > \Gamma(0) \), so the good advisor has a current benefits from announcing 0 when \( s_1 = 0 \) because she prefers a lower action after receiving a low signal. There are future reputation costs from announcing 1. So good advisor strictly prefers 0 when \( s_1 = 0 \).

**Property** \( \Lambda (1, 1) \geq \Lambda (1, 0) \).

After announcing 1, reputation is higher when the message turns out to be correct. It suffices to show that \( \frac{\phi_B(1|1)}{\phi_G(1|1)} \leq \frac{\phi_B(1|0)}{\phi_G(1|0)} \), i.e. the likelihood ratio of an announcement from the bad type to an announcement from the good type is lower when \( \omega = 1 \). This is because \( G \) never announce 1 when signal is 0, and signal is 0 with higher probability when \( \omega = 0 \).

\[
\frac{\phi_B(1|1)}{\phi_G(1|1)} = \frac{\gamma \sigma_B(1) + (1 - \gamma) \sigma_B(0)}{\gamma \sigma_G(1) + (1 - \gamma) \sigma_G(0)} = \frac{\gamma \sigma_B(1) + (1 - \gamma) \sigma_B(0)}{\gamma \sigma_G(1)} = \frac{\sigma_B(1)}{\sigma_G(1)} + \frac{1 - \gamma \sigma_B(0)}{\gamma \sigma_G(1)} < \frac{\sigma_B(1)}{\sigma_G(1)} + \frac{\gamma \sigma_B(0)}{\sigma_G(1)} < \frac{\phi_B(1|0)}{\phi_G(1|0)}.
\]

Intuitively, \( \sigma_G(0) = 0 \) and signal 0 is more likely when \( \omega = 0 \). Thus announcement 1 is even more unlikely to come from a good advisor when state is 0.

**Property** \( \Lambda (0, 1) \geq \Lambda (0, 0) \)
After an announcement of 0, reputation is higher if the message turns out to be incorrect. Suppose not. Then we have \( \Lambda(1,0) \leq \Lambda(1,1) \leq \Lambda(0,1) < \Lambda(0,0) \). So reputation costs from announcing 1 are higher when state is 0. Thus expected reputation costs are higher when signal is 0. Current benefits are independent of signal for the bad advisor. Thus the bad advisor is more likely to announce 0 when \( s_1 = 0 \). Either \( \sigma_B(0) = 0 \) or \( \sigma_B(1) = 1 \). \( \Lambda(0,1) < \Lambda(0,0) \) implies that

\[
\frac{\phi_B(0|1)}{\phi_B(0|0)} > \frac{\phi_B(0|0)}{\phi_B(0|0)},
\]

bad advisor is relatively more likely to announce 0 when state is 1. Or equivalently,

\[
\frac{\phi_B(0|0)}{\phi_B(0|1)} < \frac{\phi_B(0|0)}{\phi_B(0|0)} = \frac{(1-\gamma)(1-\sigma_G(1)) + \gamma}{\gamma(1-\sigma_G(1)) + (1-\gamma)}.
\]

If \( \sigma_B(0) = 0 \), then \( \sigma_B(0) = 0 = \sigma_G(0) \) and \( \sigma_B(1) > \sigma_G(1) \) from the implication of property 2. But then the good advisor is more likely to announce 0 when \( s_1 = 1 \) than a bad advisor. So \( \frac{\phi_B(0|0)}{\phi_B(0|1)} < \frac{\phi_B(0|0)}{\phi_B(0|1)} \), contradiction. Thus \( \sigma_B(1) = 1 \). But then \( \frac{\phi_B(0|0)}{\phi_B(0|1)} = \frac{\gamma}{1-\gamma} > \frac{\phi_B(0|0)}{\phi_B(0|1)} \). The gist is that, the good advisor has more incentive to announce 0. So a message 0 from a good advisor is less informative than a message 0 from the bad advisor.

**Property** For \( \omega \in \{0,1\} \), either \( \Lambda(0,\omega) > \lambda_1 > \Lambda(1,\omega) \) or \( \Lambda(0,\omega) = \lambda_1 = \Lambda(1,\omega) \).

Conditional on \( \omega \), expert is good with probability \( \lambda_1 \). Thus either message does not tell them apart, or \( \Lambda(0,\omega) > \lambda_1 > \Lambda(1,\omega) \).

**Property**

\[
\Lambda(0,1) \geq \Lambda(0,0) > \lambda > \Lambda(1,1) \geq \Lambda(1,0).
\]

If \( \Lambda(0,0) = \lambda_1 = \Lambda(1,0) \), then \( \Lambda(1,1) \geq \Lambda(0,1) = \lambda_1 \) and \( \Lambda(0,1) \geq \Lambda(0,0) = \lambda_1 \). But then \( \Lambda(1,1) = \Lambda(0,1) = \lambda_1 \), contradiction property 2. If \( \Lambda(0,1) = \lambda_1 = \Lambda(1,1) \), then \( \Lambda(1,0) \leq \Lambda(1,1) = \lambda_1 \) and \( \Lambda(0,0) \leq \Lambda(0,1) = \lambda_1 \). But again we must have \( \Lambda(1,0) = \lambda_1 = \Lambda(0,0) \). Contradicting property 2. So \( \Lambda(0,0) > \lambda_1 > \Lambda(1,0) \) and \( \Lambda(0,1) > \lambda_1 > \Lambda(1,1) \) and \( \Lambda(0,1) \geq \Lambda(0,0) \) and \( \Lambda(1,1) \geq \Lambda(1,0) \).

**Property** \( \sigma_G(1) > 0 \).

Suppose not. Then to have \( \Gamma(1) > \Gamma(0) \), we need \( \sigma_B(1) > \sigma_B(0) \). But then an announcement of 1 exposes the advisor’s type to be bad. Because \( \sigma_B(1) > \sigma_B(0) \), we get \( \Lambda(0,1) > \Lambda(0,0) \) because signal 1 is more likely under state 1. But the reputation costs from announcing 1 is higher under state 1, so expected reputation costs are higher under signal 1. Current benefits independent of signal for bad advisor. So \( \sigma_B(1) \leq \sigma_B(0) \), contradiction.
2.3 Short note on existence of eq in which good advisor tells the truth

Suppose DM knows that the advisor is bad and \( \sigma_B (1) = 1 \). What is DM’s posterior that \( \omega = 1 \) given that \( m = 1 \)? DM knows that conditional on the advisor having gotten signal 1, \( \omega = 1 \) with probability \( \gamma \), and \( \Pr \{ \omega = 1 | s_B = 0 \} = 1 - \gamma \).

The message itself gives DM information about the signal the bad advisor has received.

\[
\Pr \{ \omega = 1 | m = 1, B \} = \Pr \{ \omega = 1, s_B = 1 | m = 1, B \} + \Pr \{ \omega = 1, s_B = 0 | m = 1, B \}
\]

\[
= \Pr \{ s_B = 1 | m = 1 \} \Pr \{ \omega = 1 | s_B = 1, B \} + \Pr \{ s_B = 0 | m = 1 \} \Pr \{ \omega = 1 | s_B = 0, B \}
\]

\[
= \Pr \{ s_B = 1 | m = 1 \} \gamma + \Pr \{ s_B = 0 | m = 1 \} (1 - \gamma) .
\]

We have

\[
\Pr \{ s_B = 1 | m = 1 \} = \frac{\Pr \{ m = 1 | s_B = 1 \} \Pr \{ s_B = 1 \}}{\Pr \{ m = 1 | s_B = 1 \} \Pr \{ s_B = 1 \} + \Pr \{ m = 1 | s_B = 0 \} \Pr \{ s_B = 0 \}}
\]

\[
= \frac{\Pr \{ m = 1 | s_B = 1 \} \Pr \{ s_B = 1 \}}{\Pr \{ m = 1 | s_B = 1 \} \Pr \{ s_B = 1 \} + \Pr \{ m = 1 | s_B = 0 \} \Pr \{ s_B = 0 \}}
\]

\[
= \frac{1}{1 + \sigma_B (0)} .
\]

So the more often \( B \) lies, the less likely DM believes that \( B \) has gotten signal 1 when \( B \) sends message 1. Thus the more weight DM puts on \( \Pr \{ \omega = 1 | s_B = 0 \} \), and the lower DM’s posterior about \( \omega \) is after receiving \( m = 1 \) from the bad advisor. In addition, \( \Pr \{ \omega = 1 | m = 1, B \} < \gamma \) as long as \( \sigma_B (0) > 0 \).

But DM doesn’t know whether advisor is bad or good. DM believes that advisor is good with prob \( \lambda_1 \). Thus

\[
\Pr \{ \omega = 1 | m = 1 \} = \Pr \{ G | m = 1 \} \Pr \{ \omega = 1 | m = 1, G \} + \Pr \{ B | m = 1 \} \Pr \{ \omega = 1 | m = 1, B \}
\]

\[
= \frac{1}{2} \frac{\lambda_1}{\lambda_1} \gamma + \left( 1 \frac{1}{2} (1 - \lambda) \sigma_B (0) \right) \Pr \{ \omega = 1 | m = 1, B \} .
\]

The more often \( B \) lies, the less likely message 1 is coming from a good advisor, and the lower DM’s posterior about \( \omega \) is after receiving message 1 from a bad advisor. So the lower DM’s overall belief that \( \omega = 1 \) is after receiving \( m = 1 \).

Benefits of lying for bad advisor is

\[
y * (\alpha (1) - \alpha (0))
\]

\[
= y * (\Pr \{ \omega = 1 | m = 1 \} - (1 - \gamma))
\]
is decreasing in $\sigma_B(0)$, the probability with which DM believes that B lies.

Costs of lying is

$$(1 - \gamma) \left( v_B(\Lambda(m_1 = 0, \omega_1 = 1)) - \sigma_B(0) \right) + \gamma \left( v_B(\Lambda(m_1 = 0, \omega_1 = 0)) - \sigma_B(0) \right)$$

increases $\sigma_B(0)$ because the reputation after saying $m = 0$ increases with $\sigma_B(0)$ and the reputation after saying $m = 1$ decreases with $\sigma_B(0)$.

Given DM’s belief $\sigma_B(0)$, B’s best response probability of lying is

$$\left\{ \begin{array}{ll} 1 & \text{if benefits}(\sigma_B(0)) > \text{costs} \\ [0, 1] & \text{if benefits} = \text{costs} \\ 0 & \text{if benefits} < \text{costs} \end{array} \right.$$  

For $\sigma_B^*(0)$ to form a PBE, it must be a fixed point of this mapping.

References