# CAREERIST EXPERTS
# AND POLITICAL INCORRECTNESS

Chia-Hui Chen
Junichiro Ishida

March 2014

# Careerist Experts and Political Incorrectness

Chia-Hui Chen* and Junichiro Ishida†

March 7, 2014

## Abstract

While political correctness is a dominant norm in many public situations, we also observe behaviors that are apparently "politically incorrect," often from professionals and experts. This paper examines the flip side of political correctness as analyzed in Morris (2001) to shed some light on the elusive notion of political incorrectness and elucidate its equilibrium and welfare properties. We show that there are circumstances in which unbiased experts deliberately take a politically incorrect stance out of reputational concerns and identify key elements which give rise to this perverse reputational incentive. The results suggest that political incorrectness cannot necessarily be viewed as a sign of blunt honesty when informed experts have long-term reputational concerns. We also examine the welfare consequences of political incorrectness and argue that this form of information manipulation can be beneficial under some conditions.

---

*Institute of Economics, Academia Sinica. email: chchen@econ.sinica.edu.tw

†Institute of Social and Economic Research, Osaka University. email: jishida@iser.osaka-u.ac.jp

# 1    Introduction

In a seminal analysis of political correctness, Morris (2001) eloquently shows how the incentive to appear politically correct obstructs truthful information transmission. He makes this point in an environment where an uninformed decision maker needs to solicit advice from an informed expert repeatedly over time. The expert in question may be biased in favor of some politically incorrect alternative, but his predispositions are only privately known. In this dynamic context, there naturally arises an incentive for the agent to present himself as unbiased in the early stages so as to remain credible in the eyes of the decision maker. This reputational incentive is actually self-defeating, however, as it forces the expert to take a politically correct stance regardless of the true state of nature. As such, political correctness generally entails the loss of socially valuable information, illustrating why politically correct opinions are often uninformative and unreliable as a source of knowledge. Thanks in no small part to this contribution, we now have a fairly clear understanding of (at least one form of) political correctness.

The situation contrasts sharply with its counterpart, i.e., political incorrectness, which has received far less attention in the literature. Even then, the lack of attention *per se* is largely inconsequential if we can apply this same line of reasoning to its "flip side" to gain a sense of political incorrectness. To elaborate on this possibility, consider the case where the true state of nature happens to favor a politically correct alternative. In this contingency, the unbiased expert should have no incentive to take a politically incorrect stance against his belief because that can only lower his reputation, not to mention his current payoff. Given this, because any expert who makes a false recommendation is more likely to be perceived as biased, even the biased expert now has a reputational incentive to reveal the true information. Note that, unlike in the case of political correctness, reputational concerns now discipline the expert to be more truthful. According to this reasoning, political incorrectness should be regarded as a sign of blunt honesty, or "intellectual integrity," in environments where reputation matters because an informed expert would take a politically incorrect stance only when he firmly believes in it.[1]

As convincing as it may sound, however, the validity of this conclusion is not necessarily clear. At the very least, the conclusion seems rather too extreme to hold in general,[2] sug-

---

[1] This perception is perhaps exemplified most symbolically by a popular book series The Politically Incorrect Guide which presents conservative or so-called "politically incorrect" views on various topics such as Darwinism, the Constitution, the Bible, and so on.

[2] Although it is certainly not easy to quantify this claim, some people express an even harsher view against political incorrectness in general. For instance, in a widely-read political blog Crooked Timber, John Quiggin is quoted as saying "politically incorrect views are almost always incorrect in every way: literally, scientifically and morally."

1

gesting that there may be a gap to be filled in the aforementioned argument. Particularly suspicious in this regard is an implicit presumption that the reputational effect of political correctness (incorrectness) is invariably weakly positive (negative), which effectively rules out the possibility that the expert intentionally takes a politically incorrect stance out of reputational concerns. We argue that this presumption may trivialize the intricate nature of reputation formation in a dynamic setting because what "reputation" can mean in reality is potentially very broad and diverse, and the expert can gain or lose the decision maker's trust along many different dimensions. In fact, quite contrary to the original intent of the word, political correctness is now often associated with a negative connotation where people who express politically correct views are perceived as manipulative or even dishonest; put it differently, taking a politically correct stance is not necessarily a sure way to improve one's reputation, broadly defined. Given this negative perception, we may have a situation where the unbiased expert strategically deviates from the norm of political correctness to show that he is, at least, not manipulative.

In this paper, we construct a dynamic model of strategic communication to see whether and under what conditions this rough intuition would indeed survive in formal equilibrium analysis. To this end, we extend Morris (2001), which we refer to as the "original setup" for clarity, by incorporating an additional period and an additional expert type to capture a more diverse process of reputation formation. As in the original setup, the expert can be either good (unbiased) or bad (biased): if the expert is good, he has the same payoff function as the decision maker; if bad, he always wants a higher action than the decision maker. On top of these two strategic types, as another key departure from the original setup, we introduce the possibility that the expert may be inherently honest, in which case he simply reveals the true information in every opportunity he comes across.

The sequence of events within each period proceeds as follows. At the beginning of each period, the expert observes the state of nature, which takes a value of either 0 or 1, and sends a cheap-talk message, again 0 or 1, to the decision maker. Upon receiving the message, the decision maker then chooses an action from some continuous interval. The state is publicly observed after the action is taken, and the decision maker updates her belief about the expert's type conditional on all the available information. Without loss of generality, we let message 1 represent the "politically incorrect stance," i.e., the message that induces a higher (more politically incorrect) action, and say that an expert is "politically incorrect" whenever he announces 1.[3]

---

[3]Note that our notion of political correctness is defined in the *ex ante* sense (before the true state is publicly observed). It is also defined in a different way from political correctness in Morris (2001), who takes a much broader view: he defines political correctness as an act of altering what to say in order to avoid adverse

Under this setup, it is not overly surprising to see the bad type occasionally announce 1 in state 0 because he can always derive a current benefit from inducing a higher action. It is a totally different story, however, if the unbiased good type ever chooses to do so in equilibrium for some strategic reasons. For the analysis, we label this particular form of political incorrectness as *anti-political correctness*, and say that anti-political correctness arises in equilibrium whenever the good type announces 1 in state 0 with any positive probability. Since the good type derives no current benefit from misreporting in any situation, the emergence of anti-political correctness means that there must be a reputational gain from falsely announcing 1 in state 0. We emphasize this notion of anti-political correctness because it necessarily yields a profound impact on our interpretation of political incorrectness: without anti-political correctness in equilibrium, all the reputational forces would point in one direction, only working to discipline the expert to be more truthful as we discussed at the outset; if it could ever be supported in equilibrium, on the other hand, the reputational effect of political incorrectness would be reversed, resulting in qualitatively different outcomes and implications.

We obtain several results concerning the equilibrium and welfare properties of political incorrectness. First, we derive a necessary condition for our notion of anti-political correctness in a fairly general environment, and then establish as a corollary of this result that anti-political correctness can never arise in any two-period variant of our model, including the original setup, even with the addition of the honest type. There are roughly two reasons for this. First, in any two-period model, there is only one opportunity for the expert to establish his reputation. Second, since the good type and honest types are strategically equivalent in the final period, it is unambiguously better to be perceived as good *for both of the strategic types*. These two features altogether imply that there is only one route through which the expert can gain the decision maker's trust: the good type would like to separate from the bad type whereas the bad type would like to pool with the good type. Under this static reputational structure, there is simply no room for supporting anti-political correctness in equilibrium, suggesting that it inherently calls for a dynamic process of reputation formation which offers more routes to be taken by the expert.

By contrast, the two extensions of the current model allow us to depart from the dichotomous structure of the original setup and substantially enrich the process of reputation formation. With the additional period, the process is dynamic, as the expert now has multiple opportunities to build his reputation. Moreover, in the presence of the honest type, there is a potential reputational gain from announcing 0 in state 0 because that allows the

inferences.

3

expert to pool with the honest type. If the bad type's incentive to pool with the honest type is sufficiently strong, the good type may be able to separate from the bad type by falsely announcing 1 in state 0. We show that there indeed exists an equilibrium which supports this form of anti-political correctness and identify its sufficient condition. This result indicates that to the extent that these conditions hold, we cannot regard political incorrectness naively as a sign of blunt honesty since it can easily be an attempt to signal one's hidden characteristics rather than the true state of the world.

Aside from this equilibrium analysis, we also briefly discuss welfare implications of anti-political correctness with emphasis on the sorting effect it generates. Note that in evaluating the welfare impact, Morris (2001) focuses on pooling equilibria where political correctness attracts experts into a false message. In its worst case, only the babbling equilibrium survives, conveying no information about the expert's preference type despite the good type's initial intention to separate from the bad. The situation is clearly different for the case of anti-political correctness because it is a form of separating equilibrium which conveys some useful information for future interactions. Early separation of types, induced by anti-political correctness, can indeed be socially beneficial when the decision maker's expected payoff is sufficiently convex with respect to her posterior belief, i.e., in environments where precise information is more valuable. We obtain some sufficient conditions for anti-political correctness to be welfare-enhancing even though it is yet another form of information manipulation.

The paper proceeds as follows. In the remainder of this section, we briefly review the related literature. Section 2 outlines the basic model. Section 3 analyzes the general model to show that anti-political correctness never emerges in a two-period model, including the original setup, and identify key necessary conditions for anti-political correctness. Section 4 considers a more specific environment, called the model with asymmetric states, to obtain sufficient conditions for anti-political correctness and derive welfare implications. Finally, section 5 provides some concluding remarks.

**Related Literature:** The paper is broadly related to the literature which illustrates perverse effects of career concerns, initiated by Holmstrom and Ricart i Costa (1986) and Holmstrom (1999). There are roughly two strands of this literature. One strand is concerned with the case where agents attempt to signal their competence, such as the ability to acquire or interpret information, through their actions or payoff-irrelevant (cheap-talk) messages. Scharfstein and Stein (1990) show that managers ignore their private information and instead simply mimic the investment decisions of other managers under some conditions. Prendergast (1993) provides a model of "yes men" and argues that workers have an incentive to conform to the opinion of their supervisors when firms can only use subjective performance evaluation.

There are now many works along this line, showing that agents sometimes fail to make full use of their private information when they are concerned about their reputations.[4]

By contrast, this paper falls into the other strand where agents attempt to signal their payoff congruence or "unbiasedness." Earlier examples of this strand are Sobel (1985) and Benabou and Laroque (1992), who consider a repeated cheap-talk game where the sender's preference type is not known to the receiver. Both of these works assume that the sender is either honest or strategic and focus on the behavior of the strategic type. Morris (2001), on which our model is based, considers a different setting in which the sender is either good (unbiased) or bad (biased), while both types are strategic. Similarly, Ely and Valimaki (2003) consider a situation where a long-run player interacts with a sequence of short-run players who are uncertain about the long-run player's preference type. As in Morris (2001), the long-run player is either good (unbiased) or bad (biased) and attempts to signal his payoff congruence though payoff-relevant actions.[5] In terms of the type space, the current analysis can be regarded as a hybrid of these previous settings as we consider two strategic types (good and bad) as well as a commitment type (honest).

From a broader perspective, our notion of anti-political correctness can be interpreted as a form of contrarian behavior. There are now many works which analyze the origin and nature of conformism (or "herding") in economic environments (Akerlof, 1980; Banerjee, 1992; Bikhchandani et al., 1992; Bernheim, 1994). To a large extent, Morris (2001) and Ely and Valimaki (2005) are also models of conformism where agents conform to a certain standard of behavior. By contrast, our model can be seen as a model of contrarianism where (some) agents deliberately deviate from a social norm out of reputational concerns. Although the issue of contrarianism has received relatively less attention in the literature, there is a handful of works which examine why agents sometimes divert, or "anti-herd," from the majority (Avery and Chevalier, 1999; Levy 2004, 2005). The key aspect of these works is that agents differ in the accuracy of information which allows some talented ones to separate by moving against the herd to signal their competence.[6] This contrasts sharply with our model where agents move against a social norm in an attempt to signal their payoff congruence. We provide a more detailed account of this interpretation in section 4.3.

---

[4]Other examples include Prendergast and Stole (1996), Ottaviani and Sorensen (2001, 2006), Visser and Swank (2007), and Fu and Li (2014), just to name some.

[5]Ely and Valimaki (2003) consider two cases, one in which the bad type is a commitment type who always chooses a certain stage-game strategy and the other in which he is a standard strategic player, but the main conclusion is essentially the same.

[6]Effinger and Polborn (2001) consider a different setup in which agents do not know their own type. In this setup, they derive a benefit function such that an agent is most valuable if he is the only smart one, so that career concerns are not individualistic, and show that a form of contrarianism emerges in this case.

## 2  Model

**Environment:** Consider a three-period advice game with an uninformed decision maker and an expert who is perfectly informed about the current state of the world $\omega_t \in \{0,1\}$.[7] For simplicity, we assume that each state occurs equally likely with probability one half. In each period $t = 1, 2, 3$, the decision maker solicits information about the current state of the world from the expert and chooses an action $a_t \in \mathbb{R}$. Let $m_t \in \{0, 1\}$ denote the expert's message, or his "opinion," about the state in period $t$. After the action is chosen, the true state is realized and publicly observed.

**Decision maker:** In each period $t$, the decision maker's utility $u^{\omega_t}(a_t)$ depends on the realized state $\omega_t$ and his action $a_t$, where $u^{\omega_t}(a_t)$ is differentiable, concave in $a_t$ (with at least one of $u^0(a_t)$ and $u^1(a_t)$ being strictly concave), and attains a unique maximum for each $\omega_t$. Let

$$a^*(\omega) := \arg\max_a u^\omega(a) \in \mathbb{R} \cup \{-\infty, \infty\},$$

denote the optimal action under full information when the true state is known to be $\omega$. Without loss of generality, we assume $a^*(1) > a^*(0)$, so that the decision maker prefers a (weakly) higher action when the state is $\omega_t = 1$. The decision maker's total utility is given by

$$\sum_{t=1}^{3} x_t u^{\omega_t}(a_t),$$

where $x_t > 0$ is the weight attached to the payoff in period $t$. The weight is meant to capture the player's time preference as well as the salience of the decision problem at hand.

**Expert:** As in the original setup, the nature of information asymmetry is two dimensional. Aside from the true state, the expert privately observes his own type $\theta \in \{G, B, H\}$, which can be good ($\theta = G$), bad ($\theta = B$), or honest ($\theta = H$). The expert type is drawn only once at the beginning and is time-invariant. If the expert is good, he has a utility function identical to the decision maker's, so that their interests are perfectly aligned. If the expert is bad, on the other hand, he always prefers a (weakly) higher action than the decision maker, regardless of the true state, where his instantaneous utility is always given by $u^1(a_t)$.[8] The

---

[7]This is another difference from the original setup which assumes that the expert observes a possibly noisy signal. By contrast, we assume that the expert can perfectly identify the true state. The assumption is made strictly to simplify the analysis and does not affect our results in any qualitative way.

[8] Note that the payoff functions are specified in a slightly different, and more flexible, way from the original setup which, for most part, assumes $u^\omega(a) = -(a - \omega)^2$ for the good type and $u^\omega(a) = a$ for the bad type. The essence of the problem is still preserved with the current specification because the decision maker never takes an action that is too high for the bad type.

total utility of the bad type can then be expressed as

$$\sum_{t=1}^{3} y_t u^1(a_t),$$

where $y_t > 0$ is the weight attached to the payoff in period $t$. For expositional clarity, we often refer to the good and bad types as strategic. In addition to these two strategic types, we introduce an additional type, the honest, who commits to revealing the true information in every opportunity he comes across.[9] Let $\lambda_{1\theta}$ denote the prior probability that the expert is of type $\theta$ at the beginning of period 1. The prior type distribution is common knowledge.

**Political and anti-political correctness:** It is commonly understood that the bad type always has a current incentive to announce 1 regardless of the true state. To separate from the bad type, there arises a reputational incentive to announce 0 even when the true state is 1. This intention is the so-called "political correctness," where the expert intentionally lies and takes the politically correct stance (announces 0), and is analyzed extensively in Morris (2001). In this analysis, on the other hand, we focus on the possibility, which we call "anti-political correctness," where the expert announces 1 in state 0 purely out of reputational concerns. Note that the good type falsely announces 1 only when there is a reputational gain because he derives no current benefit from lying in any state. The emergence of anti-political correctness is then equivalent to the following phenomenon.

**Definition 1** *An equilibrium with anti-political correctness is the one in which the good type chooses $m_t = 1$ when $\omega_t = 0$ with any strictly positive probability.*

## 3   The general model

In this section, we consider the model with general payoff functions. The goal of this section is to establish generally that anti-political correctness cannot arise in a two-period model, including the original setup, and to illustrate why. In the course of this analysis, we also obtain necessary conditions for anti-political correctness to be sustained in equilibrium, which will be used to show its existence in the following section.

### 3.1   Preliminary: strategies and beliefs

**The expert's strategy:** The expert's strategy in each period is a mapping from the observed current state $\omega_t$ into a probability distribution over messages. Let $\mu_{t\theta}^{\omega}$ denote the expert's

---

[9]There are at least two distinct ways to interpret this honest type: in one way, the honest type can be regarded as the myopic version of the good expert with no weight on the future payoffs; in the other, it can also be the one with a sufficiently large cost of lying, be it psychological or reputational.

(type-contingent) strategy, which is defined as the probability that the type $\theta$ expert tells the truth in period $t$ given the state $\omega_t = \omega$. By assumption, $\mu_{tH}^\omega = 1$ for $\omega = 0, 1$. Let $\mu_{t\theta} := (\mu_{t\theta}^0, \mu_{t\theta}^1)$ and $M_t := (\mu_{tG}, \mu_{tB})$.

**The decision maker's strategy:** At the beginning of each period $t$, the decision maker comes in with a certain belief about the expert's type. We denote this belief by $\lambda_{t\theta}$, which indicates the probability that the expert is of type $\theta$ at the beginning of period $t$. Since $\lambda_{tG} + \lambda_{tB} + \lambda_{tH} = 1$, we let $\Lambda_t := (\lambda_{tG}, \lambda_{tH})$ represent the belief about the expert's type, which thoroughly captures his "reputation."

Without loss of generality, we focus on a class of equilibria in which the decision maker chooses a (weakly) higher action after receiving $m_t = 1$. Upon receiving a message $m_t$, the decision maker forms a belief about the current state defined as

$$p_t := \text{prob}(\omega_t = 1 \mid m_t; M_t, \Lambda_t).$$

for a given set of strategies $M_t$ and the current belief $\Lambda_t$. We in general write the belief as $p_t = \pi_t^{m_t}(M_t, \Lambda_t)$, or simply $\pi_t^{m_t}$, where

$$\pi_t^0 = \frac{\lambda_{tG}(1 - \mu_{tG}^1) + (1 - \lambda_{tH} - \lambda_{tG})(1 - \mu_{tB}^1)}{\lambda_{tH} + \lambda_{tG}(\mu_{tG}^0 + (1 - \mu_{tG}^1)) + (1 - \lambda_{tH} - \lambda_{tG})(\mu_{tB}^0 + (1 - \mu_{tB}^1))}, \tag{1}$$

$$\pi_t^1 = \frac{\lambda_{tH} + \lambda_{tG}\mu_{tG}^1 + (1 - \lambda_{tH} - \lambda_{tG})\mu_{tB}^1}{\lambda_{tH} + \lambda_{tG}(\mu_{tG}^1 + (1 - \mu_{tG}^0)) + (1 - \lambda_{tH} - \lambda_{tG})(\mu_{tB}^1 + (1 - \mu_{tB}^0))}. \tag{2}$$

The decision maker chooses an action $a_t$ that is optimal for her based on this belief $p_t$. We let $a^*(p_t)$ represent the decision maker's optimal action, which is given by

$$a^*(p) = \arg\max_a \ (1 - p)u^0(a) + pu^1(a).$$

Under the maintained assumptions, one can easily show that $a^*$ is continuous and increasing in $p$.[10] To save notation, we often write $a_t^m := a^*(\pi^m(M_t, \Lambda_t))$. In addition, we define $\overline{u}^\omega := u^\omega(a^*(\omega))$ as the payoff under complete information, and $\underline{u}^\omega := u^\omega(a^*(\frac{1}{2}))$ as the payoff under no information so that any feasible payoff is bounded in $[\underline{u}^\omega, \overline{u}^\omega]$.

**The next-period belief:** At the end of period $t = 1, 2$, the decision maker updates her belief about the expert's type after observing the realized state $\omega_t$. The decision maker's next-period belief is then determined by the realized state $\omega_t$, the message $m_t$, the set of strategies $M_t$ and the current belief $\Lambda_t$. We thus denote the next-period belief by $\lambda_{t+1\theta}^{\omega_t,m_t}(M_t, \Lambda_t)$, or simply $\lambda_{t+1\theta}^{\omega_t,m_t}$.

---

[10]The fact that $u(a_t, 0)$ and $u(a_t, 1)$ are concave implies that $a^*(p) \in [a^*(0), a^*(1)]$ is increasing. To ensure that $a^*(p)$ increases continuously, we assume that at least one of the two functions, $u(a_t, 0)$ and $u(a_t, 1)$, is strictly concave.

## 3.2 The third period

The problem in the final period is quite straightforward with no future interactions. Following Morris (2001), we restrict our attention to the unique informative equilibrium. In the absence of any reputational concerns, it is easy to see that the bad type always announces 1 while the good type always announces the true state. Under this set of strategies and the belief $\Lambda_3$, we can compute

$$\pi_3^0 = 0, \ \pi_3^1 = \frac{1}{2 - \lambda_3},$$

where $\lambda_t := \lambda_{tG} + \lambda_{tH}$. From this, we obtain the value functions for the two strategic types as follows:

$$v_{3G}(\lambda_{3G}, \lambda_{3H}) = \frac{x_3}{2}(u^0(a^*(0)) + u^1(a^*(\tfrac{1}{2-\lambda_3}))),$$

$$v_{3B}(\lambda_{3G}, \lambda_{3H}) = y_3(u^1(a^*(\tfrac{1}{2-\lambda_3})).$$

There are two key properties of the continuation payoffs that prove to be crucial. First, $\lambda_{3G}$ and $\lambda_{3H}$ are perfect substitutes in that $v_{3\theta}(\ell, \lambda_3 - \ell)$ is constant for any $1 \geq \lambda_3 \geq \ell \geq 0$. Second, because the honest type is always credible, both value functions are monotonically increasing in $\lambda_t \in [0, 1]$, thereby giving rise to the incentive to stay on the good side of the decision maker in the earlier stages. As we will see in section 3.4, these two properties play an important role in ruling out the existence of anti-political correctness when the model has only two periods.

## 3.3 The second period

The second-period problem is clearly more complicated because the good type, with reputational concerns, may now behave differently from the honest type. Define $v_{t\theta}^{\omega,m} := v_{t\theta}(\lambda_{t+1G}^{\omega,m}, \lambda_{t+1H}^{\omega,m})$, which thoroughly captures the expert's reputational concerns. Taking the expert's and decision maker's strategies as given, the good type chooses $m_2 = 1$ with positive probability only if

$$x_2 u^{\omega_2}(a_2^1) + v_{3G}^{\omega_2,1} \geq x_2 u^{\omega_2}(a_2^0) + v_{3G}^{\omega_2,0}. \tag{3}$$

Similarly, the bad type chooses $m_2 = 1$ with positive probability only if

$$y_2 u^1(a_2^1) + v_{3B}^{\omega_2,1} \geq y_2 u^1(a_2^0) + v_{3B}^{\omega_2,0}. \tag{4}$$

Each type's payoff in period $t = 1, 2$ can be decomposed into two parts. The first part is the *current payoff*, $u^{\omega_t}(a_t)$, which is determined by the match between the current state and the action chosen by the decision maker. The second part is the *continuation payoff*, $v_{t+1\theta}^{\omega_t, m_t}$, which

9

is determined by the match between the current state and the message sent by the expert. The tradeoff, if any, is clear in this environment. In state 1, the good type unambiguously derives a lower current payoff by announcing 0 in any informative equilibrium, but that may improve his reputation (a higher continuation payoff). If the increase in the continuation payoff is large enough, the good type may lie and announce 0 to protect his reputation. Now that the reputational cost of political incorrectness is even higher, the bad type may also follow the good type by announcing 0 more frequently. As argued in Morris (2001), this entire process is self-defeating, and no information may be conveyed in the worst case.

The following proposition shows that even in this extended setup with the honest type and with more general (though slightly altered) payoff functions, any informative (nonbabbling) equilibrium in period 2 still has the properties characterized in Morris (2001).

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

1. $\mu_{2G}^0 = 1$, *that is, the good type always tells the truth when the state is* 0.

2. $\mu_{2B}^1 \geq \mu_{2G}^1$ *and* $\mu_{2B}^0 \leq \mu_{2G}^0$, *that is, the bad type announces* 1 *more often than the good type.*

3. $\lambda_3^{1,0} \geq \lambda_3^{0,0} \geq \lambda_3^{1,1} \geq \lambda_3^{0,1}$ *with* $\lambda_3^{1,0} > \lambda_3^{1,1}$ *for any* $\Lambda_2$.

**Proof:** See Appendix A.

The most important result of the proposition is the first property which immediately implies the following.

**Corollary 1** *Anti-political correctness never arises when there are only two periods. In other words, an expert can only tarnish his reputation by falsely announcing* 1 *in state* 0 *in a two-period model.*

### 3.4 Necessary conditions for anti-political correctness

An equilibrium with anti-political correctness inherently requires that the bad type announce 0 in state 0 more frequently than the good type, so that there is a reputational gain for the good type to deviate from the politically correct message. At a glance, the presence of the honest type appears to give an additional incentive for the bad type to announce 0 in state 0 because that allows the bad type to pool with the honest. If the bad type chooses to do so with sufficient frequency, the good type may be able to separate from the bad type by announcing 1. Proposition 1 suggests, however, that it is generally infeasible to have this type

of separating equilibrium when there are only two periods. Before we proceed to analyze the first-period problem, we would like to discuss briefly why anti-political correctness cannot be supported as an equilibrium outcome in a two-period framework, even with the addition of the honest type.

To see why the presence of the honest type alone is not sufficient, we first establish the following statement.

**Lemma 1** *Suppose that from period $t$ on, (i) the good type always reports truthfully, and (ii) the bad type's period $t$ strategy is independent of the belief $\Lambda_t$. Then,*

$$(v_{tG}^{\omega,1} - v_{tG}^{\omega,0})(v_{tB}^{\omega,1} - v_{tB}^{\omega,0}) \geq 0, \tag{5}$$

*for any $(\lambda_{tG}^{0,1}, \lambda_{tH}^{0,1})$ and $(\lambda_{tG}^{1,1}, \lambda_{tH}^{1,1})$.*

PROOF: It suffices to show that $v_{tG}(a,b) - v_{tG}(a',b')$ always has the same sign as $v_{tB}(a,b) - v_{tB}(a',b')$. First, condition (i) directly implies that the good type is strategically equivalent to the honest type, so that $v_{t\theta}(a,b) = v_{t\theta}(a+b,0)$.[11] Second, conditions (i) and (ii) together imply that $v_{t\theta}(a+b,0)$ is monotonically (and weakly) increasing in $a+b$ ($v_{t\theta}(a+b,0)$ is constant with respect to $a+b$ if the bad type also always reports truthfully). Then, if $v_{t\theta}(a+b,0) > v_{t\theta}(a'+b',0)$, it must be $a+b > a'+b'$ because of the monotonicity. We thus have $v_{t\theta}(a,b) - v_{t\theta}(a',b') > (<)0$ for both $\theta = G, B$ if and only if $a+b > (<)a'+b'$. ∎

The property (5) means that if the reputational gain from switching to a different message is positive for one type, it must also be positive for the other type, thereby always inducing the two strategic types to move in the same direction. The intuition behind this result is as follows. First, under condition (i), the good type is strategically equivalent to the honest type, and the value function hence depends only on $\lambda_t = \lambda_{tG} + \lambda_{tH}$. The model is thus effectively reduced to the one with two expert types where the expert's reputation is thoroughly captured by a single-dimensional variable $\lambda_t$. Given this property, condition (ii) then suggests that the value function is (weakly) increasing in $\lambda_t$, i.e., a higher $\lambda_t$ is always more desirable. These two properties together imply that there is a reputational gain from announcing 1 in state $\omega$ for either type if and only if $\lambda_t^{\omega,1} > \lambda_t^{\omega,0}$.

It is fairly straightforward to verify that it is generally not feasible to have anti-political correctness when (5) is satisfied.

**Proposition 2** *Anti-political correctness cannot be supported as an equilibrium outcome in period $t-1$ if the two conditions in Lemma 1 are satisfied.*

---

[11] Note that this strategic equivalence is irrelevant in the original setup where the expert's reputation is by construction captured by a single-dimensional variable.

PROOF: Note first that a necessary condition for anti-political correctness is that given the set of strategies and beliefs, we have

$$v_{3G}^{0,1} - v_{3G}^{0,0} > 0.$$

If this holds in equilibrium, however, it follows from (5) that the bad type must have a strict incentive to announce $m_2 = 0$, so that $\mu_{2B}^0 = 0$ and $\lambda_{3G}^{0,0} + \lambda_{3H}^{0,0} = 1$. This results in a contradiction because the good type must then have a strict incentive to deviate to $m_2 = 0$.∎

This proposition illustrates why anti-political correctness cannot arise in a two-period environment, including the original setup. First, the good type always reports truthfully in the final period. Second, because the bad type always sends message 1 regardless of the current state, his strategy is also independent of his reputation. In the unique informative equilibrium which prevails in the final period, therefore, the two conditions in the lemma are necessarily satisfied, so that the reputational gain from announcing 1 always has the same sign between the two strategic types. This means that if there is ever a reputational incentive for the good type to announce 1 in state 0, the incentive must be even stronger for the bad type to do so because he can also derive a current benefit by doing so. This totally eliminates any reputational gain, thereby leaving no room for anti-political correctness when there are only two periods, regardless of whether we have the honest type or not.

## 3.5 The first period

To focus on the possibility of anti-political correctness, we restrict our attention to the case where the true state happens to be $\omega_1 = 0$. Taking the expert's and decision maker's strategies as given, the good type chooses $m_1 = 1$ with positive probability (anti-political correctness) only if

$$x_1 u^0(a_1^1) + v_{2G}^{0,1} \geq x_1 u^0(a_1^0) + v_{2G}^{0,0}.$$

Similarly, taking the expert's and decision maker's strategies as given, the bad type chooses $m_1 = 1$ with positive probability only if

$$y_1 u^1(a_1^1) + v_{2B}^{0,1} \geq y_1 u^1(a_1^0) + v_{2B}^{0,0}.$$

The following result summarizes a necessary condition for anti-political correctness.

**Proposition 3** *Given the set of strategies and beliefs, there exists an equilibrium with anti-political correctness only if*

$$(v_{2G}^{\omega,1} - v_{2G}^{\omega,0})(v_{2B}^{\omega,1} - v_{2B}^{\omega,0}) < 0, \tag{6}$$

*or, equivalently, either one of the two conditions in Lemma 1 fails to hold in period 2.*

**Proof:** Since $u^0(a^*(0)) > u^0(a^*(1))$, a necessary condition for anti-political correctness is

$$v_{2G}^{0,1} - v_{2G}^{0,0} > 0,$$

that is, lying must result in a higher continuation payoff for the good type. Now, if

$$v_{2B}^{0,1} - v_{2B}^{0,0} \geq 0,$$

the bad type has a strict incentive to send $m_1 = 1$. Then, for any $\mu_{1G}^{0,1} > 0$, we have $\lambda_{2G}^{0,0} + \lambda_{2H}^{0,0} = 1$, in which case the expert's continuation payoff is maximized. Since this is a contradiction, this means that we must have $v_{2B}^{0,1} - v_{2B}^{0,0} < 0$, so that the bad expert does not have a strict incentive to lie when $\omega_1 = 0$. The second condition follows directly from Lemma 1. ∎

Note that, because the good type may now behave differently from the honest type in period 2, the relative values of being perceived as good and honest in general differ across the two strategic types. To see this more clearly, given some set of equilibrium strategies, we can compute the value functions as follows:

$$v_{2G}(\lambda_{2G}, \lambda_{2H}) = \frac{1}{2} \sum_{\omega=0}^{1} \Big( \mu_{2G}^{\omega}(x_2 u^{\omega}(a^*(\pi_2^{\omega})) + v_{3G}^{\omega,\omega}) + (1 - \mu_{2G}^{\omega})(x_2 u^{\omega}(a^*(\pi_2^{1-\omega})) + v_{3G}^{\omega,1-\omega}) \Big),$$

$$v_{2B}(\lambda_{2G}, \lambda_{2H}) = \frac{1}{2} \sum_{\omega=0}^{1} \Big( \mu_{2B}^{\omega}(y_2 u^1(a^*(\pi_2^{\omega})) + v_{3B}^{\omega,\omega}) + (1 - \mu_{2B}^{\omega})(x_2 u^1(a^*(\pi_2^{1-\omega})) + v_{3B}^{\omega,1-\omega}) \Big),$$

where $\pi_2^0$ and $\pi_2^1$ are given by (1) and (2), respectively. It is clear from these that $\lambda_{2G}$ and $\lambda_{2H}$ are no longer perfect substitutes when $\mu_{2G}^{\omega} < 1$ for some $\omega$. As a consequence, we may have a situation in which the bad type prefers to pool with the honest type by announcing 0, while the good type prefers to separate from them by announcing 1. We dedicate the next section to exploring this possibility.

# 4 The model with asymmetric states

## 4.1 Existence of anti-political correctness

The previous section has identified a necessary condition for anti-political correctness to emerge, but it is not necessarily clear from the argument whether or not an equilibrium with anti-political correctness actually exists. Since adding one more period makes the analysis

substantially more complicated, and it is highly tedious, if possible at all, to obtain a full characterization of our general three-period setup, we here introduce an additional assumption which makes the model substantially more tractable. With this additional assumption, we obtain a sufficient condition for anti-political correctness and show that anti-political correctness can indeed be supported as an equilibrium outcome in our extended setup.

**Assumption 1** $\bar{u}^1 \to \infty$, *i.e., the maximum attainable payoff in state 1 is unbounded.*

**Lemma 2** *Suppose that* $\lambda_{2H} = 0$ *and* $\lambda_{2G} \in [0,1)$. *Then, no information is revealed in period 2 under Assumption 1.*

**Proof:** See the Appendix.


Assumption 1 clarifies our focus in two different ways. First, the uniqueness of the continuation equilibrium when $\lambda_{2H} = 0$ substantially simplifies the analysis and sharpens our predictions, which is especially important in a dynamic setup like ours. Second, and more importantly, with this assumption, the good type now has a stronger incentive to lie in the second period following $m_1 = 1$. This feature is very crucial because, as we have noted above, anti-political correctness requires (6) which is satisfied when the good type lies in some contingency and hence behaves differently from the honest type. This feature of the model immensely helps in illustrating this underlying logic.

The assumption also illuminates the essential role played by the presence of the honest type. The following proposition shows that the presence of the honest type is indeed necessary for the emergence of anti-political correctness in this setup with asymmetric states.

**Proposition 4** *Under Assumption 1, there exists no equilibrium with anti-political correctness if* $\lambda_{1H} = 0$.

**Proof:** Note first that it is not possible to achieve full separation in the model with asymmetric states because if $\mu_{2G}^{0,m} = 1$ for some $m_1 = m$, the bad expert can obtain an unbounded payoff by deviating to $m_1 = m$. Given that the continuation equilibrium in period 2 must be babbling when $\lambda_{2G} \in (0,1)$, the value functions for both types must be strictly increasing in $\lambda_{2G}$. We then have

$$(v_{2G}^{0,1} - v_{2G}^{0,0})(v_{2B}^{0,1} - v_{2B}^{0,0}) \geq 0,$$

It thus follows from Proposition 3 that anti-political correctness is not feasible in the model with asymmetric states when $\lambda_{1H} = 0$. ∎

It should be noted here that this result relies heavily on the uniqueness of the continuation equilibrium guaranteed by Lemma 2, but does not hold in general when there are multiple continuation equilibria. We show in Appendix B that it is possible to construct an equilibrium with anti-political correctness without the honest type when there are multiple continuation equilibria, because the value function $v_{2\theta}(\lambda_{2G}, 0)$ need not be monotonic in this case. We do not place much emphasis on this possibility, however, because the constructed equilibrium in Appendix B is sustained by strategies that are highly arbitrary and hence does not carry much economic insight.

We are now ready to obtain a sufficient condition under which the equilibrium with anti-political correctness exists. The following is the main result of the paper.

**Proposition 5** *Under Assumption 1, given $x_1$, $y_1$, $x_2$, $y_2$, and $y_3$, there exist $\bar{\lambda}$ and $\bar{x}$ such that an equilibrium with anti-political correctness exists if $\lambda_{1H} > \bar{\lambda}$ and $x_3 > \bar{x}$. In the equilibrium, the good type always announces 1 in period 1 while the bad type randomizes between the two messages.*

**Proof:** See the Appendix.

The proposition shows that there exists at least one route though which anti-political correctness can be supported in equilibrium. To see what happens in this equilibrium, suppose that the good type always announces 1 (anti-political correctness) in period 1. This leaves the bad type with two alternatives, either to announce 0 and pool with the honest type or to announce 1 and pool with the good type. The reputational gain from the former generally dominates the latter because the honest type is always credible, while the good type is not in period 2. The bad type thus has an incentive to announce 0 with a higher probability (while he still randomizes) not to separate from the honest type. The clustering of the bad type on message 0 then provides the good type with a different incentive because he can now separate from the bad type by announcing 1. Although this generally results in a lower payoff in period 2 because of the ensuing babbling equilibrium, the good type can capitalize on his good reputation in period 3. It can hence be optimal for the good type to announce 1 in period 1 if he cares enough about his "long-run" reputation.

## 4.2 Welfare

As in the original setup, reputational concerns lead to the loss of socially valuable information which by itself must be welfare-reducing. However, anti-political correctness also necessarily entails a positive sorting effect because it is a type of separating equilibrium which generates

15

some information about the expert's predispositions. The overall welfare effect is hence ambiguous, depending on numerous environmental factors in a complicated manner. Here, we show through a more specific example that anti-political correctness can indeed enhance welfare, despite the loss of information in period 1.

We evaluate the welfare properties of the model by the decision maker's expected payoff. Following Morris (2001), the welfare benchmark is the equilibrium with no reputational updating, i.e., the one in which the decision maker's belief remains at $\Lambda_1$. One can think of this as a case where there is a different decision maker in each period who is unable to observe the expert's past choices. Without reputational concerns, the equilibrium of this game takes a very simple form: the bad type always claims $m_t = 1$, while the good type always tells the truth. Since there is no reputational updating, the decision maker's expected payoff is the same across all three periods.

The situation draws a clear contrast to the equilibrium with anti-political correctness in which the good type separates from the bad type in period 1. Moreover, in the continuation game following $m_1 = 0$, the bad type is only pooled with the honest type. Because the bad type is perceived to be more credible and the decision maker responds more to his (possibly misrepresented) message in period 2, he tends to be more aggressive and more frequently announces 1 in state 0. Roughly speaking, therefore, the equilibrium with anti-political correctness induces early separation of types, so that the decision maker could end up with a clearer idea about the expert's type and realize a larger payoff in the end. In general, anti-political correctness is more beneficial when the decision maker's payoff depends on $\lambda_3$ in a convex manner.

We need to consider a more tightly specified model to evaluate the welfare impact of anti-political correctness more precisely. Consider an environment in which the payoff functions are specified as follows:

$$u^0(a) = -a^2 \text{ and } u^1(a) = a. \tag{7}$$

Note that the maximum payoff is unbounded in state 1, so that this specification satisfies Assumption 1. Under these payoff functions, the optimal action choice in any given period is computed as

$$a^*(p_t) = \frac{p_t}{2(1 - p_t)}.$$

In particular, since $\pi_3^0 = 0$ and $\pi_3^1 = \frac{1}{2 - \lambda_3}$, the decision maker's expected payoff in period 3

as a function of $(\lambda_{3G}, \lambda_{3H})$ can be obtained as

$$v_{3G}(\lambda_{3G}, \lambda_{3H}) = x_3 \left( \lambda_3 \frac{u^0(a^*(0)) + u^1(a^*(\pi_3^1))}{2} + (1 - \lambda_3) \frac{u^0(a^*(\pi_3^1)) + u^1(a^*(\pi_3^1))}{2} \right)$$
$$= \frac{x_3}{8(1 - \lambda_3)}, \tag{8}$$

The third-period payoff is hence convex in $\lambda_3$ and diverges to infinity as $\lambda_3 \to 1$, meaning that early separation of types can be highly beneficial in this type of environment.

Intuitively, anti-political correctness is welfare-improving when the relative weight for the third-period payoff is sufficiently large. To make this point, consider the limit case where both $x_1$ and $x_2$ approach zero, so that welfare is determined entirely by the third-period payoff. In addition, let $y_1 \to 0$ for simplicity. In this example, we can show that anti-political correctness can be supported as an equilibrium outcome if $\frac{y_3}{y_2}$ is sufficiently close to zero. In the equilibrium, the good type always claims $m_1 = 1$, while the bad type randomizes in period 1. Moreover, as $\frac{y_3}{y_2}$ approaches zero, the probability of the bad type announcing 0 in period 1 converges to one, allowing the good type to almost fully separate from the bad type. Then, as can be seen from (8), this constitutes a sufficient condition for anti-political correctness to be welfare-improving. We summarize this finding as follows.

**Proposition 6** *Suppose that the payoff functions are given by (7) and let $x_1, x_2, y_1 \to 0$. Then, the equilibrium with anti-political correctness exists and is welfare-improving if $\frac{y_3}{y_2}$ is sufficiently close to zero.*

**Proof:** See the Appendix.

## 4.3 Anti-political correctness as a form of contrarianism

Our notion of anti-political correctness can be regarded as a form of contrarian behavior or "anti-herding," and in this sense, the current analysis has some connection to the literature on conformism/contrarianism. In this section, we interpret our analysis from this perspective and briefly discuss its relation to the existing literature.

The existing literature offers at least two approaches to analyzing conformism. One approach, which is perhaps more conventional in economics, is informational, where an agent imitates the decisions of his predecessors even when his private information indicates otherwise, because actions taken by other agents provide relevant information (Banerjee, 1992; Bikhchandani et al., 1992). By contrast, the other approach assumes that agents somehow benefit from adhering to a code of behavior, or a "social norm," and emphasizes its role in inducing convergence of actions (Akerlof, 1980; Bernheim, 1994). The current analysis is more closely related to this latter notion of conformism.

As a prime example of analyses of conformism, Bernheim (1994) considers a situation where agents care about status as well as intrinsic utility. An agent's status is assumed to depend on public perceptions of his predispositions rather than on his actions themselves. In this setup, it is shown that agents with moderate preferences converge to a homogeneous standard of behavior, which can be regarded as a social norm, in order to avoid an inference that they have undesirable extreme preferences. The key aspects of his analysis are:

- There are some preference types or "predispositions" that are deemed undesirable.

- The preference types are not directly observable, which must instead be inferred from actions.

Under this setting, a social norm is defined as a set of actions which would not be taken by those undesirable types in the absence of any reputational concerns. Conformism then arises when agents abide by a particular norm and abstain from those extreme actions in order to avoid an inference that they possess undesirable predispositions.[12] According to this view, Morris (2001) and Ely and Valimaki (2003) are also models of conformism which share this feature, although they endogenize the value of reputation through repeated interactions, as we do here.[13]

In the current analysis as well as in Morris (2001), by this definition, a norm is to announce 0 because: (i) it is unambiguously undesirable to be perceived as bad; (ii) the bad type would never announce 0 in the absence of any reputational concerns. In clear contrast to Bernheim (1994) and Morris (2001), our analysis can then be seen as a model of contrarianism because agents deliberately deviate from the norm, by taking an action which would be intrinsically preferred by the bad type, i.e., announcing 1. What is interesting here is that contrarianism in this context is sustained by conformism on the part of the bad type: more forward-looking agents deviate from a social norm to differentiate from more myopic counterparts who follow the norm to secure "short-run" reputational gains. We argue that the current analysis departs from the previous literature which focuses on the case where agents attempt to signal their competence and provides a new rationale for a different form of contrarianism through which agents attempt to signal their payoff congruence.

_____

[12]To be more precise, Bernheim (1994) considers a setting in which the type space is $[0, 2]$, which represents an agent's bliss point, where types closer to either end are deemed undesirable by assumption. It is further assumed that there is some exogenous esteem function which has a unique maximum at 1. Agents then have an incentive to abstain from actions close to either end, inducing those with moderate preferences to converge towards some middle point.

[13]In Morris (2001) and Ely and Valimaki (2003), biased-preference types are deemed undesirable although the reputation cost is derived endogenously.

## 5  Conclusion

In his seminal analysis of political correctness, Morris (2001) argues that reputational concerns may induce informed experts to take a politically correct stance even when it is not the right thing to do. In this example, therefore, reputational concerns are the source of information manipulation and eventually result in the loss of socially valuable information. By the same token, one can conjecture that reputational concerns can discipline experts to be more truthful when they are supposed to take a politically correct stance. If this is indeed the case, political incorrectness should be regarded as a sign of honesty which can be taken at (almost) face value.

In this paper, to take a step towards disentangling and better understanding the general nature of political incorrectness, we provide an analytical framework which scrutinizes this rough intuition. We show that the intuition does not always hold true, as unbiased experts may deliberately take a politically incorrect stance to prove their good intentions when they are sufficiently concerned about their long-run reputations. In light of this result, we argue that political incorrectness is not necessarily a sign of honesty in long-term relationships with sufficiently strong career concerns. Moreover, we also show that a form of political incorrectness can be welfare-improving because it induces early separation of types.

As a final remark, it is important to note that our analysis sheds light on only one aspect of political incorrectness, and there are many different sides of this issue which cannot be captured by the current framework. In the future, it is of interest to see more works in this direction.

## References

Akerlof, G., 1980, A Theory of Social Custom, of Which Unemployment May Be One Consequence, *Quarterly Journal of Economics*, 94, 749-75.

Avery, C.N. and J. Chevalier, 1999, Herding over the Career, *Economics Letters*, 63, 327-33.

Banerjee, A.J., 1992, A Simple Model of Herd Behavior, *Quarterly Journal of Economics*, 107, 797-817.

Benabou, R. and G. Laroque, 1992, Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility, *Quarterly Journal of Economics*, 107, 921-58.

Bernheim, B.D., 1994, A Theory of Conformity, *Journal of Political Economy*, 102, 841-77.

Bikhchandani, S, D. Hirshleifer and I. Welch, 1992, A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades, *Journal of Political Economy*, 100, 992-1026.

Effinger, M.R. and M.K. Polborn, 2001, Herding and Anti-Herding: A Model of Reputational Differentiation, *European Economic Review*, 45, 385-403.

Ely, J.C. and J. Valimaki, 2003, Bad Reputation, *Quarterly Journal of Economics*, 118, 785-814.

Fu, Q. and M. Li, 2014, Reputation-Concerned Policy Makers and Institutional Status Quo Bias, *Journal of Public Economics*, 110, 15-25.

Holmstrom, B, 1999, Managerial Incentive Problems: A Dynamic Perspective, *Review of Economic Studies*, 66, 169-82.

Holmstrom, B. and J. Ricart i Costa, 1986, Managerial Incentives and Capital Management, *Quarterly Journal of Economics*, 101, 835-60.

Levy, G., 2004. Anti-Herding and Strategic Consultation, *European Economic Review*, 48, 503-25.

Levy, G., 2005, Careerist Judges and the Appeal Process, *RAND Journal of Economics*, 2005, 275-97.

Morris, S., 2001, Political Correctness, *Journal of Political Economy*, 109, 231-65.

Ottaviani, M. and P.N. Sorensen, 2001, Information Aggregation in Debate: Who Should Speak First? *Journal of Public Economics*, 81, 393-421.

Ottaviani, M. and P.N. Sorensen, 2006, Reputational Cheap Talk, *RAND Journal of Economics*, 37, 155-75.

Prendergast, C., 1993, A Theory of 'Yes Men,' *American Economic Review*, 83, 757-70.

Prendergast, C. and L. Stole, 1996, Inpetuous Youngsters and Jaded Old Timers: Acquring a Reputation for Learning, *Journal of Political Economy*, 104, 1105-34.

Scharfstein, D.S. and J.C. Stein, 1990, Herd Behavior and Investment, *American Economic Review*, 80, 465-79.

Sobel, J., 1985, A Theory of Credibility, *Review of Economic Studies*, 52, 557-73.

Visser, B. and O.H. Swank, 2007, On Committees of Experts, *Quarterly Journal of Economics*, 112, 337-72.

# Appendix A: the proofs

PROOF OF PROPOSITION 1: Define $\lambda_3^{\omega_2, m_2} := \lambda_{3G}^{\omega_2, m_2} + \lambda_{3B}^{\omega_2, m_2}$. It is straightforward to compute

$$\lambda_3^{1,1} = \frac{\lambda_{2H} + \lambda_{2G}\mu_{2G}^1}{\lambda_{2H} + \lambda_{2G}\mu_{2G}^1 + (1 - \lambda_{2H} - \lambda_{2G})\mu_{2B}^1}, \tag{9}$$

$$\lambda_3^{1,0} = \frac{\lambda_{2G}\left(1 - \mu_{2G}^1\right)}{\lambda_{2G}\left(1 - \mu_{2G}^1\right) + (1 - \lambda_{2H} - \lambda_{2G})\left(1 - \mu_{2B}^1\right)}, \tag{10}$$

$$\lambda_3^{0,1} = \frac{\lambda_{2G}\left(1 - \mu_{2G}^0\right)}{\lambda_{2G}\left(1 - \mu_{2G}^0\right) + (1 - \lambda_{2H} - \lambda_{2G})\left(1 - \mu_{2B}^0\right)}, \tag{11}$$

$$\lambda_3^{0,0} = \frac{\lambda_{2H} + \lambda_{2G}\mu_{2G}^0}{\lambda_{2H} + \lambda_{2G}\mu_{2G}^0 + (1 - \lambda_{2H} - \lambda_{2G})\mu_{2B}^0}. \tag{12}$$

First, we show that $\lambda_3^{0,0} \geq \lambda_3^{0,1}$, $\mu_{2G}^0 = 1$, and $\mu_{2B}^0 \leq \mu_{2G}^0$. Suppose on the contrary that $\lambda_3^{0,0} < \lambda_3^{0,1}$. Then, when $\omega_2 = 0$, the bad type earns a higher payoff by announcing 1, and hence $\mu_{2B}^0 = 0$. However, by (11) and (12), $\lambda_3^{0,0} < \lambda_3^{0,1}$ implies that

$$\lambda_{2H}(1 - \lambda_{2H} - \lambda_{2G})(1 - \mu_{2B}^0) < \lambda_{2G}(1 - \lambda_{2H} - \lambda_{2G})(\mu_{2B}^0 - \mu_{2G}^0),$$

and hence, if $\mu_{2B}^0 = 0$, $\mu_{2G}^0 < 0$, a contradiction. Therefore, $\lambda_3^{0,0} \geq \lambda_3^{0,1}$. It follows from (3) and (4) that $\mu_{2G}^0 = 1$ and $\mu_{2B}^0 \leq \mu_{2G}^0$, which implies $\lambda_3^{0,1} = 0$.[14]

Next, we show that $\lambda_3^{1,0} \geq \lambda_3^{1,1}$ and $\mu_{2B}^1 \geq \mu_{2G}^1$. Again, suppose on the contrary that $\lambda_3^{1,0} < \lambda_3^{1,1}$. If this is the case, both the good and bad types can obtain a higher payoff by announcing 1 when $\omega_2 = 1$, so that $\mu_{2B}^1 = \mu_{2G}^1 = 1$ and $\lambda_3^{1,1} = \lambda_{2H} + \lambda_{2G}$. To compute $\lambda_3^{1,0}$, we follow Morris (2001) and consider the limit of a sequence of perturbed games, in which the expert is misinformed about the state with probability $\varepsilon$. Given $\mu_G^0 = 1$, as $\varepsilon \to 0$, we have

$$\begin{aligned}
\lambda_3^{1,0} &= \lim_{\varepsilon \to 0} \frac{\varepsilon\left[\lambda_{2H} + \lambda_{2G}\mu_G^0\right]}{\varepsilon\left[\lambda_{2H} + \lambda_{2G}\mu_G^0 + (1 - \lambda_{2H} - \lambda_{2G})\mu_B^0\right]} \\
&= \frac{\lambda_{2H} + \lambda_{2G}}{\lambda_{2H} + \lambda_{2G} + (1 - \lambda_{2H} - \lambda_{2G})\mu_B^0} > \lambda_3^{1,1},
\end{aligned}$$

which is a contradiction. Given this, we have $\lambda_3^{1,0} \geq \lambda_3^{1,1}$. By (9) and (10),

$$\lambda_{2H}\left(1 - \lambda_{2H} - \lambda_{2G}\right)\left(1 - \mu_{2B}^1\right) \leq \lambda_{2G}\left(1 - \lambda_{2H} - \lambda_{2G}\right)\left(\mu_{2B}^1 - \mu_{2G}^1\right),$$

which implies $\mu_{2B}^1 \geq \mu_{2G}^1$.[15] Therefore, $\lambda_3^{1,0} \geq \lambda_3^{1,1}$ and $\mu_{2B}^1 \geq \mu_{2G}^1$.

Finally, we show that $\lambda_3^{1,0} \geq \lambda_3^{0,0} \geq \lambda_3^{1,1} \geq \lambda_3^{0,1}$ and $\lambda_3^{1,0} > \lambda_3^{1,1}$ through the following four steps:

---

[14] $(\omega_2 = 0, m_2 = 1)$ is off the equilibrium path when $\mu_{2G}^0 = \mu_{2B}^0 = 1$, and we assume $\lambda_3^{0,1} = 0$ in this case.

[15] As shown above, $(\omega_2 = 1, m_2 = 0)$ is off the equilibrium path when $\mu_{2G}^1 = \mu_{2B}^1 = 1$, and we assume $\lambda_3^{1,0} = 1$ in this case.

1. If $\mu_{2B}^0 < 1$, $\lambda_3^{0,1} = 0$, meaning that $\lambda_3^{1,1} \geq 0 = \lambda_3^{0,1}$. If $\mu_{2B}^0 = 1$, $\mu_{2G}^0 = 1$ too, so we have

$$\lambda_3^{1,1} \geq \lambda_3^{0,1} = 0,$$

where $\lambda_3^{0,1}$ is as defined in footnote 14. Therefore, $\lambda_3^{1,1} \geq \lambda_3^{0,1}$.

2. Let $q_\omega$ denote the probability of observing $m_2 = 1$ in state $\omega_2 = \omega$. Since

$$q_0 \lambda_3^{0,1} + (1 - q_0)\lambda_3^{0,0} = q_1 \lambda_3^{1,1} + (1 - q_1)\lambda_3^{1,0}, \tag{13}$$

it must be the case that $\lambda_3^{1,1} \leq \lambda_3^{0,0}$.

3. We show that we cannot have $\lambda_3^{0,0} > \lambda_3^{1,0} \geq \lambda_3^{1,1} \geq \lambda_3^{0,1}$. Suppose on the contrary that this is the case. Then, by (4), the bad type has a stronger incentive to announce 0 when $\omega_2 = 0$. This means that $\mu_{2B}^0 \geq 1 - \mu_{2B}^1$, and more specifically either $\mu_{2B}^0 = 1$ or $\mu_{2B}^1 = 1$. If $\mu_{2B}^0 = 1$, it is implied that $q_0$ in (13) is 0. But then, (13) can never hold under the presupposed condition. If $\mu_{2B}^1 = 1$, $\lambda_3^{1,0} = 1 \geq \lambda_3^{0,0}$, and therefore we have a contradiction. This shows that $\lambda_3^{1,0} \geq \lambda_3^{0,0} \geq \lambda_3^{1,1} \geq \lambda_3^{0,1}$.

4. Lastly, we prove that $\lambda_3^{1,0} > \lambda_3^{1,1}$. Suppose that $\lambda_3^{1,0} = \lambda_3^{1,1}$, which implies that $\lambda_3^{1,0} = \lambda_3^{0,0} = \lambda_3^{1,1}$. If $\mu_{2G}^1 = \mu_{2B}^1 = 1$, then as defined in footnote 15, $\lambda_3^{1,0} = 1 > \lambda_3^{1,1} = \lambda_{2H} + \lambda_{2G}$, a contradiction. Next consider the case where $\mu_{2G}^1 < 1$. Given $\lambda_3^{1,0} = \lambda_3^{0,0} = \lambda_3^{1,1}$, (13) implies that either $\lambda_3^{0,1} = \lambda_3^{0,0}$ or $q_0 = 0$, which can only be satisfied when $\mu_{2B}^0 = 1$, and therefore, $\lambda_3^{0,1} = 0$ by footnote 14, and $\lambda_3^{1,0} = \lambda_3^{0,0} = \lambda_3^{1,1} = \lambda_{2H} + \lambda_{2G}$. Since $\mu_{2G}^1 < 1$ and $\mu_{2G}^0 = \mu_{2B}^0 = 1$, $\pi_2^1 > \pi_2^0$ and $a^*(\pi_2^1) > a^*(\pi_2^0)$. But then, when $\omega_2 = 1$, the good type has a strict incentive to choose $m_2 = 1$, so $\mu_{2G}^1 = 1$, a contradiction. We thus have $\lambda_3^{1,0} > \lambda_3^{1,1}$.

∎

PROOF OF LEMMA 2: Suppose that $\omega_2 = 1$. The necessary conditions for truth telling are

$$x_2 u^1(a^*(\pi_2^1)) + v_{3G}^{1,1} \geq x_2 u^1(a^*(\pi_2^0)) + v_{3G}^{1,0},$$

$$y_2 u^1(a^*(\pi_2^1)) + v_{3B}^{1,1} \geq y_2 u^1(a^*(\pi_2^0)) + v_{3B}^{1,0},$$

where

$$v_{3G}^{1,m} = \frac{x_3}{2}(\overline{u}_G^0 + u_G^1(a^*(\tfrac{1}{2-\lambda_3^{1,m}}))), \quad v_{3B}^{1,m} = y_3 u_B(a^*(\tfrac{1}{2-\lambda_3^{1,m}})).$$

These conditions can thus be written as

$$\frac{2x_2}{x_3}(u^1(a^*(\pi_2^1)) - u^1(a^*(\pi_2^0))) \geq u^1(a^*(\tfrac{1}{2-\lambda_3^{1,0}})) - u^1(a^*(\tfrac{1}{2-\lambda_3^{1,1}})),$$

$$\frac{y_2}{y_3}(u^1(a^*(\pi_2^1)) - u^1(a^*(\pi_2^0))) \geq u^1(a^*(\tfrac{1}{2-\lambda_3^{1,0}})) - u^1(a^*(\tfrac{1}{2-\lambda_3^{1,1}})).$$

Since $\lambda_3^{1,0} > \lambda_3^{1,1}$ in any informative equilibrium, when $\lambda_{2H} = 0$, we must have $\mu_{2B}^1 = \mu_{2G}^1 = 1$ or $\mu_{2B}^1 > \mu_{2G}^1$. When $\mu_{2B}^1 > \mu_{2G}^1$, we generically have either $\mu_{2B}^1 = 1$ or $\mu_{2G}^1 = 0$.[16]

It follows from Proposition 1 that if the continuation equilibrium in the second period is informative, $\mu_{2G}^0 = 1$. Suppose first that $\mu_{2G}^0 > 0$ which also implies $\mu_{2B}^1 = 1$. Note that if $\mu_{2B}^0 = 1$, then $\pi^1 = 1$, $\pi^0 < 1$, $\lambda_3^{0,0} < 1$ and $\lambda_3^{0,1} = 0$. The bad type then has a strict incentive to announce 1 if $\bar{u}^1$ is sufficiently large. Since this is a contradiction, we must have $\mu_{2B}^0 < 1$ and $\pi^1 < 1$. However, since $\mu_{2B}^1 = 1$, $\lambda_3^{1,0} = 1$, which means that the bad type can ensure a payoff of $\bar{u}^1$ in the third period by announcing $m_2 = 0$. We must therefore have $\mu_{2B}^1 < 1$ if $\bar{u}^1$ is sufficiently large, a contradiction. This shows that we cannot have $\mu_{2G}^0 > 0$ in any informative equilibrium.

The only remaining possibility is $\mu_{2G}^0 = 1$ and $\mu_{2G}^1 = 0$, but we can also rule out this possibility. Given that $\mu_{2G}^0 = 1$ and $\mu_{2G}^1 = 0$, any informative equilibrium requires that $\mu_{2B}^1 > 1 - \mu_{2B}^0$, i.e., the bad type must have a stronger incentive to announce 1 in state 1. Since the bad type's incentive compatibility constraint can be written as

$$\frac{y_2}{y_3}(u^1(a^*(\pi_2^1)) - u^1(a^*(\pi_2^0))) \geq u^1(a^*(\tfrac{1}{2-\lambda_3^{\omega,0}})) - u^1(a^*(\tfrac{1}{2-\lambda_3^{\omega,1}})),$$

for $\omega = 0, 1$, this condition implies that

$$u^1(a^*(\tfrac{1}{2-\lambda_3^{0,0}})) - u^1(a^*(\tfrac{1}{2-\lambda_3^{0,1}})) \geq u^1(a^*(\tfrac{1}{2-\lambda_3^{1,0}})) - u^1(a^*(\tfrac{1}{2-\lambda_3^{1,1}})),$$

Given that $\mu_{2G}^0 = 1$ and $\mu_{2G}^1 = 0$, $\lambda_3^{\omega,1} = 0$ for $\omega = 0, 1$,[17] and this condition is reduced to

$$u^1(a^*(\tfrac{1}{2-\lambda_3^{0,0}})) \geq u^1(a^*(\tfrac{1}{2-\lambda_3^{1,0}})).$$

This is a contradiction, however, because $\mu_{2B}^1 > 1 - \mu_{2B}^0$ implies that $\lambda_3^{1,0} > \lambda_3^{0,0}$. This means that the equilibrium must be babbling if $\bar{u}^1$ is sufficiently large.

---

[16]When $\frac{y_2}{y_3} = \frac{2x_2}{x_3}$, both of the strategic types may adopt mixed strategies where any combination of $(\mu_{2B}^1, \mu_{2G}^1)$ is feasible. We will ignore this non-generic case throughout the proof.

[17]$(\omega_2 = 0, m_2 = 1)$ may be off the equilibrium path, but we define $\lambda_3^{0,1} = 0$ in this case (see footnote 14). Alternatively, we can also show that $\lambda_3^{0,1} = 0$ by considering the limit of a sequence of perturbed games (as in the proof of Proposition 1) in which the expert is misinformed about the current state with probability $\varepsilon$.

This result proves that there only exists the babbling equilibrium when $\lambda_{2H} = 0$ and $\lambda_{2G} \in [0, 1)$, if the attainable payoff in state 1 is unbounded. ∎

PROOF OF PROPOSITION 5: We construct an equilibrium in which the good type always chooses $m_1 = 1$ in state $\omega_1 = 0$. Note that, given this strategy, the bad type must randomize between the two messages because there would always be a profitable deviation otherwise. Let $\hat{\mu}_{2\theta}^{\omega_2}$ denote the strategy in the continuation game following $m_1 = 0$, and $\hat{\mu}_{2\theta} := (\hat{\mu}_{2\theta}^0, \hat{\mu}_{2\theta}^1)$. Similarly, let $\hat{\lambda}_{3\theta}^{\omega_2, m_2}$ and $\hat{v}_{3\theta}^{\omega_2, m_2}$ denote the third-period belief and third-period value function, respectively, contingent on $(\omega_2, m_2)$ in the continuation game following $m_1 = 0$, and $\hat{\lambda}_3^{\omega_2, m_2} := \hat{\lambda}_{3G}^{\omega_2, m_2} + \hat{\lambda}_{3H}^{\omega_2, m_2}$.

Given that $\lambda_{1G} > 0$, $\mu_{1G}^0 = 0$, and $\mu_{1B}^0 \in (0, 1)$, both $\pi_1^0$ and $\pi_1^1$ are continuous in $\mu_{1B}^0 \in (0, 1)$ and, moreover, are bounded by a number less than 1. Let $\bar{p}$ be the upper bound. Moreover there exists $\lambda'$ such that

$$\pi_1^1 > \frac{1}{2} > \pi_1^0,$$

for $\lambda_{1H} > \lambda'$. Note also that $\lambda_{2G}^{0,0} = 0$, $\lambda_{2H}^{0,0} > \frac{\lambda_{1H}}{1-\lambda_{1G}}$, $\lambda_{2G}^{0,1} > \frac{\lambda_{1G}}{1-\lambda_{1H}}$, and $\lambda_{2H}^{0,1} = 0$. We now need to consider two possibilities, depending on $m_1$.

1. The continuation game following $m_1 = 0$.

Note first that the continuation equilibrium in period 2 following $m_1 = 0$ must be informative.[18] Given that there are only the honest and bad types, the bad type has no incentive to lie in state 1, so that $\hat{\mu}_{2B}^1 = 1$.[19] Furthermore, $\hat{\mu}_{2B}^0 \in (0, 1]$; otherwise, there exists a profitable deviation for the bad type if $\hat{\mu}_{2B}^0 = 0$. We then obtain

$$\pi_2^0 = 0, \ \pi_2^1 = \frac{1}{\lambda_{2H}^{0,0} + (1 - \lambda_{2H}^{0,0})(2 - \hat{\mu}_{2B}^0)}.$$

The beliefs regarding the expert's type are given

$$\hat{\lambda}_3^{1,0} = 0, \ \hat{\lambda}_3^{1,1} = \lambda_{2H}^{0,0}, \ \hat{\lambda}_3^{0,0} = \frac{\lambda_{2H}^{0,0}}{\lambda_{2H}^{0,0} + (1 - \lambda_{2H}^{0,0})\hat{\mu}_{2B}^0}, \ \hat{\lambda}_3^{0,1} = 0.$$

The value function for the bad type is now computed as

$$v_{2B} = \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( \mu_{2B}^\omega (y_2 u^1(a^*(\hat{\pi}_2^\omega)) + \hat{v}_{3B}^{\omega,\omega}) + (1 - \mu_{2B}^\omega)(y_2 u^1(a^*(\hat{\pi}_2^{1-\omega})) + \hat{v}_{3B}^{\omega,1-\omega}) \right), \quad (14)$$

---

[18]Suppose that the continuation equilibrium following $m_1 = 0$ is also babbling. If that is the case, then the second-period payoff is constant regardless of $m_1$ for both types since the continuation equilibrium following $m_1 = 1$ is also babbling (see Lemma 2). Anti-political correctness cannot then arise because it is effectively reduced to a two-period model.

[19]We assume that since the honest type never lies, $\lambda_{3H}^{1,0} = 0$ even though it is off the equilibrium path.

where $\hat{\pi}_2^{\omega_2} := \pi_2^{\omega_2}(\hat{\mu}_{2G}, \hat{\mu}_{2B}; \lambda_{2G}^{0,0}, \lambda_{2H}^{0,0})$ and $\hat{v}_{3\theta} := v_{3\theta}(\hat{\lambda}_{3G}^{\omega_2, m_2}, \hat{\lambda}_{3H}^{\omega_2, m_2})$. Since the bad type feels indifferent between the two messages in state 0 if $\hat{\mu}_{2B}^1 \in (0,1)$, (14) can be written as

$$v_{2B} = \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( y_2 u^1(a^*(\hat{\pi}_2^\omega)) + \hat{v}_{3B}^{\omega,\omega} \right). \tag{15}$$

Now suppose that the good type deviates and sends $m_1 = 0$ in the first period. Given the continuation equilibrium following $m_1 = 0$, $\hat{\mu}_{2G}^1 = \hat{\mu}_{2G}^0 = 1$. The value function for the good type is then obtained as

$$v_{2G} = \frac{1}{2} \sum_{\omega = \{0,1\}} \left( x_2 u^\omega(a^*(\hat{\pi}_2^\omega)) + \hat{v}_{3G}^{\omega,\omega} \right).$$

2. The continuation game following $m_1 = 1$.

By Lemma 2, we know that the continuation equilibrium in period 2 must be babbling. We thus have

$$\pi_2^0 = \pi_2^1 = \frac{1}{2}, \ \lambda_3 = \lambda_{2G}^{0,1}.$$

The value functions are given, respectively, by

$$v_{2B} = y_2 \underline{u}^1 + v_{3B}(\lambda_{2G}^{0,1}, 0), \tag{16}$$

$$v_{2G} = \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( x_2 \underline{u}^\omega + v_{3G}(\lambda_{2G}^{0,1}, 0) \right),$$

where $\underline{u}^\omega$ is the minimum payoff in state $\omega$ as defined earlier (the payoff under no information).

We are now ready to prove the proposition. We first compare the bad type's expected second-period payoffs in (15) and (16):

$$\frac{1}{2} y_2 \left( u^1(a^*(\frac{1}{\lambda_{2H}^{0,0} + (1 - \lambda_{2H}^{0,0})(2 - \hat{\mu}_{2B}^0)})) + u^1(a^*(0)) \right) \text{ and } y_2 \underline{u}^1.$$

Given that $u^1(a^*(\lambda))$ increases unboundedly in $\lambda$, there exists a $\bar{\lambda} \geq \lambda'$ such that if $\lambda_{2H}^{0,0} > \frac{\bar{\lambda}}{1 - \lambda_{1G}}$,

$$\frac{1}{2} y_2 \left( u^1(a^*(\frac{1}{\lambda_{2H}^{0,0} + (1 - \lambda_{2H}^{0,0})(2 - \hat{\mu}_{2B}^0)})) + u^1(a^*(0)) \right) > y_1 u^1(a^*(\bar{p})) + y_2 \underline{u}^1. \tag{17}$$

Next, note that if $\lambda_{1H} > \bar{\lambda}$, then $\lambda_{2H}^{0,0} > \frac{\lambda_{1H}}{1 - \lambda_{1G}}$ and $\lambda_{2G}^{0,1} = \frac{\lambda_{1G}}{1 - \frac{\lambda_{1H}}{\lambda_{2H}^{0,0}}}$. Therefore, as $\lambda_{2H}^{0,0}$ increases from $\frac{\lambda_{1H}}{1 - \lambda_{1G}}$ to 1, $\lambda_{2G}^{0,1}$ decreases from 1 to $\frac{\lambda_{1G}}{1 - \lambda_{1H}}$, the value of (15) increases to $\infty$,

and the value of (16) decreases from $\infty$. This means that there exist $\lambda_{2H}^{0,0} \in (\frac{\lambda_{1H}}{1-\lambda_{1G}}, 1)$ and $\lambda_{2G}^{0,1} = \frac{\lambda_{1G}}{1-\lambda_{1H}-\frac{\lambda_{2H}^{0,0}}{\lambda_{1H}}}$ such that the bad type is indifferent between the two messages. That is, given $\lambda_{2H}^{0,0}$ and the corresponding $\mu_{1B}^0$,

$$y_1 u^1(a^*(\pi_1^0)) + \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( y_2 u^1(a^*(\hat\pi_2^\omega)) + \hat v_{3B}^{\omega,\omega} \right) = y_1 u^1(a^*(\pi_1^1)) + y_2 \underline{u}^1 + v_{3B}(\lambda_{2G}^{0,1}, 0).$$

It follows from (17) that

$$v_{3B}(\lambda_{2G}^{0,1}, 0) > \frac{1}{2} \sum_{\omega \in \{0,1\}} \hat v_{3B}^{\omega,\omega}.$$

This in turn implies that

$$v_{3G}(\lambda_{2G}^{0,1}, 0) > \frac{1}{2} \sum_{\omega \in \{0,1\}} \hat v_{3G}^{\omega,\omega}.$$

Furthermore, if $x_3$ is large enough,

$$x_1 u^0(a^*(\pi_1^0)) + \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( x_2 u^\omega(a^*(\hat\pi_2^\omega)) + \hat v_{3G}^{\omega,\omega} \right) < x_1 u^0(a^*(\pi_1^1)) + \frac{1}{2} \sum_{\omega \in \{0,1\}} \left( x_2 \underline{u}^\omega + v_{3G}(\lambda_{2G}^{0,1}, 0) \right).$$

Therefore, when $\lambda_{1H}$ and $x_3$ are large enough, the good type chooses $m_1 = 1$ while the bad type randomizes between the two messages in state $\omega_1 = 0$. ∎

PROOF OF PROPOSITION 6: We first explicitly construct an equilibrium in which, given $\omega_1 = 0$, the good type always announces 1 while the bad type announces 0 with probability $\mu_{1B}^0 \in (0, 1)$ in period 1. We then derive a sufficient condition under which anti-political correctness is indeed welfare-improving.

**Period 3:** As we have seen, the equilibrium in the final period is straightforward: the good type always tells the truth while the bad type always announces 1. We then obtain

$$v_{3G}(\lambda_{3G}, \lambda_{3H}) = \frac{x_3}{8(1 - \lambda_3)}, \quad v_{3B}(\lambda_{3G}, \lambda_{3H}) = \frac{y_3}{2(1 - \lambda_3)},$$

for any prior belief $\Lambda_3$.

**Period 2:** There only exists the babbling equilibrium in the continuation game following $m_1 = 1$, so we focus on the continuation game following $m_1 = 0$.

We first consider the problem for the bad type. It is clear that the bad type always announces 1 in state 1 because that can raise both the current payoff and the continuation payoff. In state 0, on the other hand, the bad type tells the truth only if

$$y_2 u^1(a_2^0) + v_{3B}(0, \lambda_{3H}^{0,0}) \geq y_2 u^1(a_2^1) + v_{3B}(0, 0),$$

26

where

$$a_2^1 = \frac{1}{2(1-\lambda_{2H}^{0,0})(1-\mu_{2B}^0)}, \quad \lambda_{3H}^{0,0} = \frac{\lambda_{2H}^{0,0}}{\lambda_{2H}^{0,0} + (1-\lambda_{2H}^{0,0})\mu_{2B}^0}.$$

Since $u^1(0) = 0$ and $v_{3B}(0,0) = \frac{1}{2}$, this condition can be written as

$$\frac{y_3(\lambda_{2H}^{0,0} + (1-\lambda_{2H}^{0,0})\mu_{2B}^0)}{2(1-\lambda_{2H}^{0,0})\mu_{2B}^0} \geq \frac{y_2}{2(1-\lambda_{2H}^{0,0})(1-\mu_{2B}^0)} + \frac{y_3}{2},$$

which is further simplified to

$$y_3\lambda_{2H}^{0,0}(1-\mu_{2B}^0) \geq y_2\mu_{2B}^0.$$

It is easy to see that this condition must hold with equality in equilibrium. The bad type therefore always randomizes in this contingency where

$$\mu_{2B}^0 = \frac{y_3\lambda_{2H}^{0,0}}{y_2 + y_3\lambda_{2H}^{0,0}}.$$

Now suppose that the good type deviates and announces 0 in period 1. In this case, it is clear that the good type always tells the truth because that can raise both the current payoff and the continuation payoff.

**Period 1:** We are now ready to check when anti-political correctness can be supported in equilibrium. First, the incentive compatibility constraint for the good type is given by

$$v_{2G}^{0,1} \geq v_{2G}^{0,0},$$

which can be written as

$$v_{3G}\left(\frac{\lambda_{1G}}{\lambda_{1G} + (1-\lambda_{1G}-\lambda_{1H})(1-\mu_{1B}^0)}, 0\right) \geq \frac{v_{3G}\left(0, \frac{\lambda_{2H}^{0,0}}{\lambda_{2H}^{0,0} + (1-\lambda_{2H}^{0,0})\mu_{2B}^0}\right) + v_{3G}(0, \lambda_{2H}^{0,0})}{2},$$

where

$$\lambda_{2H}^{0,0} = \frac{\lambda_{1H}}{\lambda_{1H} + (1-\lambda_{1G}-\lambda_{1H})\mu_{1B}^0}.$$

A sufficient condition for this is

$$\frac{\lambda_{1G}}{\lambda_{1G} + (1-\lambda_{1G}-\lambda_{1H})(1-\mu_{1B}^0)} \geq \frac{\lambda_{2H}^{0,0}}{\lambda_{2H}^{0,0} + (1-\lambda_{2H}^{0,0})\mu_{2B}^0} = \frac{\lambda_{1H}}{\lambda_{1H} + (1-\lambda_{1G}-\lambda_{1H})\mu_{1B}^0\mu_{2B}^0},$$

which holds if $\mu_{1B}^0$ is sufficiently close to one.

Second, the incentive compatibility constraint for the bad type is given by

$$v_{2B}^{0,1} = v_{2B}^{0,0},$$

which can be written as

$$\frac{y_2}{2} + v_{3B}\big(\frac{\lambda_{1G}}{\lambda_{1G}+(1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B})}, 0\big) = v_{3B}\big(0, \frac{\lambda_{1H}}{\lambda_{1H}+(1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}\mu^0_{2B}}\big).$$

This is further reduced to

$$y_2 + \frac{y_3(\lambda_{1G} + (1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B}))}{(1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B})} = \frac{y_3(\lambda_{1H} + (1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}\mu^0_{2B})}{(1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}\mu^0_{2B}}, \qquad (18)$$

where

$$\mu^0_{2B} = \frac{y_3\lambda_{1H}}{y_2(\lambda_{1H} + (1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}) + y_3\lambda_{1H}}. \qquad (19)$$

Substituting (19) into (18) and rearranging, we obtain

$$y_2 + \frac{y_3\lambda_{1G}}{(1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B})} = \frac{y_2(\lambda_{1H} + (1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}) + y_3\lambda_{1H}}{(1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}}. \qquad (20)$$

Note that the left-hand side is strictly increasing in $\mu^0_{1B}$ while the right-hand side is strictly decreasing, which implies that, for any given $(y_2, y_3)$, there exists a unique solution which solves (20).

We next show that the equilibrium with anti-political correctness exists for any given $(y_2, y_3)$ if $\frac{y_3}{y_2} \to 0$. To see this, we rewrite (20) as

$$\frac{\frac{y_3}{y_2}\lambda_{1G}}{1-\lambda_{1G}-\lambda_{1H}} = \frac{\lambda_{1H}(1-\mu^0_{1B}) + \frac{y_3}{y_2}\lambda_{1H}}{(1-\lambda_{1G}-\lambda_{1H})\mu^0_{1B}}.$$

We can then see that $\mu^0_{1B} \to 1$ and the equilibrium with anti-political correctness exists if $\frac{y_3}{y_2} \to 0$.

Given this explicit solution, it is straightforward to compare the expected payoffs. First, the expected payoff without reputational updating is

$$v_{3G}(\lambda_{1G}, \lambda_{1H}) = \frac{x_3}{8(1-\lambda_{1G}-\lambda_{1H})}.$$

On the other hand, the expected payoff with anti-political correctness when the expert happens to be good is

$$v_{3G}\big(\frac{\lambda_{1G}}{\lambda_{1G}+(1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B})}, 0\big) = \frac{x_3(\lambda_{1G} + (1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B}))}{8(1-\lambda_{1G}-\lambda_{1H})(1-\mu^0_{1B})}.$$

It is clear that the expected payoff with anti-political correctness becomes unbounded as $\frac{y_3}{y_2} \to 0$ and hence $\mu^0_{1B} \to 1$, so that anti-political correctness is welfare-improving. ∎

# Appendix B: the emergence of anti-political correctness with two types

In this appendix, we show by construction that it is indeed possible to have anti-political correctness in equilibrium even in the absence of the honest type. Consider the following example.

**Example:** Suppose that

$$u^0(a) = -a^2 \text{ and } u^1(a) = -(1-a)^2,$$

Let $\lambda_{3H} = 0$, $\lambda_{3G} \in \left(\frac{1}{2}, \frac{7}{8}\right)$, $\frac{x_1}{x_2} < \frac{49}{162}$, and $\frac{y_1}{y_2} < \frac{5}{18}$.

We can show that both types randomize in period 1 in this specified example. Now suppose that

$$\frac{x_3}{x_2} = \frac{\frac{7}{18} + \varepsilon_G}{\frac{7}{81}} \text{ and } \frac{y_3}{y_2} = \frac{\frac{5}{18} - \varepsilon_B}{\frac{11}{36}},$$

where

$$\varepsilon_G := \frac{x_1}{x_2}\left(u^0(a^*(\pi_1^1)) - u^0(a^*(\pi_1^0))\right) < \frac{x_1}{x_2},$$

$$\varepsilon_B := \frac{y_1}{y_2}\left(u^0(a^*(\pi_1^1)) - u^0(a^*(\pi_1^0))\right) < \frac{y_1}{y_2}.$$

Then there exists an equilibrium with anti-political correctness in which: (i) in the continuation game following $m_1 = 0$, the good type tells the truth in period 2 while the bad type always claims $m_2 = 1$; (ii) in the continuation game following $m_1 = 1$, the second-period equilibrium is babbling.

One can readily check that neither type has an incentive to deviate. The constructed equilibrium fully exploits the fact that there exist multiple continuation equilibria in period 2, in particular that the babbling equilibrium always exists under any circumstance in cheap-talk models. For this reason, we can always construct a situation in which different equilibria (are expected to) occur after different messages. To be more concrete, anti-political correctness is sustained in this particular case because the expert (somehow) holds a pessimistic expectation that an informative babbling equilibrium ensues when he chooses $m_1 = 1$. The same tradeoff then arises, in which the expert must trade off the short-run (second-period) reputation against the long-run (third-period) reputation.