Political Correctness as Anti-Herding

Melania Nica [*]

New College of the Humanities

January 2015

Abstract

This is a political correctness model where a decision maker has to take decisions based on the advice of a possibly biased expert when states are unverifiable. I find that in equilibrium an expert motivated by career advancement reports not only against her possible bias but also against the public prior on the state of the world. This result is similar to the concept of anti-herding, which is developed under the assumption of asymmetric information on an expert's ability (rather than their preferences).

Keywords: Political Correctness, Anti-herding, Unverifiable states

JEL Codes: D82, D83, G30, L20

Incomplete; please do not quote

[*] New College of the Humanities , Department of Economics, 19 Bedford Square, London WC1B 3HH, United Kingdom. *Email*: melania.nica@nchum.org. *Tel*: +44 207 637 4550.

I am grateful to Gilat Levy, Francesco Nava, Ronny Razin and Pasquale Schiraldi for helpful discussions. All errors remain mine.

1 Introduction

Many times we rely on expert's advice when we lack the knowledge or wisdom for taking proper decisions. However, how shall we interpret expert reports on questions for which there is no clear answer? Could we at least believe experts from fields with a higher degree of certainty?

In 1543, Nicolaus Copernicus' treatise On the Revolution of the Celestial Spheres was published when he was lying on his death bed. Contrary to the popular view that the Earth was the center of the universe Copernicus put forward a revolutionary theory - the Sun not the Earth was at the center of the universe. This idea was not new as it was proposed in the 3rd century BC by Aristarchus of Samos and advanced further by Martianus Cappela, the bishop Nicole Oresme and the cardinal Nicholas of Cusa in the in 5th, 14th and 15th century respectively. At that time, however the geocentric view based on Ptolemy's work was favored as it was closer to the Christian view of the world. Many historians have argued that Copernicus Copernicus waited to publish his work only at the end of his life as he feared ridicule and loss of reputation among his peers who were more inclined to accept the Ptolemaic representation of the universe. One could argue that it was fear of being declared heretic and not loss of reputation that made him postpone the publication, but then why the Catholic Church waited for 73 years after his death to ban On the Revolution of the Celestial Spheres? We could therefore agree at least partially that Copernicus distorted his academic discourse to not include his new theory on astronomy in order to not part with his reputed position among the scholars of the time.[1]

Before publishing his theory, in 1536, Copernicus found a most improbable supporter from the most probable biased side - the Church - in the Cardinal of Capua Nicholas Schönberg. In one of his letters to Copernicus he writes: "I had learned that you formulated a new cosmology. In it you maintain that the earth moves; that the sun occupies the lowest, and thus the central, place in the universe [...]. With the utmost earnestness I entreat you, most learned sir, unless I inconvenience you, to communicate this discovery of yours to scholars" and then he concludes: "If you gratify my desire in this matter, you will see that you are dealing with a man who is zealous for your reputation and eager to do justice to so fine a talent." So we see that Cardinal of Capua, a man of the Church, acts also in an astonishing manner disavowing the Church's beliefs (in the geocentric theory) by favoring the revolutionary theory of Copernicus.[2]

[1] Copernicus had a doctorated in cannon law, but also professed as a physician, governor, diplomat, and economist. Please see Armitage, A: *Copernicus, the founder of modern astronomy,* Publisher: T Yoseloff, 1957

[2] Copernicus theory was finally accepted by the 18th century.

This historical exposition captures the very important starting point of this paper: there are situations where there is impossible to know or at least verify the reports of an expert. Furthermore, the expert could be biased and even have a predisposition for accolades and good reputation. How, then, do we interpret her report? Does she truthfully report her information or her belief? Will she distort her report as both Copernicus (by restraining to disclose his work) and the Cardinal of Capua (by acting against his belief) seemed to have done at some point. Had the fact that the overall accepted view was that the Earth was the center of the Universe and not otherwise affected their behavior?

In order to explore the questions above, I construct a reputation forming game involving a decision maker and an expert who can favor a particular ideology. The decision maker is uncertain on whether the expert is biased or not. The expert reports over two periods about the state of the world to the decision maker who then takes an action based on this report. The expert is informed about the state (but imperfectly), while at the same time the decision maker is not able to verify it. However, the decision maker is aware of a common prior on the state of the world. The preference for a good reputation is captured by the relative value that the expert puts on the future versus the present - so I center this model in a career concerns framework.

The corporate finance/career concerns literature has recognized that people take inefficient decisions when driven by career advancement. This was firstly explored by Fama (1980) and Holmstrom (1982). In the context of the wider literature on career concerns, this paper is related in a behavioral sense to the two strands in which there is asymmetric information about *ability* or *misalignment of preferences*. In terms of ability, Trueman (1994), Avery and Chevalier (2001), Effinger and Polborn (2001) and Levy (2004) show that managers might excessively contradict public information (or anti-herd) to distinguish themselves from the rest and increase their reputation. On the other hand, Scharfstein and Stein (1990), Prendergast and Stole (1996), Ottaviani and Sorensen (2001), Prat (2005) to name just few, explore the behavior of managers when they ignore their own information and herd on the others' actions for the purpose of being perceived informed. In these papers, the uncertain types are in the dimension of ability, unlike Sobel (1985), Benabou and Laroque (1992) and Morris (2001) where the uncertainty is on the alignment of preferences with a principal.

This paper is related to the second class of models as it deals with asymmetric information about the misalignment of preferences. There is also a conceptual link with Morris (2001) which shows that experts motivated by career advancement might distort their reports, so that their listeners would not infer that they are biased. Morris' result however is built on the fact that the decision maker compares expert's report with a realized state when building expert's reputation. In this model as the state of the world of the world is uncertain, this comparison is not viable anymore and the decision maker has to make use of the public view on the state of the world. Similar to Morris, I find that the experts report against their possible bias for reputational reasons. However, in this uncertain environment there is a further incentive in place: in order to build their reputation experts report also against the public prior on the state; furthermore declaring against one's possible bias is stronger when the public thinks the opposite. This result is similar with the concept of anti-herding developed by Levy (2004) and others in the career concerns strand of asymmetric information about the ability rather than alignment of preferences.

Morris's reasoning for the political correctness is based on Loury (1994) who develops a syllogism for political correctness as reputational distortion due to the inherent inclination of members of a community to adhere to communal values. People declare as their fellows as to not offend the community and remain in good standards with their peers. Failing to do so results in the "odds that the speaker is not in fact faithful to communal values as estimated by a listener otherwise uniformed about his views to increase." So, Morris even though he accepts that the motivation of his model is narrower in scope than Loury's argument, he adheres to political correctness as reputational distortion due to conformity to social norms as in Bernhem (1994).

In this model contrary to Morris but building on Morris I aim to show that people act in a political correct manner not only to disavow their individual bias but also to show that they hold different views that their peers. By not adhering to an accepted dogma they build their reputation of being of a good type.

I set this study within the cheap talk framework introduced by Crawford and Sobel(1982) where information is transmitted from an expert to a decision maker through costless signals. A decision maker has to take a decision however over two periods based on the reports of an expert about whose preference he is uncertain. A state of the world, 0 or 1, is realized and there is a public prior on what this realization is. Once the expert is informed (partially) on the state of the world, she reports (costlessly) to the decision maker who takes an action between 0 and 1 which reflects his belief in the state being 1. The expert is either good with an objective of being as close as possible to the true state of the world, or bad - biased in favor of 1. In this model, the expertise is based on partially informative signals, and expert's initial reputation is publicly known. The expert is concerned about the decision maker's belief on her being of a good type as higher reputation translates to a higher chance of influencing the decision maker in the future.

The game is solve by backward induction. In the last period, I find that in an informative equilibrium the expert irrespective of being good or bad reports as per her preferences as there are no career concerns in place to distort her views

In the first period, the expert trades off her respective current preference against the incentive to distort her information for reputational reasons. Depending on expert's initial reputation, signal precision, and her relative preference for the future, I show the existence of truthtelling, informative and non-informative equilibria. In a truthtelling equilibrium the good expert discloses fully her signal while the bad one only partially. The informative equilibrium occurs when good expert's career concerns become more important and she also starts to report only partially the true signals. A limiting case is the babbling equilibrium when the good expert never gives a report consistent with her perceived bias, and the bad expert pools on this strategy.

Furthermore more I show that in both truthtelling and informative equilibrium the expert tends to declare against her individual perceived bias in order to show that she cannot be biased herself (the political correctness effect) but also when the prior is in favor of her bias (the anti-herding effect).

The rest of the paper is organized as follows. In the next section I describe the model. As the game is set over two periods, in section 3 I find and characterize the equilibrium in the last stage game; in section 4 I characterize and show the existence of the equilibrium. All proofs are not in the text but in the appendix.

2 Model

There are two players in this game: a decision maker D, and an expert R. The game is played over two periods, $t \in \{1, 2\}$.

There is an underlying state of the world x_t which can take values of 0 and 1. The prior probability that the state is 0 is τ - $\Pr(x_1 = 0) = \tau$ and $\Pr(x_1 = 1) = 1 - \tau$ - with $\tau \in (0, 1)$.. The states of the world are drawn independently each time. The decision maker is not able to verify the state of the world in either period. However, the expert receives a noisy but informative private signal about the true state of the world each period: $s_t \in \{0, 1\}$. The signal has precision $p = \Pr[s_t = x_t | x_t] > \frac{1}{2}$.

The decision maker receives report r_t about x_t from the expert R and based on this report, he takes an action $a_t \in [0, 1]$. His objective is to be as close as possible to the true state of the world, so I set his expected payoff to be: $-\mu_1 E(x_1 - a_1)^2 - \mu_2 E(x_2 - a_2)^2$.

There are two types of experts: 'good' (G) and 'bad' (B) and the decision maker D is uncertain of their type. D's prior probability that R is of type G is $\lambda_1 \in (0, 1)$.

The good expert has preferences aligned with the decision maker, which is reflected in his payoff structure. If the expert R is good, her payoff is: $-\mu_1^G E\left[(x_1 - a_1)^2 | s_1\right] - \mu_2^G E\left[(x_2 - a_2)^2 | s_2\right]$.

The bad expert is biased towards state 1. If R is bad, she has a higher utility when the action taken by D is closer to 1. Her payoff is $\mu_1^B a_1 + \mu_2^B a_2$.

The experts could value the present different than the future by assigning different weights to current and future payoffs: $\mu_1^k > 0$,and $\mu_2^k > 0$ with $k \in \{G, B\}$. These weights reflect different time preferences between experts and allow for situations in which any of the parties involved could value the future payoff more than the current one.

After observing report r_1, D updates his beliefs on the type of the experts and on the state of the world x_1. Expert's posterior reputation is denoted λ_2 and the belief on the state of the world $\Gamma(x_1|r_1) = \Pr(x_1|r_1)$. For simplicity of notations I denote the posterior belief that the state of the world is 1 with $\Gamma(r_1)$.

If the state of the world were verifiable, the decision maker could update the reputation of the expert by comparing the report of the expert with the realized state. When the state is unknown, the updating is based only on the report, keeping in mind their initial reputation.

In the second period ($t = 2$) the game is repeated, with the state of the world x_2 independent of x_1. In this model, everything apart from the type of the experts and their private signals is known by everyone.

The strategy profile for the players is $(\pi_{kt}(s_t), a_t(r_t))$, where $\pi_{kt}(s_t)$ is R's probability of reporting 1 when the signal is s_t and $a_t(r_t)$ is the action taken by the decision maker given r_t. It is important to note that the experts' strategies represent the probability that their report is the same as their potential bias. The relevant state variable for this game is experts's reputation Λ_t. The posterior belief on the state of the world Γ_t is not carried forward to the future as the state of the world is *i.i.d.* over time.

Definition 1 *A strategy profile $(\pi_{kt}(s_t), a_t(, r_t))$ is an equilibrium if (a) the experts's reports given their signals maximize their respective payoffs given the posterior reputational beliefs, (b) the decision maker's action maximizes his expected payoff given his posterior probability on the state of the world and (c) the posterior probabilities on the type of the expert and the state of the world are derived according to Bayes' rule.*

As the game is set over two periods the equilibrium outcomes will be determined by backward induction. In each stage game I will use the strategy profile without a time subscript for notational ease.

3 The Second Stage - No Reputational Concerns

In the last period R enter with reputations λ_2. This is a cheap talk game where the expert's report does not enter her payoff directly but indirectly through the influence she has on

D's belief about the state of the world and consequently through D's action. As for any cheap talk game there exists always a babbling equilibria. Below, I characterize however the informative equilibrium of the game.

An informative equilibrium is an equilibrium in which expert's report is correlated with the state of the world for any r_2.

Proposition 1 *There exists an informative equilibrium where the decision maker's optimal action is $a_2^*(r_2) = \Gamma(r_2)$. The good expert' s optimal strategies are $\pi_G(1) = 1$, $\pi_G(0) = 0$. The bad expert's strategies are $\pi_B(s_2) = 1$, for any $s_2 \in \{0, 1\}$.*

The above equilibrium strategies reflect the fact that in the last period the good expert declares her signals while the bad expert's report is consistent with her respective bias.

The idea behind this proposition is the fact that in an informative equilibrium the message sent carries some information to the decision maker. Essentially, if the decision maker observes 1 from R (the message is informative) he will choose a higher action than if he had observed 0, thus the bad expert R will have a strict incentive to declare 1 while the good expert R will have a strict incentive to truthfully reveal her signal.

The optimal action of the decision maker for all possible reports is:

$$a_2^*(r_2) = \begin{cases} \frac{(1-p)(1-\tau)}{(1-p)(1-\tau)+p\tau} & \text{if } r_2 = 0 \\ \frac{[\lambda_2 p+(1-\lambda_2)](1-\tau)}{1-\lambda_2(1+2p\tau-p-\tau)} & \text{if } r_2 = 1 \end{cases}$$

Next I look at how the individual reputational change affects the expert's expected payoff in the second period.

For a good type R the value of reputation acquired from the first period is her ex-ante expected payoff $-E\left[(x_2 - a_2^*)^2 | \Lambda_2\right]$ and it is calculated as follows:

$v^G(\lambda_2) = -\sum\limits_{x_2}\sum\limits_{n} \Pr(x_2) \Pr\left[s_2^R = n | x_2\right] (x_2 - a_2(r_2 = n))^2.$

In the above expression $x_2, n \in \{0, 1\}$. The state of the world is drawn independently each time, so a good expert entering the second period could face either a state 0 with probability τ or 1 with probability $1 - \tau$, moreover there is also uncertainty on the signal received by the expert given the state of the world.

The bad expert however is biased towards 1, so irrespective of her signal, her expected payoff feature these biases. As a result R's reputational value is:

$v^B(\lambda_2) = \sum\limits_{x_2} \Pr(x_2) a_2(r_2 = 1).$

Result 1 *The value of reputation of the expert (irrespective of her type) is strictly increasing and continuos in her posterior reputation.*

This is an important result as it triggers the action of the experts in the first period: irrespective of her type the expert has incentives to acquire good reputation in the first period so that her voice is heard in the next period.

4 First Stage Game

The first period game is similar with the second period game with the exception that the expert has reputational concerns for the second period of the game. The prior probability of the expert being good is λ_1.

Experts' total payoff functions account for both current and future payoffs, taking into account their relative time preference. I represent the total payoffs in terms of the relative weight of the first period payoff i.e. $\mu^k \equiv \frac{\mu_1^k}{\mu_2^k}$ with $k \in \{G, B\}$.

Experts' total payoff is the sum of the first stage payoff weighted by the appropriate time preference and the second stage expected payoff (which I called in the previous section experts' value of reputation).

The good expert's total payoff is $\mu^G u_G(r_1, s_1) + v_G^k(\lambda_2)$ where $k \in \{G, B\}$. Their current payoff $u_G(r_1, s_1)$ is $-E(x_1 - a_1^*(r_1)|s_1)^2$ and captures the objective of the good experts to take an action as close as possible to the state of the world.[3]

The bad expert's total payoff account for her preference for the state 1. As a result bad type expert has a total payoff of $\mu^B u_B(r_1) + v_B(\lambda_2)$ where $u_B(r_1) = a_1(r_1)$.

4.1 Reputation Formation

The expert enters the first stage game with an initial prior on her reputation λ_1. After she sends her report, the decision maker updates his belief on her type.

Expert's posterior reputation and the posterior belief on state are obtained by Bayesian updating given only the report provided by the expert as D is not able to verify the state of the world. There is no comparison with a counterpart or a state.

The expert of type $k \in (G, B)$ reports r_1 with probability $\phi_k(r_1)$. This probability takes into account the fact that the state of the world could be either 0 or 1. $\phi_k(r_1) = \phi_k(r_1|x = 1)\Pr(x_1 = 1) + \phi_k(r_1|x_1 = 0)\Pr(x_1 = 0)$.

R's posterior reputation is:

$$\lambda_2(r_1) = \frac{\lambda_1\phi_G(r_1)}{\lambda_1\phi_G(r_1) + (1 - \lambda_1)\phi_B(r_1)}$$

[3] $u_G(r_1, s_1 = 1) = -E(x_1 - a_1^*(r_1)|s_1 = 1)^2 = -\frac{1}{2}p + pa_1(r_1) - \frac{1}{2}a_1(r_1)^2$

while posterior probability that the state of the world is 1 $\Pr(x_1 = 1|r_1)$ denoted as $\Gamma(r_1)$ is:

$$\Gamma(r_1) = \frac{\Pr(r_1|x_1 = 1)\Pr(x_1 = 1)}{\Pr(r_1|x_1 = 1)\Pr(x_1 = 1) + \Pr(r_1|x_1 = 0)\Pr(x_1 = 0)}$$

Furthermore we know that:

$$\Pr(r_1 = 1|x = 1) = \left[\begin{array}{c} p(\lambda_1\pi_G(1) + (1-\lambda_1)\pi_B(1)) \\ +(1-p)(\lambda_1\pi_G(0) + (1-\lambda_1)\pi_B(0)) \end{array}\right]$$

$$\Pr(r_1 = 1|x = 0) = \left[\begin{array}{c} p(\lambda_1\pi_G(0) + (1-\lambda_1)\pi_B(0)) \\ +(1-p)(\lambda_1\pi_G(1) + (1-\lambda_1)\pi_B(1)) \end{array}\right]$$

$$\Pr(r_1 = 0|x = 1) = \left[\begin{array}{c} p(\lambda_1(1-\pi_G(1)) + (1-\lambda_1)(1-\pi_B(1))) \\ +(1-p)(\lambda_1(1-\pi_G(0)) + (1-\lambda_1)(1-\pi_B(0))) \end{array}\right]$$

$$\Pr(r_1 = 0|x = 0) = \left[\begin{array}{c} p(\lambda_1(1-\pi_G(0)) + (1-\lambda_1)(1-\pi_B(0))) \\ +(1-p)(\lambda_1(1-\pi_G(1)) + (1-\lambda_1)(1-\pi_B(1))) \end{array}\right]$$

4.2 First Stage Equilibrium

Similar with the second period game, D does not observe the state and as a result his optimal action is his posterior belief about the state of the world.

$$a_1^*(r_1) = \Gamma(r_1)$$

Proposition 2 *Any informative equilibrium* $(\pi_k, \Gamma, \lambda_2)$ *is characterized by:*

1. *When the good expert R observes signal $s_1 = 0$, she always announces* 0 - $\pi_G(0) = 0$*; truthtelling is always optimal for R when her signal is* 0.

2. *the equilibrium reputations (posteriors) are such that* $\lambda_2(0) \geq \lambda_2(1)$ *and* $\frac{d\lambda_2(0)}{d\tau} < 0$

The first result is that if a good expert gets a signal opposite her potential bias, she will report it truthfully. The logic behind this is the fact that if $s_1 = 0$ there is no benefit from lying for a good expert.

The second result says that there are also incentives to report against one's bias for reputational reasons as declaring against one's perceived bias increases the probability that the expert is good.

Furthermore, R's reporting against her potential bias decreases with the probability that the state is 0. In particular the intensity of declaring 0 for the purpose of disavowing one's bias decreases with probability that the state of the world is 0. Basically, at low levels of τ, by doing 0 R says: "because the state is more likely to be 1 rather than 0, I report 0 and thus agree with a low prior on state 0, to show the decision maker that I am not biased as I am as far as possible from 1 which has a high probability of being the true state."

So there is a clear connection with the herding literature. In particular, R is more likely to report 0 if the prior on the 0 state is low. This means that the expert contradicts public information by reporting 0 when the the prior on state 1 is high. This would be equivalent with the anti-herding idea developed by Levy (2004) when careerist experts contradict public information, but here it is applied to experts with possible misalignment of preferences. This result is in the opposite direction to what the political correctness literature has described as a reputational distortion due to the inherent inclination of members of a community to adhere to communal values for fear of not being ostracized in the future.

Contrary to the literature and Morris (2001) this result thus shows that people act in a political correct manner to signal that they hold different views than their community. Thus political correctness is not herding but anti-herding.

4.3 Informative Equilibrium Existence

First I investigate whether this game supports a *full truthtelling equilibrium* where the expert, irrespective of her type, reports truthfully her signal: $\pi_G(1) = 1$, $\pi_G(0) = 0$ and $\pi_B(1) = 1$, $\pi_B(0) = 0$. However, it is easy to see that there does not exist such an equilibrium as the bad expert has incentives to deviate to reporting 1.

Claim 1 *There is no informative equilibrium with the expert following full truthtelling strategies.*

This is due to the fact that if such an equilibrium exists the posterior reputations are equal with the priors. But this implies that there is no reputational cost for the bad expert of announcing her biases. Thus regardless of her signal the bad expert always reports 1. But this is a contradiction to truthtelling. So, there is no full truthtelling equilibrium.

The next proposition looks at the equilibrium existence of the first stage game:truthtelling equilibrium (when the good expert always reports her signals), informative equilibrium when the experts (irrespective of their type) distort their reports with positive probability but information is transmitted to the decision maker and non-informative equilibrium when no information is transmitted to the decision maker.

Proposition 3 *For any $\lambda_1 \in (0,1)$ there exist $\bar{\mu}^G, \underline{\mu}^G \in (0,1)$ such that*

1. *if $\mu^G(\lambda_1, \mu^B) > \bar{\mu}^G$ there is a unique truthtelling equilibrium where $\pi_G(1) = 1, \pi_G(0) = 0, \pi_B(1) = 1$ and $\pi_B(0) \in (0,1]$.*

2. *if* $\underline{\mu}^G < \mu^G\left(\lambda_1, \mu^B\right) \leq \bar{\mu}^G$ *there exists an unique informative equilibrium* $\pi_G(1) \in (0,1], \pi_G(0) = 0, \pi_B(1) = 1$ *and* $\pi_B(0) \in (0,1]$.

3. *if* $\mu^G\left(\lambda_1, \mu^B\right) \leq \underline{\mu}^G$ *the equilibria of the games are non-informative.*

The truthtelling equilibrium (point 1 in the above proposition) describes a situation when the good R reports her signal with probability 1, while the bad R reports her signal only when the signal coincide with her bias. This equilibrium exists as long as the good expert does not value the future high enough in order to distort her reports.

Political correctness as anti-herding in a truthtelling equilibrium is reflected in the actions of the bad expert. In particular a bad expert has a higher inclination to reveal her signal when signal is 0 but there is a higher prior on the state being 1.

In terms of the historical example presented at the beginning this would correspond to the Cardinal's support of Copernicus claim. We see thus, that Cardinal of Capua is disciplined by his reputational concerns to favoring the heliocentric theory (on which he has information that it is correct) instead of supporting the opposite view of the church. He is politically correct (by going against his bias) but even more so since the geocentric view was generally accepted at that time. In this way he builds his reputation of being a true man of science and not just a simple follower of the Church.

When the good expert's career concerns start becoming important she starts developing incentives to distort her report as well. This is the case when the good expert' signal is her potential bias and the future benefit from lying is higher than the current benefit from telling the truth. The bad expert still tells the truth with a positive probability for reputation building reasons. This reflects the informative equilibrium (point 2 in the above theorem) when the good R reports her signal truthfully when the signal is opposite the perceived bias, while the bad R reports truth only when the signal coincide with her bias.

While the direct political correctness effect - as reporting against the perceived bias - on the action of good experts could be observed in many historical examples including the case of Copernicus when he chose not to report his scientific results, however political correctness as anti-herding is a more fine effect to detect. This is due to the fact that it reflects the perverse effect of misreporting own information for the purpose of showing that one is different than the others. Had President Nixon chosen to open dialog with China in a period when the public was opposing it just for political reason? Was he proving in this manner that he was not biased against the left (himself being a Republican) but also different than the general accepted view? Other examples are when President Trueman fired General MacArthur during the Korean War [4] or when George W. Bush signed a nuclear deal with North Korea in 2007

[4] These examples were pointed out by Gilat Levy in *Anti-herding and Strategic Consultation.*

even though he included North Korea on the axis of evil in 2002 and the public did not favor this agreement.

The non-informative equilibrium arrises in the situation in which political correctness takes full hold of good experts behavior. As a result the good expert never declares her possible bias while the bad expert pools on the action of the good expert.

4.4 Conclusion

In this paper I extend Morris (2001) by allowing the states of the world to be unverifiable. Morris' political correctness result is built however on the direct comparison between expert's report with the realized state. As the state of the world of the world is uncertain, this comparison is not viable anymore and the decision maker compares the report with the public view on the state of the world. Similar to Morris, I find that the experts report against their possible bias for reputational reasons. However, there is a further incentive in place: in order to build their reputation experts report also against the public prior on the state. Furthermore declaring against one's possible bias is more intense when the public thinks the opposite. So this model depicts political correctness as an anti-herding result.

This paper lies at the congruence of two bodies of research: the career concerns literature with uncertain misalignment of preferences between a decision maker and agent as in Morris (2001) and the career concerns literature with uncertain level of expertise as in Prendergast and Stole (1996), Ottaviani and Sorensen (2001), Levy (2004). While these strands of literature have developed separately, this paper builds a unifying framework for both of them.

References

[1] Avery, C.N., and J. A. Chevalier, 1999, Herding over Career, *Economic Letters*, 63, 327–333.

[2] Benabou, R. and G. Laroque, 1992, Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility, *Quarterly Journal of Economics*, 107 (3), 921-958.

[3] Bernheim, B. D., 1994, A Theory of Conformity, *Journal of Political Economy*, 102, 841-877.

[4] Crawford, V. P., and J. Sobel, 1982, Strategic Information Transmission, *Econometrica*, 50, 1431-1451.

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

[6] Levy, G., 2004, Anti-Herding and Strategic Consultation, *European Economic Review*, 48, 503-525.

[7] Loury, G., 1994. Self-censorship in public discourse, *Rationality and Society,* 6, 428-61.

[8] Mailath,G. J , and L. Samuelson, 2001, Who Wants a Good Reputation?, *Review of Economic Studies*, 68, 415-41.

[9] Morris, S., 2001, Political Correctness, *Journal of Political Economy*, 109, 231-265..

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

[11] Prat, A. 2005, The Wrong Kind of Transparency, *American Economic Review*, 95(3), 862-877.

[12] Prendergast, C., and L. Stole, 1996, Impetuous Youngsters and Jaded Old-Timers: Acquiring a Reputation for Learning, *Journal of Political Economy*, 104, 1105-34.

[13] Scharfstein, D S., and J. C. Stein, 1990, Herd Behavior and Investment, *American Economic Review*, 80, 465-479.

[14] Sobel, J., 1985, A Theory of Credibility, *Review of. Economic Studies*, 52, 557–73.

[15] Trueman, B., 1994, Analyst Forecasts and Herding Behavior, *Review of Financial Studies*, 7, 97-124.

5 Appendix

5.1 Second Proposition 1

D believes that if $r_2 = 0$, R is good while if $r_2 = 1$ R is good with probability λ_2. Based on these beliefs we could compute by Bayes' rule also probability of the state being 1 in period 2.

- $\Pr(x_2 = 1|r_2 = 0) = \frac{(1-p)(1-\tau)}{(1-p)(1-\tau)+p\tau}$
- $\Pr(x_2 = 1|r_2 = 1) = \frac{(\lambda_2 p+(1-\lambda_2))(1-\tau)}{[\lambda_2 p+(1-\lambda_2)](1-\tau)+[(\lambda_2(1-p)+(1-\lambda_2))]\tau} = \frac{[\lambda_2 p+(1-\lambda_2)](1-\tau)}{1-\lambda_2(1+2p\tau-p-\tau)}$

The decision maker's payoff in the last period is $-E(x_2 - a_2)^2$, so for message r_2 the optimal action of the principal is:

$$a_2^*(r_2) = \Pr(x_2 = 1|r_2)\,1 + \Pr(x_2 = 0|r_2)\,0 = \Pr(x_2 = 1|r_2)$$

$$a_2^*(1) = \frac{[\lambda_2(p-1)+1](1-\tau)}{1-\lambda_2(1+2p\tau-p-\tau)}$$

$$a_2^*(0) = \frac{(1-p)(1-\tau)}{(1-p)(1-\tau)+p\tau}$$

In this case it is easy to see there is no incentive for R to deviate from full truthtelling equilibrium.

5.2 Proof of Result 2

For a good type expert her expected payoff at the beginning of period 2 given decision maker posterior belief on the experts reputation is:

$$v^G(\lambda_2) = -E[(x_2 - a_2^*)|\lambda_2]$$

$$\begin{aligned}E\left[\left(x_2-a_2^*\right)|\lambda_2\right] &= (1-\tau)\,p\left(1-\Pr\left(x_2=1|r_2=1\right)\right)^2+\\ &\quad \tau\left(1-p\right)\left(0-\Pr\left(x_2=1|r_2=1\right)\right)^2+\\ &\quad (1-\tau)\left(1-p\right)\left(1-\Pr\left(x_2=1|r_2=0\right)\right)^2\\ &\quad \tau p\left(0-\Pr\left(x_2=1|r_2=0\right)\right)^2\\ &= (1-\tau)\,p\left(1-a_2^*\left(r_2=1\right)\right)^2+\tau\left(1-p\right)\left(a_2^*\left(r_2=1\right)\right)^2\\ &\quad +(1-\tau)\left(1-p\right)\left(1-a_2^*\left(r_2=0\right)\right)^2+\tau p\left(\left(a_2^*\left(r_2=0\right)\right)\right)^2\end{aligned}$$

For a bad type her expected value at the beginning of period 2 is:

$$v^B\left(\lambda_2\right)=E\left[a_2^*\left(r_2=1\right)|\lambda_2\right]$$

Result 2 *The value of reputation for the bad expert is increasing in her posterior reputation:*

$$\frac{da_2^*\left(r_2=1\right)}{d\lambda_2}=\frac{\left(2p-1\right)\left(1-\tau\right)\tau}{\left[1-\lambda_2\left(1+2p\tau-p-\tau\right)\right]^2}>0$$

Result 3 *The value of reputation for the good expert is increasing in her posterior reputation:*

$$\begin{aligned}\frac{d}{d\lambda_2}E\left[\left(x_2-a_2^*\right)|\lambda_2\right] &= -\left(1-\tau\right)p\frac{\left(\lambda_2\left(1-p\right)+\left(1-\lambda_2\right)\right)\tau}{1-\lambda_2\left(1+2p\tau-p-\tau\right)}\frac{da_2^*\left(r_2=1\right)}{d\lambda_2}+\\ &\quad \tau\left(1-p\right)\left(\frac{\left[\lambda_2 p+\left(1-\lambda_2\right)\right]\left(1-\tau\right)}{1-\lambda_2\left(1+2p\tau-p-\tau\right)}\right)\frac{da_2^*\left(r_2=1\right)}{d\lambda_2}\\ &= -\frac{\left(1-\tau\right)\tau\left(1-\lambda_2\right)\left(2p-1\right)}{1-\lambda_2\left(1+2p\tau-p-\tau\right)}\frac{da_2^*\left(r_2=1\right)}{d\lambda_2}\end{aligned}$$

So:

$$\frac{dv^G\left(\lambda_2\right)}{d\lambda_2}=-\frac{dE\left[\left(x_2-a_2^*\right)|\lambda_2\right]}{d\lambda_2}=\frac{\left(1-\tau\right)\tau\left(1-\lambda_2\right)\left(2p-1\right)}{1-\lambda_2\left(1+2p\tau-p-\tau\right)}\frac{da_2^*\left(r_2=1\right)}{d\lambda_2}>0$$

5.3 Proof of Claim 3

Assume that such an equilibrium existed then $\lambda_2\left(r_1\right)=\lambda_1$. But this implies that there is no reputational cost for R of announcing 1. On the other hand D will take a higher action after R sending a 1 message. Because a biased R prefers a higher action there is a strict incentive to send $r_1=1$. Thus regardless of the signal the bad R will send message 1: $\pi_B\left(1\right)=\pi_B\left(0\right)=1$ which is a contradiction.

5.4 Proof of Proposition 4

Proof 1. I will prove the first point by contradiction.

Suppose not and $\lambda_2(1) > \lambda_2(0)$; in this situation a bad R has a both a higher reputation by declaring 1 and a higher current payoff for any $s_1 = \{0,1\}$; thus the biased R will always say 1 and $\pi_B(0) = \pi_B(1) = 1$ resulting in $\phi_B(0) = \phi_B(1)$. Then,

$$\lambda_2(r_1) = \frac{1}{1 + \frac{1-\lambda_1}{\lambda_1}\frac{1}{\phi_G(r_1)}}$$

Now, in order to have $\lambda_2(1) > \lambda_2(0)$ then $\phi_G(0) < \phi_G(1)$ must be satisfied. However this is not possible as a 0 report from R implies that R is of a good type, and thus $\phi_G(0) > \phi_G(1)$ always.

Hence, $\lambda_2(0) \geq \lambda_2(1)$. ■

Proof 2. Now,

$$\lambda_2(r_1) = \frac{\lambda_1\phi_G(r_1)}{\lambda_1\phi_G(r_1) + (1-\lambda_1)\phi_B(r_1)}$$

where

$$\phi_k(r_1) = \phi_k(r_1|x=0) + [\phi_k(r_1|x=1) - \phi_k(r_1|x=0)](1-\tau)$$

For notational ease I denote $\phi_k(r_1|x=0)$ with A_k and $[\phi_k(r_1|x=1) - \phi_k(r_1|x=0)]$ with B_k

$$\lambda_2(\tau) = \frac{\lambda_1[A_G + B_G(1-\tau)]}{\lambda_1[A_G + B_G(1-\tau)] + \left(1-\lambda_1^R\right)[A_B + B_B(1-\tau)]}$$

or

$$\lambda_2(\tau) = \frac{1}{1 + \frac{1-\lambda_1^R}{\lambda_1^R}\frac{A_B + B_B(1-\tau)}{A_G + B_G(1-\tau)}}$$

thus

$$\frac{d\lambda_2(\tau)}{d\tau} = -\frac{1}{\left(1 + \frac{1-\lambda_1^R}{\lambda_1^R}\frac{A_B + B_B(1-\tau)}{A_G + B_G(1-\tau)}\right)^2}\frac{1-\lambda_1^R}{\lambda_1^R}\frac{df(\tau)}{d\tau}$$

where $f(\tau) = \frac{A_B + B_B(1-\tau)}{A_G + B_G(1-\tau)}$.

Now, $\frac{df(\tau)}{d\tau} = \frac{B_G A_B - B_B A_G}{(A_G + B_G(1-\tau))^2}$ and it is positive if $\frac{B_G}{A_G} \geq \frac{B_B}{A_B}$.

Returning to the original notations this means $\frac{\phi_G(r_1|x=1) - \phi_G(r_1|x=0)}{\phi_G(r_1|x=0)} \geq \frac{\phi_B(r_1|x=1) - \phi_B(r_1|x=0)}{\phi_B(r_1|x=1)}$. However this is always true as long as $\pi_B(1) \geq \pi_G(1)$ and $\pi_B(0) \geq \pi_G(0)$ which is implied by point 1.

Thus we can conclude that$\frac{d\lambda_2(r_1)}{d\tau} < 0$. ■

5.5 Proof of Equilibrium Existence

Let's assume that this equilibrium exists: $\pi_G(1) = 1$; $\pi_G(0) = 0$. It cannot be the case that the bad expert also tells the truth always. In any informative equilibrium we know that the posterior reputation of an R expert after announcing 0 must be higher . Thus $\lambda_2(0) \geq \lambda_2(1)$, translates into $\pi_B(1) \geq \pi_G(1)$ and $\pi_B(0) \geq \pi_G(0)$ with one strict inequality. But if good R tells the truth in equilibrium this implies $\pi_B(1) = 1$ and $\pi_B(0) > 0$.

Now, I look for the equilibrium strategy of the bad expert.

Suppose that the expert observes signal 0. Her current utility from lying would be $\mu_B a_1^*(1)$ while her current utility from telling the truth would be $\mu_B a_1^*(0)$. The net expected benefit from lying when observing signal 0 is:

$$\Pi_B(s_1 = 0) = \mu_B\left(a_1^*(1) - a_1^*(0)\right)$$

$$a_1^*(1) = \frac{\Pr(r_1 = 1|x = 1)(1-\tau)}{\Pr(r_1 = 1|x = 1)(1-\tau) + \Pr(r_1 = 1|x = 0)\tau}$$

$$a_1^*(0) = \frac{\Pr(r_1 = 0|x = 1)(1-\tau)}{\Pr(r_1 = 0|x = 1)(1-\tau) + \Pr(r_1 = 0|x = 0)\tau}$$

Under the assumption of truthtelling equilibrium:

$\Pr(r_1 = 1|x = 1) = [p + (1-p)(1-\lambda_1)\pi_B(0)]$

$\Pr(r_1 = 1|x = 0) = [(1-p) + p(1-\lambda_1)\pi_B(0)]$

The reputational cost of lying when observing signal 0 is:

$$\Pi R_B(s_1 = 0) = v^B(\lambda_2(r_1 = 0)) - v^B(\lambda_2(r_1 = 1))$$

where $v^B(\lambda_2(r_1)) = \frac{[\lambda_2(r_1)(p-1)+1](1-\tau)}{1-\lambda_2(r_1)(1+2p\tau-p-\tau)}$.

Her equilibrium strategy $\pi_B(0)$ is determined by the indifference condition between the net current benefit versus the future reputational costs:

$$\Pi_B(s_1 = 0) = \Pi R_B(s_1 = 0)$$

In order to complete the proof, now I look under which time preference parameter the good expert is indeed telling the truth. We already know that when the signal is 0 the good expert always tells the truth, however there might be a distortion when the signal is 1 for political correct reasons. In order not to have this distortion it is necessary and sufficient that the net current gain from telling the truth is greater than the reputational costs of telling the truth when the signal is 1. For any parameter λ_1 we can find thresholds for the time preference parameter μ_R such that $\mu_R > \bar{\mu}_R$ there exist a truthtelling equilibrium when the good expert tells the truth. $\bar{\mu}_R$ as solution to the indifference condition of the good expert.

The net current benefit of telling the truth is:

$$\Pi_G(s_1 = 1) = \mu_B\left(a_1^*(1) - a_1^*(0)\right)$$

while the net current cost of telling the truth is:

$$\Pi R_G(s_1 = 1) = v^G\left(\lambda_2(r_1 = 0)\right) - v^B\left(\lambda_2(r_1 = 1)\right)$$

Thus $\bar{\mu}_R$ is solution to:

$$\Pi_G(s_1 = 1) = \Pi R_G(s_1 = 1)$$

5.6 Proof of Proposition 6

To determine the equilibrium existence in the general case I follow the same procedure as in the truthtelling equilibrium with the difference that I capture both the bad experts' discipline effect but also the good expert's political correctness.

For $\mu_R < \bar{\mu}_R$ the good expert distorts her signal when 0 and she reports it truthfully if 0: $\pi_G(0) = 0$ but $\pi_G(1) > 0$.

The equilibrium strategies $\pi_G^*(1)$, $\pi_B^*(0) = \pi_B^*(1)$ are solution of the system of equations:

$$\begin{aligned}\Pi_B(s_1 = 0) &= \Pi R_B(s_1 = 0)\\ \Pi_G(s_1 = 1) &= \Pi R_G(s_1 = 1)\end{aligned}$$

It's important to realize that, $\pi_G(0) = 0$ so:

$$\Pr(r_1 = 1|x = 1) = \begin{bmatrix} p\left(\lambda_1\pi_G(1) + (1-\lambda_1)\pi_B(1)\right) \\ +(1-p)(1-\lambda_1)\pi_B(0) \end{bmatrix}$$

$$\Pr(r_1 = 1|x = 0) = \begin{bmatrix} p(1-\lambda_1)\pi_B(0) \\ +(1-p)\left(\lambda_1\pi_G(1) + (1-\lambda_1)\pi_B(1)\right) \end{bmatrix}$$

The non-informative equilibrium arrises in the situation in which political correctness takes full hold of good experts behavior. As a result no expert ever declares her possible bias. The lower weight bounds $\underline{\mu}^G$ which trigger this type of non-informative equilibrium are given by the indifference conditions $\Pi_G(s_1 = 1) = \Pi R_G(s_1 = 1)$ evaluated at babbling strategies $\pi_G(0) = \pi_G(1) = \pi_B(1) = \pi_B(0)$.