
Munich Personal RePEc Archive

Discrimination and Freedom of Speech: Is there a Benefit from Political Correctness?

Gomez-Ruano, Gerardo

June 2012

Online at https://mpra.ub.uni-muenchen.de/93885/
MPRA Paper No. 93885, posted 13 May 2019 15:31 UTC

Discrimination and Freedom of Speech
Is there a Benefit from Political Correctness?

Gerardo Gomez-Ruano *†

June 2012

Abstract

This paper shows how Political Correctness may eliminate market-discrimination (i.e., wage-gaps). The rationale for this effect is that speech may permit others to gauge the viability of discriminatory actions in a social context. Political Correctness can, hence, increase the welfare of the discriminated group *despite* speech being per-se harmless. The paper contributes to the literature on Political Correctness, by pointing out non-trivial benefits from it; to the literature on discrimination, by suggesting an alternative mechanism for discrimination; and to legal theory, by exposing some trade-offs involved with the freedom of expression. The model is consistent with real-life phenomena like the unequal "forbidding of words", the social segregation and integration patterns, and the failure of hate-speech laws to reduce discrimination.

JEL: J15, J7, K42

Keywords: Political Correctness, Free Speech, Discrimination, Imperfect Enforcement.

1 Introduction

Political Correctness can be defined as informal censorship of speech:[1] it is not legally enforced, it does not follow any written rules, and it is usually backed up by sanctions that are socially enforced—like ostracism or economic embargoing.

*Banco de México. Email: ggomezr@alum.bu.edu

†I would like to thank Kevin Lang, participants at the Boston University Microeconomics Theory Workshop, participants of the seminar at Banco de México as well as Carlos Lever.

[1]While the phenomenon of informal censorship is centuries old, the term "political correctness" is relatively new. For an account of its probable origins—around 1940—see Feldstein (1997).

What makes Political Correctness such an interesting phenomenon is its deep involvement with beliefs shared by the community—beliefs about the qualities of those who stick to the "correct" expressions and those who don't.

The complexity of the Political Correctness phenomenon is portrayed in Loury (1994). Therein one can read about the subtleties of language in the post-war Germany, and about the contradictions involved in the "Sanctions Debate" (when the struggle of blacks in South Africa required the U.S. government to impose trade sanctions), among other historical anecdotes of the Political Correctness phenomenon.

Loury noted in his essay that he had mostly ignored any possible benefits of Political Correctness, and that these remained to be investigated. As for the costs, he did not provide any explicit modeling of the phenomenon; instead, he pointed out that the theory of conformity proposed by Bernheim (1994) could be adapted for this purpose.[2]

The main contribution of this paper lies precisely in proving the existence of benefits from Political Correctness under non-trivial assumptions.[3] This is done by showing how socially-enforced censorship of speech may reduce, or even eliminate, output-wasteful discrimination in a particular instance, namely at the workplace.

The paper also contributes to the search for alternative formulations of prejudice. As Charles and Guryan (2007) report, the Becker-mechanism of discrimination accounts for only one quarter of the wage-gap, and is a particular form of racial animus: an aversion for cross-racial contact. The mechanism presented herein provides testable implications, and might account for part of the unexplained wage-gap. Unlike the two most popular discrimination theories—distaste for association and statistical discrimination—this paper's theory requires both, assumptions on preferences *and* on technology. If found to be consistent with the data—this would encourage further research along both types of assumptions.

Last but not least, this article also contributes to the law literature concerning the tension between free-speech and non-discrimination,[4] and more generally to the legal theory, by exposing some trade-offs involved with freedom of expression.

[2] A few years later, Morris (2001) modeled the costs of Political Correctness, as portrayed by Loury, in a game where an informed advisor wishes to convey her valuable information to an uninformed action-taker with identical preferences; the conflict arises because the action-taker believes there is a positive probability that the advisor is actually "bad" (i.e., has the opposite preferences). As a consequence, reputational concerns will give the "good advisor" an incentive to send sometimes false signals. If these reputational concerns are sufficiently important, the result goes, no information is conveyed in equilibrium (just like Loury contended).

[3] The benefits are shown to exist *even if language is per se harmless.*

[4] See, for example, Coliver et al (1992).

The acknowledgement of this tension between free-speech and non-discrimination dates back to (at least) 1951: the year of the controversial Supreme Court case Dennis v. United States.[5] The preamble to this case is that american leaders of the Communist Party were indicted and convicted "for willfully and knowingly conspiring [...] to teach and advocate the overthrow and destruction of the Government of the United States by force and violence". The Supreme Court was then asked to determine whether the Smith Act (under which the leaders where convicted) violated the First Amendment (which guarantees the right to free speech).

While the court was divided in the decision, most Justices agreed that free speech is not an absolute right, and their differences of opinion had more to do with *where* to draw the line. Implicit in the main opinion of the case was the argument that due to the stealthy nature of communist activists ("the Communist Party is a highly disciplined organization, adept at infiltration into strategic positions, use of aliases, and double meaning language; [...] the Party is rigidly controlled"), the state could not wait until right before a putsch ("Obviously, the words cannot mean that, before the Government may act, it must wait until the putsch is about to be executed").Overall, the statements from the case suggest that *it is the limited monitoring capacity of the state what makes speech less of an absolute right*: in a world with perfect monitoring/law enforcement, the state could permit any utterances yet guarantee that these would never end up in acts of crime; in a world with imperfect monitoring/law enforcement though, limiting speech may limit crime as well, thus freedom of speech becomes relative to the benefits of restriction.

This paper captures the concerns of the Supreme Court by incorporating imperfect monitoring/enforcement and proving how these concerns may indeed be effectively alleviated by the restriction of speech.

The two pillars for the mechanism that is posited by the paper are partial-misanthropic preferences and imperfect enforcement.[6] The model provided in the paper should be seen *only as an illustration of this mechanism.*

The particular game-theoretic model is as follows. It is assumed that there are two groups of workers in the population, and that they both have distinct identifiable traits. Following the conventions in the discrimination literature, suppose there is a group of black workers and a group of white workers. Furthermore, suppose that *not all but some* of the white workers dislike black workers (henceforth "racist whites"). This trait (being

[5]All the corresponding quotations come from Dennis v. United States, 341 U.S. 494 (1951), both the Syllabus and the main Opinion of the case.

[6]The "partial-misanthropic" qualifier refers to the fact that the individual is not entirely misanthropic but just misanthropic with regard to a proper subset of humankind. A more colloquial / less pretentious term would perhaps be "spiteful preferences", or simply "hatred" as in Glaeser (2005).

racist) is private information though. In other words, there are both racist whites and non-racist whites, but it is impossible to tell them apart. A key premise of the mechanism is that enforcement of property rights is imperfect. In the particular case of this model, enforcement is imperfectly done by the workers themselves: in general, all the workers like to enforce property rights, and will denounce any law-breaker that they catch; but racist whites differ in that they do not denounce the sabotage of blacks, and they themselves like to sabotage blacks.[7]

Typically, the three types of workers are found in the workplace, yet they do not necessarily know each other's types (again, it is not possible to tell if whites are racists or not just by appearance).

Intuitively, if the white population is known to be mostly racist, racists will "feel at home" and tend to sabotage black workers freely. But, if the white population is more evenly split between racist and non-racist types, those few racists in the population will face a dilemma. There is still a chance that their white coworkers are racist as well (they are only a sample of the population). But if they choose to sabotage and their white coworkers are not racist, they will likely get caught and be punished. They would therefore be much better off if they could only tell the type of their white coworkers. One way to do this is by noticing whether these coworkers use racial slurs in their daily conversations.

The first result of the paper shows that if racists stick to a racist speech-code (or, put simply, a speech-code that differentiates them from non-racists), they will—some of the time—indeed be able to sabotage blacks with a very small chance of getting caught. Hence, blacks will end up being sabotaged some of the time, thus producing (and earning) less, on average.

Again, the intuition is straightforward: racists will only sabotage if they "sense" enough support from the rest of the white workers. This gauging of the support occurs precisely by paying attention at the messages that their white coworkers send. A racist white is likely to send a racist message, a non-racist is likely to send a non-racist message.

In real life, the racist or non-racist character of messages is predetermined by historical evidence.[8] The model will therefore take this character of messages as given. Thus obviating the artifact of having nature choose which message is considered racist.

Given the aforementioned result, it becomes tempting to eliminate sabotage by having a "safety quota" of black workers in every firm (if that firm is to have black workers at

[7]The idea that sabotage can take place at the workplace is not new. See, for example, Lazear (1989).

[8]Indeed, as expressed in the main opinion of Dennis v. United States: "Nothing is more certain in modern society than the principle that [...] a name, a phrase, a standard has meaning only when associated with the considerations which gave birth to the nomenclature."

all). The second result of the paper formalizes this possibility.

Unfortunately, safety quotas are likely to produce partial segregation, thus undermining or limiting social cohesion. If only speech was absent, blacks would avoid the type of discrimination mentioned before; but speech cannot be eliminated by decree. Could it be censored by citizens themselves? And if so, would non-discrimination still follow?

The answer is found in the third result of the paper, which gives sufficient conditions for discrimination to be eliminated when the white non-racist workers selectively censor speech by sanctioning "politically-incorrect" utterances.[9]

The reduction or elimination of discrimination obtained does not follow mechanically though; it depends crucially on the preexisting levels of racism (both the extent and intensity). In fact, in some circumstances, Political Correctness could "backfire" and further increase discrimination. The last result of the paper covers this possibility.

Thus, the model provides some testable implications for empirical research; and, as usual, applies to any type of discrimination based on *identifiable* traits, be it gender, race, height, etc.[10]

Together, the results of the paper are quite useful in explaining recurrent phenomena, like the following.

Forbidden Words It is often contended by self-proclaimed non-racists that "it is not fair" when they are sanctioned for using the same words that blacks use all the time without consequence. These individuals consider that their rights are violated, and that these violations are selective—hence discriminating—since they do not apply to the first discriminated group: "how come black people can use the 'N' word and I can't?"

This observation is consistent with the model. In it, blacks need not be censored because their speech has no further consequence; it does not reveal anything. On the other hand, in the model, even if a white worker is not racist she cannot utter the "forbidden word(s)" because no one knows if she is actually racist or not, and therefore no one knows if she is—or not—actually "encouraging" other racists to sabotage.

[9]The result assumes that only non-racist whites censor/sanction, thus avoiding any concerns of directed speech (e.g. what if racists only talk when no blacks are around?). Needless to say, the result only strengthens if blacks censor/sanction as well.

[10]Identifiable traits are not limited to the phenotype; nationality, religion, sexual preferences, socioeconomic background, and others can often be (if *only* noisily) identified from accents, tastes, clothes, manners, and so on. Often, individuals could "pass" for other types, but they may not want to do so despite the bearing.

Integration vs Segregation As pointed out some paragraphs before, one of the advantages of Political Correctness is that it allows for integration. This should imply at least some correlation between regimes of self-censorship and integration. The historical accounts for the U.S. seem to support this assertion: when the use of racial slurs was quite common and went unpunished, there was clearly more segregation than now, when racial slurs are not as common and tend to be—at least socially—punished.

Relevance of Workplace-Size Although the mechanism can be explained with only three workers (and the reader has probably figured this out by now), the model we employ for illustration includes the size of the workplace (i.e., the number of workers in a common area). Intuitively, the "sanction" of "incorrect" messages can be smaller, the smaller the crowd is (i.e., very subtle when in *petit comité*); this is mirrored by the model and is just one example of workplace-size effects that can be found.[11]

The Failure of Hate-Speech Laws to Reduce Discrimination As we have argued before, speech cannot be eliminated by decree. *It is practically impossible to enforce hate-speech laws because speech is ubiquitous.* And so, except for mainstream media, the presence of hate-speech cannot be effectively eliminated by traditional enforcement instruments (e.g., police surveillance). Censorship by the citizens themselves, however, might get around this.

Coliver (1992) offers plenty of evidence that hate-speech laws have not worked (particularly in countries with considerable preexisting discrimination), sometimes even ending up in more discrimination.

The structure of the paper is as follows. Section 2 is a brief preamble to the game-theoretical model. Section 3 presents the game itself. Section 4 states the result that in a setting with imperfect enforcement and free speech, on-the-job discrimination is close to unavoidable; it also presents the short result that the use of quotas would eliminate on-the-job discrimination. Section 5 has non-racist white workers selectively censor speech, and presents the main result which establishes sufficient conditions for discrimination to be eliminated. Section 6 touches upon the paradoxical fact that Political Correctness might further hinder the discriminated group. Section 7 concludes.

[11] Throughout the paper we employ only one set of 'benchmark' assumptions regarding both preferences and sanctioning technology. The model, however, is flexible enough so as to allow future experimentation and/or calibration of these assumptions.

2 The Story

The model of this paper is a stylized version of a day at work. First, it assumes that there are three types of workers: non-racist blacks (BN) non-racist whites (WN), and racist whites (WR). They all have the same productive potential and only differ in two ways: their skin color, and their preferences with regard to their coworkers.

In the usual taste-based models, racists experience disutility from having contact with blacks. In this model, however, racists get disutility from nearby blacks' wellbeing.[12] This is modeled by having racists' preferences equal non-racists' preferences minus a term reflecting blacks' well-being. Thus, if $U_N(\cdot)$ is the payoff of a non-racist player, then a racist player will have a payoff $U_{WR}(\cdot) = U_N(\cdot) - \beta\ \overline{U_B(\cdot)}$, where $\overline{U_B(\cdot)}$ is the average payoff for players of black type and β is the intensity of racism.[13]

The utility function $U_N(\cdot)$ has both market goods/bads and non-market goods/bads as arguments. In this particular model, market goods/bads are in the form of a paycheck, and non-market goods/bads are in the form of punishments/scoldings (at the workplace) by the boss and (social) sanctions (at the workplace) by coworkers.

The model begins with the hiring of a finite number of workers at random[14] according to the distribution (F_{BN}, F_{WN}, F_{WR}), where F_θ is the percentage of the population of type θ.

Next, all workers share the same workplace, they see each other and are able to identify each other's skin color.

Following that, workers produce one unit of output each.

At any workplace, employees have conversations from time to time; in the model this is introduced as a simultaneous talk, where everyone sends a public message.[15] Under Political Correctness, which will not be covered until section 5, the utterance of a message that is considered politically incorrect (for historical/exogenous reasons) leads to a social sanction $c > 0$ by every non-racist white. Until that section though, $c = 0$ so that there is absolute freedom of speech.[16]

[12] As to where do these type of (partial-misanthropic) preferences come from, Glaeser (2005) gives an interesting suggestion. Grossly speaking, whenever there is a politically relevant and socially isolated group, there are incentives (e.g., for politicians) to supply hate-creating stories about them. Glaeser provides some historical evidence supporting this. In his model the utility function is "U = Income Net of Taxes and Transfers + Expected Damage from the Out-Group − [Other terms]"; this specification is isomorphic to ours.

[13] The average payoff was chosen because it is scale invariant (doubling the number of players does not affect payoffs) and discriminated-group-share invariant. This delivers a parsimonious benchmark.

[14] That is, there is no discrimination in the hiring decision.

[15] That is, each message is heard by every player.

[16] The paper employs a binary message space. The results carry on for more general message spaces.

Also at any workplace, employees have to leave their working station unattended for a while (to go to the bathroom, to go get some coffee, etc.). These moments represent an opportunity for racists to sabotage. In the model, this opportunity is simply modeled as a choice to wreck δ units of someone's work:[17] specifically, worker i chooses the action a_i to attack player j ($a_i = j$) or to not attack anyone at all ($a_i = p$). The symbol p stands for "peaceful action".[18]

The decision to sabotage is not trivial though, since there is a risk of being caught and denounced, which translates into receiving a punishment tantamount to not being paid any wage at all.

Regarding the risk of getting caught, the main premise is that workers generally do not like to sabotage each other and denounce any law-breakers that they catch, the exception being that racist-whites do not denounce the sabotage of blacks and they themselves like to sabotage blacks.

There are many different plausible specifics satisfying the aforementioned premise and delivering a similar result. We employ the one set of specifics that seemed most straightforward and tractable to us: (a) Anyone sabotaging a white worker will be caught with probability one, and (b) anyone sabotaging a black worker will be caught with a probability equal to the share of non-racist workers.

Finally, each worker gets paid according to her/his observed output (which could be less than originally produced if the worker was sabotaged). That is, we assume workers are paid according to their individual, observable performance. Of course there are many times when the performance of the group (as a whole) matters, but this paper will abstract from those cases.

3 The Game

There are n players, and nature moves both at the beginning and at the end of the game. The set of players is symbolized by $\mathcal{N}$, thus $\mathcal{N} = \{ 1, 2, \ldots, n \}$. Each player is of one of these types: BN, WN, WR. The first letter of the type is observable (public information), while the second is not (private information).

The order of the moves is as follows.

1. Nature chooses the vector $\boldsymbol{\theta}$ from $\boldsymbol{\Theta} \equiv \{BN, WN, WR\}^n$ according to the common prior.

[17] The terms "sabotage", "attack", and "wreck" will be used interchangeably.

[18] The game considers specifically one-on-one sabotage. Multiple sabotage—with the correspondingly higher risk—yields similar results.

The common prior is a discrete probability distribution given by the mapping[19] $P_{\boldsymbol{\theta}} : 2^{\boldsymbol{\Theta}} \longrightarrow [0, 1]$.

In what follows, $P_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}))$ stands for the probability of the set $\mathcal{S}$ of vectors which satisfy the logical statement $L(\boldsymbol{\theta})$, with $\mathcal{S} \subseteq 2^{\boldsymbol{\Theta}}$. Similarly, $P_{\boldsymbol{\theta}}(L_1(\boldsymbol{\theta})|L_2(\boldsymbol{\theta}))$ stands for the probability of the set $\mathcal{S}'$ of vectors which satisfy logical statement $L_1(\boldsymbol{\theta})$ and is a subset of the set $\mathcal{S}''$ of vectors which satisfy logical statement $L_2(\boldsymbol{\theta})$, with $\mathcal{S}'$ and $\mathcal{S}''$ such that $\mathcal{S}' \subseteq \mathcal{S}'' \subseteq 2^{\boldsymbol{\Theta}}$.[20]

Let $\mathbf{x}_{-i}$ stand for the "subvector" that results from extracting the ith component from vector $\mathbf{x}$, for any vector $\mathbf{x}$.

The common prior satisfies:

(a) $P_{\boldsymbol{\theta}}(\theta_i = BN|\boldsymbol{\theta}_{-i} = \bar{\boldsymbol{\theta}}_{-i}) = F_{BN}$ for every $i \in \mathcal{N}$ and any $\bar{\boldsymbol{\theta}}_{-i} \in \{BN, WN, WR\}^{n-1}$.

(b) $P_{\boldsymbol{\theta}}(\theta_i = WN|\boldsymbol{\theta}_{-i} = \bar{\boldsymbol{\theta}}_{-i}) = F_{WN}$ for every $i \in \mathcal{N}$ and any $\bar{\boldsymbol{\theta}}_{-i} \in \{BN, WN, WR\}^{n-1}$.

(c) $P_{\boldsymbol{\theta}}(\theta_i = WR|\boldsymbol{\theta}_{-i} = \bar{\boldsymbol{\theta}}_{-i}) = F_{WR}$ for every $i \in \mathcal{N}$ and any $\bar{\boldsymbol{\theta}}_{-i} \in \{BN, WN, WR\}^{n-1}$.

(d) (F_{BN}, F_{WN}, F_{WR}) is a probability distribution, i.e., $F_{BN}, F_{WN}, F_{WR} \geq 0$ and $F_{BN} + F_{WN} + F_{WR} = 1$.

In other words, the type of player i is i.i.d. with probabilities F_{BN}, F_{WN}, F_{WR} for each corresponding type.

Let $I(L)$ be the indicator function, which takes the value one when the logical statement L is true and zero otherwise. Let $v_i \equiv I(\theta_i = BN)$ and let $\boldsymbol{v} \equiv (v_1, v_2, \ldots, v_n)$.
After Nature's move, information is released as follows:
Player i knows θ_i and $\boldsymbol{v}$; for all $i \in \mathcal{N}$.

In other words, players know their own type and whether the other players are black or white.

2. Players simultaneously move as follows:
Player i sends a message $m_i \in \{M_1, M_2\}$; for all $i \in \mathcal{N}$.

After all players simultaneously move, information is released as follows:
Player i knows the entire vector $\boldsymbol{m} \equiv (m_1, m_2, \ldots, m_n)$; for all $i \in \mathcal{N}$.

3. Players simultaneously move as follows:
Player i chooses an action $a_i \in \mathcal{A} \equiv \mathcal{N} \cup \{p\}$; for all $i \in \mathcal{N}$.

[19] We write 2^S to denote the set of all sets found in S, for any set S. Formally, $2^S \equiv \{X \mid X \subseteq S\}$. This is sometimes known as the power set of S.

[20] By the definition of conditional probability, one has that $P_{\boldsymbol{\theta}}(L_1(\boldsymbol{\theta})|L_2(\boldsymbol{\theta})) = \frac{P_{\boldsymbol{\theta}}(L_1(\boldsymbol{\theta}) \text{ and } L_2(\boldsymbol{\theta}))}{P_{\boldsymbol{\theta}}(L_2(\boldsymbol{\theta}))}$.

Let $\boldsymbol{a} \equiv (a_1, a_2, \ldots, a_n)$.
No information is released.

4. Nature moves by choosing a vector $\boldsymbol{b} \in \{0,1\}^n$ according to the $\boldsymbol{\theta}$-dependant, discrete probability function $P_{\boldsymbol{b}} : 2^{\{0,1\}^n} \longrightarrow [0,1]$.

   Let $\#\mathcal{S}$ stand for the cardinality (the size) of set $\mathcal{S}$, for any set $\mathcal{S}$. $P_{\boldsymbol{b}}$ satisfies:[21]

   - $P_{\boldsymbol{b}}(b_i = 1 \mid \boldsymbol{b}_{-i} = \bar{\boldsymbol{b}}_{-i}) = 1 - P_{\boldsymbol{b}}(b_i = 0 \mid \boldsymbol{b}_{-i} = \bar{\boldsymbol{b}}_{-i}) = \frac{\#\{j \in \mathcal{N} | \theta_j \neq WR\}}{n}$ for every $i \in \mathcal{N}$ and any $\bar{\boldsymbol{b}}_{-i} \in \{0,1\}^{n-1}$.

   In other words, b_i is an i.i.d. Bernoulli draw with success probability $\frac{\#\{j \in \mathcal{N} | \theta_j \neq WR\}}{n}$.

   After nature moves, terminal nodes are reached and payoffs are given according to the Payoff Function that follows.

Define $\mathcal{N}_\theta = \{\, j \in \mathcal{N} \mid \theta_j = \theta \,\}$ to shorten the notation. *Without loss of generality, and for the whole paper, let M_1 be the exogenously/historically determined "politically incorrect" message.*

The Payoff function for player i, $U_i : \{BN, WN, WR\}^n \times \{0,1\}^n \times \{M_1, M_2\}^n \times \mathcal{A}^n \longrightarrow \mathbb{R}$, is given by:

$$
\begin{aligned}
U_i(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) = & \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] \Big(1 - I(a_i \neq p)\big[I(a_i \in \mathcal{N}_{BN}) b_i + I(a_i \notin \mathcal{N}_{BN})\big]\Big) \\
& - I(\theta_i \neq BN \text{ and } m_i = M_1)\, c\, \frac{\#\mathcal{N}_{WN}}{\max\{1, \#\{j \in \mathcal{N} | \theta_j \neq BN \text{ and } m_j = M_1\}\}} \\
& - I(\theta_i = WR) \frac{\beta}{\#\mathcal{N}_{BN}} \sum_{\theta_j = BN} U_j(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) \; .
\end{aligned}
$$

This is the general payoff function. It will be described in greater detail in sections 4 and 5.

The structure of the game, including the number of players and the probabilities for both of nature's moves, is common knowledge.

Only type-symmetric equilibria in pure strategies are considered in the paper, and the proofs employ Weak Perfect Bayesian Equilibrium.[22] We focus on "type-symmetric" equilibria—which means that individuals of the same type have the same strategies—because we care about the behavior of agents as a function of their type (and nothing

[21] We use the same notational convention as for $P_{\boldsymbol{\theta}}$ above.

[22] Nonetheless the proofs can be extended to use Sequential Equilibrium.

else). That is, we only care about the behavior of a worker given that he is a racist-white, a non-racist white, or a non-racist black. The paper focuses only on pure strategies for simplicity.

3.0.1 The Wage and its Expectation, the Speech Regimes, and the Parameter Space

The wage of player i, written w_i, is the first factor in the first summand of the payoff function. That is,

$$w_i = \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] .$$

This is what i's observable output amounts to at the end of the game: one unit of output minus that destroyed by the others.[23]

Both the social sanction for sending a politically incorrect message and the punishment for sabotaging are assumed non-pecuniary, for matters of wage comparison.[24]

Let $\mathbf{\Omega} \equiv \{BN, WN, WR\}^n \times \{0,1\}^n$, which is the set of all the possible two moves that nature can make. We write $\tilde{\boldsymbol{\omega}} \equiv (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{b}})$ for any element of $\mathbf{\Omega}$, and we may call the former a state and the latter the state-space. We write $\boldsymbol{\omega} \equiv (\boldsymbol{\theta}, \boldsymbol{b})$ for the chosen moves by nature, and we may call it the realized state.

The expected wage for players of type θ is defined as the expectation of the average wage for the players of type θ in state $\tilde{\boldsymbol{\omega}}$, taken over all $\tilde{\boldsymbol{\omega}} \in \mathbf{\Omega}$ according to the probability distribution given in the first part of this section.[25] That is,

$$E[w_\theta] = E_{\boldsymbol{\omega}} \left[\frac{1}{\#\mathcal{N}_\theta} \sum_{\theta_i = \theta} w_i\right] = \int_{\mathbf{\Omega}} \left[\frac{1}{\#\mathcal{N}_\theta} \sum_{\theta_i = \theta} w_i\right] dP(\tilde{\boldsymbol{\omega}}) ,$$

where the measure P is composed of both $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{b}}$.

[23] The non-negativity constraint was left out for simplicity. The results carry on regardless.

[24] Notice non-racist whites will sanction politically incorrect whites (white senders of message M_1) even if that includes themselves. This was left for the sake of simplicity; all results go through if one eliminates self-sanctioning. Similarly, notice sabotage is punished even if done to oneself. Again, this was left for the sake of simplicity; all results remain the same if one eliminates the punishments in these cases.

[25] Of course, wages are determined by the strategies, the parameters, and the moves by nature, so—strictly speaking—the left-hand-side should have the strategies, the prior, and all the parameters as arguments; this has been left out to keep the notation short.
Thus, expected wages will vary depending on the equilibrium being considered, and this definition will allow us to make comparisons between different equilibria and/or parameters.

The set of possible values for the social sanction (i.e., the non-negative real numbers) can be partitioned into two sets (or regimes)—each with an intuitive interpretation: the "freedom of speech" regime (where $c = 0$) and the "political correctness" regime (where $c > 0$). These two cases are dealt with separately in sections 4 and 5 respectively.

The remaining exogenous variables ($\beta, \delta, F_{BN}, F_{WN}, F_{WR}$ and n) are called the parameters. *The set of parameters satisfying the following restrictions is called the parameter-space*:

- $\mathbf{F}$ has full support,
- $0 < \beta, \delta < 1$, and
- $n \geq 3$;

where the distribution $\mathbf{F} = (F_{BN}, F_{WN}, F_{WR})$ is said to have full support if all types have a strictly positive probability; that is, if $F_{BN}, F_{WN}, F_{WR}, > 0$.

4 Free Speech

Under Free Speech, $c = 0$, which means there is no censorship; the payoff functions are much simpler. For both of the non-racist types the payoff becomes

$$U_i(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) = \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] \Big(1 - I(a_i \neq p)\big[I(a_i \in \mathcal{N}_{BN}) b_i + I(a_i \notin \mathcal{N}_{BN})\big]\Big) .$$

The first factor is simply the wage that i gets. As previously said, this is the observed output at the end of the game—one minus delta times the number of workers that sabotaged i.

The second factor is one minus the existence of a non-pecuniary punishment for sabotage (i.e., if it exists the value is one, otherwise it's zero). This punishment takes place when i sabotages ($a_i \neq i$) a black worker ($a_i \in \mathcal{N}_{BN}$) with probability $\Pr[b_i = 1] = \frac{\#\{j \in \mathcal{N} | \theta_j \neq WR\}}{\#\mathcal{N}}$, and when i sabotages ($a_i \neq i$) a white worker ($a_i \notin \mathcal{N}_{BN}$) with probability one.

On the other hand, the racist's payoff is

$$\begin{aligned} U_i(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) = & \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] \Big(1 - I(a_i \neq p)\big[I(a_i \in \mathcal{N}_{BN}) b_i + I(a_i \notin \mathcal{N}_{BN})\big]\Big) \\ & - \frac{\beta}{\#\mathcal{N}_{BN}} \sum_{\theta_j = BN} U_j(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) , \end{aligned}$$

consistent with section 2's expression: $U_{WR}(\cdot) = U_N(\cdot) - \beta\overline{U_B(\cdot)}$.

The next proposition establishes the aforementioned perverse quality of the free-speech regime.

Proposition (Open Doors for Discrimination under Free Speech)**.** *For the whole parameter-space:*
If $\beta\delta \geq 1/n$, then there exists an equilibrium where the expected wage of a black type is strictly less than that of any white type.

Notice that the proposition remains true no matter how small the share of the population with racist preferences is.

Surely, one could argue that this outcome cannot be an economic equilibrium because "in real life" under-payed black workers would switch to better work environments. But this will not be the case in an economy where mobility is costly, and/or where the expected benefit from moving to another job is not big enough as to offset this moving cost.[26]

All that is necessary for this equilibrium to hold is for *WR*s to send a different message than for *WN*s—indeed this is an equilibrium that is separating in messages from the white workers. This separation is possible because there is *free speech*, i.e., there is no punishment for sending either message or, more precisely, for sending a different message than the one other types send. Given this revealing mechanism and the fact that types are distributed randomly, there is always a positive probability that a black type will be surrounded by racist whites. Since racist whites reveal themselves, they will know that they constitute the majority and that the chance of "getting caught" is minimal, thus making it worthwhile to sabotage the black type.

One would naturally expect that avoiding situations where racist whites are a majority could eliminate discrimination in the present setting. Unfortunately, racist whites are indistinguishable from their non-racist counterparts; therefore employers would have to restrict the share of *all* white workers (or alternatively require a minimum share of black workers). Let $\hat{F}_\theta = \frac{\#\mathcal{N}_\theta}{\#\mathcal{N}}$. The next claim formalizes this possibility.

Claim 1 (Effectiveness of Quotas)**.** *For the whole parameter-space:*
Let employers be restricted to hire such that either zero or more than $\sqrt{\beta\delta n}$ of their n workers are black. Then, discrimination is eliminated, i.e., wages are equal for all types in any equilibria and any realized state.

It is therefore sufficient to have the share of black workers be weakly greater than $\sqrt{\beta\delta}$ or equal to zero. The intuition for the claim is that there is "safety in numbers". But,

[26]The notion that mobility or information costs perpetuate differentials due to discrimination is formalized in Black (1995).

importantly, a policy like this implies—in general—partial segregation since, whenever $0 < F_{BN} < \sqrt{\beta\delta/n}$, not all of the firms will be able to have a mixture of blacks and whites; some will have only whites despite this being an unlikely outcome under random hiring with a high n, for example. If there are any reasons to avoid segregation, then those same reasons would make quotas undesirable as well.

5 Political Correctness

Under the Political Correctness regime, $c > 0$, which means that non-racist whites will sanction any white worker (they cannot know which one is or not racist) that utters the exogenously/historically determined "politically incorrect" message M_1. The fact that $c > 0$ is not enough for discrimination to disappear though

The payoff functions become more complicated to take the sanction into account. For non-racist black types the payoff function stays as before, but for non-racist whites it gets a new, second line:

$$U_i(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) = \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] \Big(1 - I(a_i \neq p)\big[I(a_i \in \mathcal{N}_{BN}) b_i + I(a_i \notin \mathcal{N}_{BN})\big]\Big)$$
$$- I(\theta_i \neq BN \text{ and } m_i = M_1)\, c\, \frac{\#\mathcal{N}_{WN}}{\max\{1, \#\{j \in \mathcal{N} | \theta_j \in \{WN, WR\} \text{ and } m_j = M_1\}\}} .$$

The second line stands for the social sanction that *any* white worker who sends a politically incorrect message gets from every non-racist white that is present. Because more than one worker may send the politically incorrect message, it is assumed that sanctions are evenly split across all the whites who utter M_1. Racists whites have—on top of this—the usual disutility they get from blacks' well-being:

$$U_i(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) = \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] \Big(1 - I(a_i \neq p)\big[I(a_i \in \mathcal{N}_{BN}) b_i + I(a_i \notin \mathcal{N}_{BN})\big]\Big)$$
$$- I(\theta_i \neq BN \text{ and } m_i = M_1)\, c\, \frac{\#\mathcal{N}_{WN}}{\max\{1, \#\{j \in \mathcal{N} | \theta_j \in \{WN, WR\} \text{ and } m_j = M_1\}\}}$$
$$- I(\theta_i = WR) \frac{\beta}{\#\mathcal{N}_{BN}} \sum_{\theta_j = BN} U_j(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{m}, \boldsymbol{a}) .$$

Define $F_{\mathcal{W}} \equiv F_{WR} + F_{WN}$ to simplify notation. The main result of the paper follows.

Theorem (Effectiveness of Political Correctness)**.** *For the whole parameter-space: If*

$$\frac{F_{WR}}{F_{\mathcal{W}}} + \beta\delta \leq \frac{1}{2} \tag{LER}$$

and

$$c > n \cdot \beta\delta \cdot \left(\frac{F_{WR}}{F_{WN}}\right) \tag{HES}$$

then no discrimination exists, i.e., wages are equal for all types in any realized state of any equilibria.

This result implies that even in the worse possible states, where all white players have racist preferences and are many times the number of black players, no racist white player will dare to perform an aggressive action. Notice that the sufficient lower bound (HES) for the social sanction c is strictly positive, increasing in F_{WR}, and decreasing in F_{WN}.

Thus, Political Correctness eliminates resource-wasteful discrimination if two conditions are met: there is a low enough level of racism (LER) both in extension and intensity, and there is a high enough sanction (HES) for being politically incorrect.

6 Harmful Political Correctness

Before concluding the paper, we have to mention the paradoxical fact that Political Correctness may actually be harmful. Specifically, for high enough levels of racism, a restriction of speech can result in an even higher amount of sabotage and—therefore—higher wage-gaps.

Claim 2. *There exist parameters for which Political Correctness is detrimental. That is, there are parameters for which the Political Correctness regime has a lower expected wage for type BN than the Freedom of Speech regime under some equilibria.*

Intuitively, the randomness in hiring allows the existence of states where, despite the share of racists being high enough to attack without any further information, racists will be too few in the sample, and hence—under full revelation—will not attack. This effect may go both ways, but—in general—does not cancel out. The proof shows this for a particularly extreme case: that with $n = 3$. The reader can imagine how a big enough $F_{WR}/F_{\mathcal{W}}$ might push all racists to attack in states where, under full revelation, they would not have.[27]

[27]For $n = 3$, sabotage can only occur with one black and two white workers, and a racist worker will not attack if he finds out that the other white worker is not racist. See subsection D.1 in the Appendix.

7 Conclusion

This paper showed how political correctness, an implicit restriction of speech, may eliminate discrimination. Thus, a benefit for the discriminated group was obtained despite the harmless nature of speech. In addition, the model was shown to exhibit behavior consistent with real-life phenomena like the unequal "forbidding of words", the facilitating of integration without the use of quotas, and the failure of hate-speech laws.

Unlike Loury (1994) and Morris (2001), the paper equates Political Correctness to a regime where "incorrect speech" is sanctioned, and not to a message-pooling equilibrium; in addition, it defines Freedom of Speech as a regime where no speech is punished at all. This distinction is subtle but indispensable in order to realize that, while a social sanction is indeed a device that may prevent the existence of a white-message-separating equilibrium, it may not always have the "desired effect": On the one hand, if the sanction is not strong enough, message-pooling will not necessarily follow. On the other hand, preventing a white-message-separating equilibrium can—paradoxically—result in a "worse" outcome. Therefore, setting the "Political Correctness vs Freedom of Speech" and the "Message-Pooling vs Message-Separating" dichotomies apart ends up in a useful approach.[28] There is more to the phenomenon than a simple comparison between message-pooling and message-separating equilibria.

Although the paper treated the decision to sanction as exogenous, it could be endogenized by having non-racists receive part of the extra output, as long as the cost of sanctioning remains low enough. This observation, together with the fact that the model does not require the discriminated group to be a minority, suggests that the relevant measure for discrimination is not that of the discriminated group's population-share (i.e., whether they are a minority) but that of the output-loss due to the group's discrimination.

It is extremely important to keep in mind that the censorship discussed herein is enforced by society. As previously remarked, governments and their judicial branches often want to censor "hate-speech" to reduce discrimination, but because of the ubiquitous and hard-to-monitor nature of speech, government-imposed censorship policies that are not widely supported by the population are doomed to fail, and have done so.[29]

When racism has considerable presence among the population, government intervention through the media, education and other preference-modifying methods could prove far more useful than the passing of hate-speech laws.

Finally, it is worth pointing out that the pillars of the paper's mechanism are nothing but a special case of a more general, almost tautological, hypothesis: discrimination arises

[28] Certainly, equating Freedom of Speech to a message-separating equilibrium would give the misleading impression that a babbling equilibrium is absent whenever there is free speech.

[29] Again, see Coliver (1992).

whenever preferences are not 'difference-ignorant' and the enforcement of 'difference-ignorant' behavior is imperfect. This is worth pointing out because it means *other similar though less blatant mechanisms are relevant*—and perhaps more plausible—as well.[30]

[30]Like a simple lack of cooperation instead of outright sabotage.

References

[1] Bernheim, B. D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102, 841–77.

[2] Black, D. A. (1995): "Discrimination in an Equilibrium Search Model," *Journal of Labor Economics*, 13, 309–334.

[3] Charles, K. K., and J. Guryan (2007): "Prejudice and Wages: An Empirical Assessment of Becker's *The Economics of Discrimination*," *Journal of Political Economy*, 116, 773–809.

[4] Coliver, S., K. Boyle, and F. D'Souza, eds. (1992): *Striking a Balance: Hate Speech, Freedom of Expression and Non-discrimination*. London: ARTICLE 19 and Human Rights Center at the University of Essex.

[5] Coliver, S. (1992): "Hate Speech Laws: Do They Work?," in *Striking a Balance: Hate Speech, Freedom of Expression and Non-discrimination*, ed. by S. Coliver, K.Boyle, and F. D'Souza. London: ARTICLE 19 and Human Rights Center at the University of Essex.

[6] Feldstein, R. (1997): *Political Correctness: a Response from the Cultural Left*. Minneapolis, MN: University of Minnesota Press.

[7] Glaeser, E. L. (2005): "The Political Economy of Hatred," *Quarterly Journal of Economics*, 120, 45–86.

[8] Lazear, E. P. (1989): "Pay Equality and Industrial Politics," *Journal of Political Economy*, 97, 561–580.

[9] Loury, G. (1994): "Self-Censorship in Public Discourse," *Rationality and Society*, 6, 428–461.

[10] Morris, S. (2001): "Political Correctness," *Journal of Political Economy*, 109, 231–265.

A Notation for the Proofs

$n_\theta \equiv \#\mathcal{N}_\theta$. In other words, n_θ is the cardinality of the set $\mathcal{N}_\theta$.

$\eta_\theta(\tilde{\boldsymbol{\theta}}_{-i}) \equiv \sum_{j=1}^{n-1} I(\tilde{\theta}_{-i,j} = \theta)$, where $I(\cdot)$ is the indicator function. That is, $\eta_\theta(\tilde{\boldsymbol{\theta}}_{-i})$ is the number of players of type θ in the vector $\tilde{\boldsymbol{\theta}}_{-i}$, for any vector $\tilde{\boldsymbol{\theta}}_{-i} \in \{BN, WN, WR\}^{n-1}$.

$n_{\mathcal{W}} \equiv n_{WR} + n_{WN}$. So $n_{\mathcal{W}}$ is simply the number of white players. Notice $n_{\mathcal{W}} = n - n_{BN} = n - \sum_{j=1}^{n} v_j$, thus $n_{\mathcal{W}}$ is known whenever $\boldsymbol{v}$ is known.

$\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i}) \equiv \eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i}) + \eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})$. Thus, $\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})$ is the amount of white types in the vector $\tilde{\boldsymbol{\theta}}_{-i}$.

$\mathcal{N}_{\mathcal{W}} \equiv \mathcal{N}_{WN} \cup \mathcal{N}_{WR}$. So $\mathcal{N}_{\mathcal{W}}$ is simply the set of white players. It is clearly known whenever $\boldsymbol{v}$ is known.

$\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \equiv \frac{\#\{j \in \mathcal{N} | \bar{m}_j = M_1 \text{ and } \bar{v}_j = 0\}}{n}$. So $\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ is the number of white players—according to $\bar{\boldsymbol{v}}$—who sent the politically incorrect message—according to $\bar{\boldsymbol{m}}$—divided by the total number of players.

$\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}) \equiv \{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta} \mid \tilde{\theta}_i = \bar{\theta}_i \text{ and } I(\tilde{\theta}_j = BN) = \bar{v}_j \text{ for every } j \in \mathcal{N}\}$. That is, $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$ is the information set where player i of type $\bar{\theta}_i$ finds himself after observing $\bar{\boldsymbol{v}}$ (i.e., at the end of stage 1 of the game described in section 3). This information set contains only those nodes $\tilde{\boldsymbol{\theta}}$ consistent with the information player i has up to that point: $\bar{\theta}_i$ and $\bar{\boldsymbol{v}}$.

$\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \equiv \{(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{m}}) \in \boldsymbol{\Theta} \times \{M_1, M_2\}^n \mid \tilde{\theta}_i = \bar{\theta}_i \text{ and } I(\tilde{\theta}_j = BN) = \bar{v}_j \text{ for every } j \in \mathcal{N} \text{ and } \tilde{\boldsymbol{m}} = \bar{\boldsymbol{m}}\}$. Thus $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ is the information set where player i of type $\bar{\theta}_i$ finds himself after observing $\bar{\boldsymbol{v}}$ and "listening" $\bar{\boldsymbol{m}}$ (i.e., at the end of stage 2 in the game described in section 3). This information set contains only those nodes $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{m}})$ consistent with the information player i has up to that point: $\bar{\theta}_i$, $\bar{\boldsymbol{v}}$, and $\bar{\boldsymbol{m}}$.

The rest of the notation employed has been introduced in the main text.

B Proof of the Proposition

The proof consists of three parts. First we propose an equilibrium. Second, we prove that it is indeed an equilibrium. Third, we prove that, under the stated conditions, the equilibrium has the stated implications.

There are many equilibria that can be used to prove the proposition. However, for the sake of simplicity, we present just one that is particularly tractable.

B.1 The Proposed Equilibrium

B.1.1 Strategies

Strategies are as follows. Let $m_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$ be the message that player i sends when at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$. Then it is given by

$$m_i(\bar{\theta}_i, \bar{\boldsymbol{v}}) = m(\bar{\theta}_i, \bar{\boldsymbol{v}}) = \begin{cases} M_1, & \text{if } \bar{\theta}_i = WR; \\ M_2, & \text{otherwise.} \end{cases}$$

Hereafter $\boldsymbol{m}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{v}}) \equiv (m(\bar{\theta}_1, \bar{\boldsymbol{v}}), m(\bar{\theta}_2, \bar{\boldsymbol{v}}), \ldots, m(\bar{\theta}_n, \bar{\boldsymbol{v}}))$.

Let $a_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ be the action that player i takes at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$. Then

$$a_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = a(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \begin{cases} \min \mathcal{N}_{BN}, & \text{if } \bar{\theta}_i = WR \text{ and } n_{BN} > 0 \text{ and } \widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}}; \\ p, & \text{otherwise,} \end{cases}$$

where the players use the relations $\mathcal{N}_{BN} = \{j \in \mathcal{N} \mid v_j = 1\}$, $n_{BN} = \#\mathcal{N}_{BN}$, and $\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \equiv \frac{\#\{j \in \mathcal{N} | \bar{m}_j = M_1 \text{ and } \bar{v}_j = 0\}}{n}$.

B.1.2 Beliefs

Beliefs are as follows. Let $\pi_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}})$ be the probability that player i assigns to being at node $\tilde{\boldsymbol{\theta}} \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$, given that he is at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$. This probability is given by

$$\pi_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}) = \frac{F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})}}{F_{\mathcal{W}}^{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})}} \binom{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})}{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})}^{-1}$$

where $\binom{e}{d}^{-1} = \binom{d}{e} = \frac{d!}{e!(d-e)!}$, for any numbers d, e.

Let $\pi_i'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ be the probability that player i assigns to being at node $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{m}}) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$, given that he is at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ indeed. This probability is given by

$$\pi_i'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \begin{cases} 1, & \text{if } m(\tilde{\theta}_j, \bar{\boldsymbol{v}}) = \bar{m}_j \text{ for all } j \notin \mathcal{N}_{BN}; \\ 0, & \text{otherwise.} \end{cases}$$

B.2 Proof that B.1 is an Equilibrium

For B.1 to be a Weak Perfect Bayesian Equilibrium, beliefs have to be weakly consistent, and strategies have to be sequentially rational.

B.2.1 Weak Consistency of Beliefs

Weak consistency of beliefs means that beliefs have to be consistent with Bayes' rule whenever possible (whenever information sets are reached with positive probability).

Consider first the beliefs given by $\pi_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \overline{\boldsymbol{v}})$. For $\tilde{\boldsymbol{\theta}} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}})$, they have to satisfy

$$\pi_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \overline{\boldsymbol{v}}) = \frac{\Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}]}{\Pr[\boldsymbol{\theta} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}}) \mid P_{\boldsymbol{\theta}}]} \text{ , whenever } \Pr[\boldsymbol{\theta} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}}) \mid P_{\boldsymbol{\theta}}] > 0.$$

Meaning that $\pi_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \overline{\boldsymbol{v}})$ has to be equal to the probability of reaching node $\tilde{\boldsymbol{\theta}}$ divided by the probability of reaching information set $\mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}})$—of which $\tilde{\boldsymbol{\theta}}$ is an element, whenever this information set has a positive probability of being reached. With both probabilities conditional on the prior $P_{\boldsymbol{\theta}}$.

Because of the assumed—common—prior for the game (see section 3), all non-empty information sets $\mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}})$ have a positive probability of being reached.

Given the independence of each player's type, and the rules of probability, it follows that, for any $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$,

$$\begin{aligned}
&\Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}] \\
&\quad = F_{\tilde{\theta}_i} F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})} F_{BN}^{\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})} \left[\frac{n!}{1!\,\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})!} \right]^{-1}.
\end{aligned}$$

Where the term before the brackets is the probability that nature's chosen vector $\boldsymbol{\theta}$ has the same numbers of each type as vector $\tilde{\boldsymbol{\theta}}$, and the second term is one over the number of such—equally likely—vectors. This second term is necessary because there is only one such vector that has $\tilde{\boldsymbol{\theta}}$'s same ordering.

Similarly, it follows that, for any non-empty set $\mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}})$,

$$\begin{aligned}
&\Pr[\boldsymbol{\theta} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}}) \mid P_{\boldsymbol{\theta}}] \\
&\qquad = F_{\bar{\theta}_i} F_{\mathcal{W}}^{n-1-\sum_{j\neq i}\bar{v}_j} F_{BN}^{\sum_{j\neq i}\bar{v}_j - 1} \left[\frac{n!}{1!(n-1-\sum_{j\neq i}\bar{v}_j)!\,(\sum_{j\neq i}\bar{v}_j - 1)!} \right]^{-1}.
\end{aligned}$$

Using both probabilities, we have that

$$\frac{\Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}]}{\Pr[\boldsymbol{\theta} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}}) \mid P_{\boldsymbol{\theta}}]}$$
$$= \frac{F_{\tilde{\theta}_i} F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})} F_{BN}^{\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})} \left[\frac{n!}{1!\,\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})!}\right]^{-1}}{F_{\bar{\theta}_i} F_{\mathcal{W}}^{n-1-\sum_{j\neq i}\bar{v}_j} F_{BN}^{\sum_{j\neq i}\bar{v}_j - 1} \left[\frac{n!}{1!(n-1-\sum_{j\neq i}\bar{v}_j)!\,(\sum_{j\neq i}\bar{v}_j - 1)!}\right]^{-1}},$$

or, simplifying a bit,

$$\frac{F_{\tilde{\theta}_i} F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})} F_{BN}^{\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})} \left[\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})!\right]}{F_{\bar{\theta}_i} F_{\mathcal{W}}^{n-1-\sum_{j\neq i}\bar{v}_j} F_{BN}^{\sum_{j\neq i}\bar{v}_j - 1} \left[(n-1-\sum_{j\neq i}\bar{v}_j)!\,(\sum_{j\neq i}\bar{v}_j - 1)!\right]}.$$

Since $\tilde{\boldsymbol{\theta}} \in \mathcal{I}_i(\bar{\theta}_i, \overline{\boldsymbol{v}})$, it must be that $\bar{\theta}_i = \tilde{\theta}_i$, $\sum_{j\neq i}\bar{v}_j - 1 = \eta_{BN}(\tilde{\boldsymbol{\theta}}_{-i})$, and $n-1-\sum_{j\neq i}\bar{v}_j = \eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})$. Therefore, we have

$$\frac{F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})} \left[\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})!\,\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})!\right]}{F_{\mathcal{W}}^{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})} \left[\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})!\right]}.$$

Using the fact that $\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i}) \equiv \eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i}) + \eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})$ we get

$$\frac{F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})}}{F_{\mathcal{W}}^{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})}} \left[\frac{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})!}{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})!\,\Big(\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i}) - \eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})\Big)!}\right]^{-1},$$

but for d, e integers with $d \geq e \geq 0$ we have that $\frac{d!}{e!(d-e)!} \equiv \binom{d}{e}$. Hence the previous expression becomes

$$\frac{F_{WR}^{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})} F_{WN}^{\eta_{WN}(\tilde{\boldsymbol{\theta}}_{-i})}}{F_{\mathcal{W}}^{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})}} \binom{\eta_{\mathcal{W}}(\tilde{\boldsymbol{\theta}}_{-i})}{\eta_{WR}(\tilde{\boldsymbol{\theta}}_{-i})}^{-1}.$$

This is identical to the beliefs in B.1.2.

Now consider the beliefs given by $\pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$. For $(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ they have to satisfy

$$\pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \frac{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) = (\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}\ , \tag{1}$$

whenever $\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] > 0$. Meaning that $\pi'_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ has to be equal to the probability of reaching node $(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}})$, under the pure substrategies $\boldsymbol{m}(\cdot, \cdot)$, divided by the probability of reaching information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$—of which $(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}})$ is an element, under the pure substrategies $\boldsymbol{m}(\cdot, \cdot)$, whenever the information set has a positive probability of being reached. With both probabilities conditional on the prior $P_{\boldsymbol{\theta}}$.

By lemma 1—on page 27—we have that under equilibrium B.1, whenever the information set has a positive probability of being reached, the beliefs / left hand side of equation (1) become

$$\pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \begin{cases} 1, & \text{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}}; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

So, we need to show that the right hand side of equation (1) is equal to this, whenever the information set has a positive probability of being reached. Let $\boldsymbol{V}(\boldsymbol{x}) \equiv (I(x_1 = BN), I(x_2 = BN), \ldots, I(x_n = BN))$ for any $\boldsymbol{x} \in \boldsymbol{\Theta}$. Ignoring the restriction $\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] > 0$, we have that, for any $(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \in \boldsymbol{\Theta} \times \{M_1, M_2\}^n$,

$$\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) = (\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] = \begin{cases} \Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}], & \text{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \boldsymbol{V}(\tilde{\boldsymbol{\theta}})) = \bar{\boldsymbol{m}}; \\ 0, & \text{otherwise.} \end{cases}$$

Now, let $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ be the element of $\boldsymbol{\Theta}$ that is uniquely defined by[31]

1. $\boldsymbol{V}(\boldsymbol{\vartheta}) = \bar{\boldsymbol{v}}$, and
2. $m_j(\vartheta_j, \bar{\boldsymbol{v}}) = \bar{m}_j$ for all $j \in \mathcal{N}_{\mathcal{W}}$.

Unlike previously, some of the non-empty information sets have zero probability of being reached, because of the—pure—strategies. Ignoring the restriction $\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] > 0$, we have that, for any non-empty set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$,

$$\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] = \begin{cases} \Pr[\boldsymbol{\theta} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}], & \text{if } \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}}; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

[31] See lemma 2—on page 27—for the uniqueness of $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$.

Since $\mathbf{F}$ has full support and $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \in \boldsymbol{\Theta}$ always exists uniquely,[32] it follows that

$$\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}] > 0 \iff \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}} \tag{4}$$

Therefore, *when* $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ *is reached with positive probability*, we have that

$$\frac{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) = (\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]} = \begin{cases} \frac{\Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}]}{\Pr[\boldsymbol{\theta} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}, & \text{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \boldsymbol{V}(\tilde{\boldsymbol{\theta}})) = \bar{\boldsymbol{m}}; \\ 0, & \text{otherwise;} \end{cases}$$

and

$$\boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}}\,.$$

Together both previous equations imply

$$\frac{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) = (\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]} = \begin{cases} \frac{\Pr[\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \mid P_{\boldsymbol{\theta}}]}{\Pr[\boldsymbol{\theta} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}, & \text{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \boldsymbol{V}(\tilde{\boldsymbol{\theta}})) = \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}); \\ 0, & \text{otherwise.} \end{cases}$$

But, by lemma 3 on page 28, we know that whenever $\tilde{\boldsymbol{\theta}} \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$, then$\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) = \bar{\boldsymbol{v}}$ and $\boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \boldsymbol{V}(\tilde{\boldsymbol{\theta}})) = \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}) \implies \tilde{\boldsymbol{\theta}} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$. Therefore

$$\frac{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) = (\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]}{\Pr[(\boldsymbol{\theta}, \boldsymbol{m}(\boldsymbol{\theta}, \boldsymbol{v})) \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \mid P_{\boldsymbol{\theta}}]} = \begin{cases} 1, & \text{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}}; \\ 0, & \text{otherwise.} \end{cases}$$

Which is the same as the beliefs in equation (2).

B.2.2 Sequential Rationality

Sequential rationality means that all players make their moves such that their expected payoff—conditional on their beliefs and everyone else's strategies—is maximized at every information set. In the case of this game, checking sequential rationality does not involve any special technique and is easy, though very tedious, to perform thoroughly. We therefore provide a more heuristic proof.

Note that, for the sake of space, we will proceed heuristically to show sequential rationality since the methods involved are conventional and straightforward.

For player i, of type $\bar{\theta}_i \neq WR$, at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$, it is straightforward to show that the action $a_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ is optimal.

If i chooses an aggressive action—chooses $a_i \neq p$—then, in the best scenario—where i is not caught/punished—he gets a payoff that is exactly equal to the payoff that he would get—with certainty—by choosing the peaceful action—$a_i = p$. Therefore, the peaceful action is optimal.

[32] See lemma 2 for the existence of $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$.

For player i, of type $\bar{\theta}_i = WR$, at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$, it is slightly more involved to show the optimality of its action $a_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$.

Consider the case where $n_{BN} > 0$ and $\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}}$. Then—given the strategies—we have that

$$\begin{aligned} &\frac{n_{WR}}{n} \geq 1 - \frac{\beta\delta}{n_{BN}}\,, \\ \Leftrightarrow\ &\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}\,. \end{aligned}$$

Now, it turns out that the benefit that i gets—with certainty—from sabotaging *any* black worker is precisely the left hand side of this last inequality: $\frac{\beta\delta}{n_{BN}}$. On the other hand, the expected cost is simply the probability that i gets caught times his wage—which under the strategies is equal to one. Therefore the expected cost is $1 - \frac{n_{WR}}{n}$. We know that $\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}$ from the last equation, therefore the aggressive action is optimal in these cases *and only* in these cases. Hence the peaceful action is optimal in the remaining cases.

For player i, of type $\bar{\theta}_i \neq WR$, at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$, it's easy to show that $m_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$ is optimal.

For $\theta_i = BN$ the choice of message is inconsequential, therefore optimal. For $\theta_i = WN$ the choice of message has some effects, but only on third parties. That is, it affects black workers and racist white workers but non-racist white workers are not affected at all. Therefore the message is optimal as well.

For player i, of type $\bar{\theta}_i = WR$, at information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$, it is straightforward to show the optimality of its message $m_i(\bar{\theta}_i, \bar{\boldsymbol{v}})$.

First, notice that—relative to the politically correct message—sending the politically incorrect message has no cost—with certainty. On the other hand though, the expected benefits—relative to sending the politically correct message—are positive since for some states that occur with positive probability this will decrease blacks' average utility even further (remember racists attack depending on how great the number of white players sending the politically incorrect message is).

B.3 Proof of the Implications from the Equilibrium

So far we have proposed an equilibrium and shown that it is indeed an equilibrium, but we haven't shown that the cases under which sabotage takes place actually happen with positive probability.

Given equilibrium B.1, we want to show that the expected wage of a black type is strictly less than that of any white type (racist or non-racist). Recall that the wage of player i is given by

$$w_i = \left[1 - \delta \sum_{j \in \mathcal{N}} I(a_j = i)\right] ,$$

and the expected wage of type θ is given by

$$E[w_\theta] = E_{\boldsymbol{\omega}} \left[\frac{1}{\#\mathcal{N}_\theta} \sum_{\theta_i = \theta} w_i\right] = \int_{\boldsymbol{\Omega}} \left[\frac{1}{\#\mathcal{N}_\theta} \sum_{\theta_i = \theta} w_i\right] dP(\tilde{\boldsymbol{\omega}}) \ .$$

Under the strategies of equilibrium B.1, the wage w_i for any player i of type $\theta_i \neq BN$ is equal to one, in *any* state:

$$\theta_i \neq BN \Rightarrow w_i = 1 \quad \forall \boldsymbol{\omega} \in \boldsymbol{\Omega} \, .$$

Therefore the expected wage for type $\theta \neq BN$ equals one as well:

$$E[w_\theta] = \int_{\boldsymbol{\Omega}} \left[\frac{1}{\#\mathcal{N}_\theta} \sum_{\theta_i = \theta} 1\right] dP(\tilde{\boldsymbol{\omega}}) = 1, \text{ for } \theta \neq BN \, .$$

For player i of type $\theta_i = BN$ though, we have that in states where $n_{WR} > 0$ and $\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}$, the wage is less than one:

$$\theta_i = BN \text{ and } n_{WR} > 0 \text{ and } \frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n} \implies w_i < 1 \, .$$

Otherwise the wage is trivially equal to one. Hence, all we need to show is that the set of states with $n_{WR} > 0$ and $\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}$ (and $n_{BN} > 0$ of course) has a positive probability. Let $\boldsymbol{\Omega}_a$ be the set of such states, clearly $\boldsymbol{\Omega}_a \subseteq \boldsymbol{\Omega}$. It is therefore sufficient if we show that a non-empty subset of $\boldsymbol{\Omega}_a$ has positive probability.

Consider the states where exactly one player is black, and the rest are racist whites. Since $\mathbf{F}$ has full support and n is finite, these states have a positive probability: $F_{BN} F_{WR}^{n-1} > 0$. Moreover, these states are indeed in $\boldsymbol{\Omega}_a$ because $n_{BN} = 1 > 0$, $n_{WR} = n - n_{BN} = n - 1 > 0$ since $n > 2$, and $\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}$ since

$$\begin{aligned}
&\beta\delta \geq \frac{1}{n} \text{ by assumption of the proposition,} \\
\Leftrightarrow &\frac{\beta\delta}{n_{BN}} \geq \frac{n_{BN}}{n} \text{ since these states have } n_{BN} = 1, \\
\Leftrightarrow &\frac{\beta\delta}{n_{BN}} \geq \frac{n - n_{WR}}{n} \text{ since these states have } n = n_{BN} + n_{WR},
\end{aligned}$$

therefore

$$\frac{\beta\delta}{n_{BN}} \geq 1 - \frac{n_{WR}}{n}, \text{ which proves that these states are in } \boldsymbol{\Omega}_a.$$

Therefore $\boldsymbol{\Omega}_a$ has positive probability and the expected wage for type $\theta = BN$ is less then one, i.e., the expected wage of blacks is less than that of any white type.

B.4 Lemmas employed in the Proof of the Proposition

Lemma 1. *If $I_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ is reached with positive probability under equilibrium B.1, then*

$$\pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \begin{cases} 1, & \textit{if } \boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{v}}) = \bar{\boldsymbol{m}}; \\ 0, & \textit{otherwise.} \end{cases}$$

Proof. Under equilibrium B.1 the beliefs—which are presented in B.1.2—are given by

$$\pi'_i(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \pi'(\tilde{\boldsymbol{\theta}} \mid \bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) = \begin{cases} 1, & \text{if } m(\tilde{\theta}_j, \bar{\boldsymbol{v}}) = \bar{m}_j \text{ for all } j \notin \mathcal{N}_{BN}; \\ 0, & \text{otherwise.} \end{cases}$$

That is, there is only one node in $I_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ with a positive probability. Hence, all we need to show is that this node must comply with "$m(\tilde{\theta}_j, \bar{\boldsymbol{v}}) = \bar{m}_j$ for all $j \in \mathcal{N}_{BN}$" as well. But this must be true since the information set is reached with positive probability, i.e., all the observed messages are sent according to the equilibrium strategy: $m_j = m(\tilde{\theta}_j, \bar{\boldsymbol{v}}) \quad \forall j$.

If $I_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ is reached with positive probability it must therefore be that $\bar{m}_j = m(\tilde{\theta}_j, \bar{\boldsymbol{v}}) \quad \forall j \in \mathcal{N}$ for the only node that is reached with positive probability in $I_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$. □

Lemma 2. *Let $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ be an element of $\boldsymbol{\Theta}$ that satisfies*

1. $\boldsymbol{V}(\boldsymbol{\vartheta}) = \bar{\boldsymbol{v}}$, *and*
2. $m_j(\vartheta_j, \bar{\boldsymbol{v}}) = \bar{m}_j$ *for all* $j \in \mathcal{N}_{\mathcal{W}}$.

Then, $\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ exists and is unique for any $(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \in \{0,1\}^n \times \{M_1, M_2\}^n$.

Proof. We can always construct such a $\boldsymbol{\vartheta}$, and it can only be done one way:

First, to comply with $\boldsymbol{V}(\boldsymbol{\vartheta}) = \bar{\boldsymbol{v}}$, we have to set $\vartheta_j = BN$ whenever $\bar{v}_j = 1$. Second, to comply with $m_j(\vartheta_j, \bar{\boldsymbol{v}}) = \bar{m}_j$ for all $j \in \mathcal{N}_{\mathcal{W}}$, we have to set

$$\vartheta_j = \begin{cases} WR, & \text{if } \bar{v}_j = 0 \text{ and } m_j = M_1; \\ WN, & \text{if } \bar{v}_j = 0 \text{ and } m_j = M_2. \end{cases}$$

This assigns a unique type to every $j \in \mathcal{N}$. Therefore $\boldsymbol{\vartheta}$ exists and is unique. □

Lemma 3. *Let* $\tilde{\boldsymbol{\theta}} \in \mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$*, then* $\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) = \bar{\boldsymbol{v}}$ *and*

$$\boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \boldsymbol{V}(\tilde{\boldsymbol{\theta}})) = \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}}) \Longrightarrow \tilde{\boldsymbol{\theta}} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})\,.$$

Proof. $\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) = \bar{\boldsymbol{v}}$ follows immediately from the definition of the information set $\mathcal{I}_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$ on page 19.

$\boldsymbol{V}(\tilde{\boldsymbol{\theta}}) = \bar{\boldsymbol{v}}$ implies $\vartheta_j = \tilde{\theta}_j$ for all j with $\bar{v}_j = 1$. That is, for those js both components are *BN*.

On the other hand, $\boldsymbol{m}(\tilde{\boldsymbol{\theta}}, \bar{\boldsymbol{v}}) = \boldsymbol{m}(\boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}), \bar{\boldsymbol{v}})$ implies that for every j with $\bar{v}_j = 0$:

$$\tilde{\theta}_j = \vartheta_j\,.$$

That is, under function $\boldsymbol{m}(\cdot, \cdot)$, the type of all the white players is fully determined by the message they send. Therefore $\tilde{\theta}_j = \vartheta_j$ for all j, and so $\tilde{\boldsymbol{\theta}} = \boldsymbol{\vartheta}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$. □

C Proof of Claim 1

If $n_{BN} = 0$ then, by necessity, there is no discrimination. So let us focus on the cases where $n_{BN} > \sqrt{\beta\delta n}$.

We have

$$\begin{aligned} & n_{BN} > \sqrt{\beta\delta n}\,, \\ \Rightarrow\ & n_{BN}^2 > \beta\delta n \quad \text{since } \sqrt{\beta\delta n} > 0, \\ \Leftrightarrow\ & \frac{n_{BN}}{n} > \frac{\beta\delta}{n_{BN}}\,, \end{aligned}$$

therefore

$$\frac{n_{BN} + n_{WN}}{n} \geq \frac{n_{BN}}{n} > \frac{\beta\delta}{n_{BN}}\,.$$

The expression on the far left is the expected cost—relative to the peaceful action—of sabotaging a black player after the cheap-talk stage, *regardless* of the equilibrium. The expression on the far right is the expected benefit—relative to the peaceful action—of sabotaging a black player after the cheap-talk stage , *regardless* of the equilibrium. Therefore in any (weak perfect bayesian) equilibrium the black players will not be sabotaged, since this would go against sequential rationality.

White players are not sabotaged either, since the expected benefit is simply replaced by zero *regardless* of the equilibrium.

D Proof of the Theorem

First, in subsection D.1, we show that—for any equilibrium—there is no discrimination in the states of the world with $n_{\mathcal{W}} = 1$. Therefore we have that discrimination can only occur in states of the world with $n_{\mathcal{W}} \geq 2$.

Notice that the set of all type-symmetric weak perfect bayesian equilibria can be partitioned into the set of white-message-separating equilibria (those equilibria where racist whites send a different message from non-racist whites *in at least one* state of the world with $n_{\mathcal{W}} \geq 2$) and the set of white-message-pooling equilibria (those equilibria where both racist and non-racist whites send the same message *in every* state of the world with $n_{\mathcal{W}} \geq 2$).

In subsection D.2, we show that there is no equilibrium, in the white-message-pooling class, where the aggressive action is chosen under "LER" (low enough racism).

Finally, in subsection D.3, we show that under "HES (high enough sanction) and LER" the white-message-separating class is empty.

D.1 Discrimination can only occur in states of the world with $n_{\mathcal{W}} \geq 2$

If $n_{\mathcal{W}} = 0$ then, by necessity, there is no discrimination. So consider the remaining case where $n_{\mathcal{W}} = 1$. Clearly, in this case the only white player does not face any uncertainty with regard to $\boldsymbol{\theta}$ (he knows his type and that of the others: *BN*).

Relative to the peaceful action, the expected benefit that a racist player gets from sabotaging any black player is $\frac{\beta\delta}{n-1}$, while the expected cost he faces is $\frac{n-1}{n}$. Therefore sequential rationality requires the choice of the aggressive action if and only if

$$\frac{\beta\delta}{n-1} \geq \frac{n-1}{n},$$
$$\Leftrightarrow \beta\delta \geq \frac{(n-1)^2}{n}$$

but $\frac{(n-1)^2}{n} \geq \frac{4}{3}$ since $n \geq 3$ by assumption. Therefore sequential rationality requires the choice of the aggressive action if and only if

$$\beta\delta \geq \frac{4}{3},$$

but this is impossible since $0 \leq \beta, \delta \leq 1$.

Corollary. *Under $n = 3$ and a fully white-message-separating equilibrium: If state $\tilde{\omega}$ has $n_{BN} = n_{WN} = n_{WR} = 1$, then no sabotage occurs.*

D.2 The class of white-message-pooling equilibria has no discrimination under "LER"

Let $\theta_i = WR$ and $n_{\mathcal{W}} \geq 2$. Under white-message-pooling, information set $\mathcal{I}_i(\theta_i, \overline{\boldsymbol{v}}, \overline{\boldsymbol{m}})$ is equal to information set $\mathcal{I}_i(\theta_i, \overline{\boldsymbol{v}})$: Player i has the same information because racist whites and non-racist whites are indistinguishable from their messages.

Let u_i^a be the payoff that i gets from sabotaging a black player, and let u^p be the payoff that i gets from being peaceful. Sequential rationality implies that i will choose the aggressive action if and only if

$$\begin{aligned} & E[u_i^a - u_i^p \mid \mathcal{I}_i(WR, \overline{\boldsymbol{v}})] \geq 0 \\ \Leftrightarrow E\left[\frac{\beta\delta}{n_{BN}} - b_i\left[1 - \delta\sum_{j\in\mathcal{N}} I(a_j = i)\right]\middle| \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right] & \geq 0\,; \end{aligned}$$

notice that sequential rationality also implies that no white player is sabotaged, since the expected benefit is zero and the expected cost is positive given $n \geq 3$, therefore

$$E\left[\frac{\beta\delta}{n_{BN}} - b_i\middle| \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right] \geq 0\,;$$

the only uncertainty remaining is that due to b_i, so

$$\begin{aligned} \frac{\beta\delta}{n_{BN}} - E\left[b_i \mid \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right] & \geq 0\,, \\ \Leftrightarrow \frac{\beta\delta}{n_{BN}} & \geq E\left[b_i \mid \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right]\,; \end{aligned}$$

by the law of iterated expectations we have

$$\frac{\beta\delta}{n_{BN}} \geq E\left[E[b_i \mid n_{WR}] \mid \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right]\,;$$

the expectation of a Bernoulli r.v. is simply its probability, so we have

$$\begin{aligned} \frac{\beta\delta}{n_{BN}} & \geq E\left[1 - \frac{n_{WR}}{n}\middle| \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right]\,, \\ \Leftrightarrow \frac{\beta\delta}{n_{BN}} & \geq 1 - \frac{1}{n}E\left[n_{WR} \mid \mathcal{I}_i(WR, \overline{\boldsymbol{v}})\right]\,; \end{aligned}$$

the relevant information that player i possesses at information set $\mathcal{I}_i(WR, \bar{\boldsymbol{v}})$ is $n_{\mathcal{W}}$ itself and $n_{WR} \geq 1$, so we may express the previous inequality as

$$\begin{aligned} \frac{\beta\delta}{n_{BN}} &\geq 1 - \frac{1}{n} E\left[n_{WR} \mid n_{\mathcal{W}} \text{ and } n_{WR} \geq 1\right], \\ \Leftrightarrow \frac{\beta\delta}{n_{BN}} &\geq 1 - \frac{1}{n}\left[1 + (n_{\mathcal{W}} - 1)\frac{F_{WR}}{F_{\mathcal{W}}}\right]. \end{aligned}$$

It turns out that this inequality contradicts the assumed condition "LER":

$$\begin{aligned} &\frac{F_{WR}}{F_{\mathcal{W}}} + \beta\delta \leq \frac{1}{2}, \\ &\quad \Leftrightarrow \beta\delta \leq 1 - \frac{1}{2} - \frac{F_{WR}}{F_{\mathcal{W}}}, \\ &\quad \Rightarrow \beta\delta < 1 - \frac{1}{3} - \frac{F_{WR}}{F_{\mathcal{W}}}, \\ &\quad \Rightarrow \beta\delta < 1 - \frac{1}{n} - \frac{F_{WR}}{F_{\mathcal{W}}} \text{ since throughout the paper we have assumed } n \geq 3, \\ &\quad \Leftrightarrow \beta\delta < \left(1 - \frac{F_{WR}}{F_{\mathcal{W}}}\right) - \frac{1}{n}, \\ &\quad \Rightarrow \beta\delta < \left(1 - \frac{F_{WR}}{F_{\mathcal{W}}}\right) + \frac{1}{n}\left(2\frac{F_{WR}}{F_{\mathcal{W}}} - 1\right) \text{ since } \frac{F_{WR}}{F_{\mathcal{W}}} > 0 \text{ by assumption}, \\ &\quad \Leftrightarrow \beta\delta < 1 - \frac{1}{n}\left[1 + (n-2)\frac{F_{WR}}{F_{\mathcal{W}}}\right], \\ &\Rightarrow \frac{\beta\delta}{n_{BN}} < 1 - \frac{1}{n}\left[1 + (n_{\mathcal{W}} - 1)\frac{F_{WR}}{F_{\mathcal{W}}}\right] \text{ since discrimination requires } n - 1 \geq n - n_{BN} = n_{\mathcal{W}}. \end{aligned}$$

Therefore sequential rationality and condition "LER" assure that aggressive actions are never chosen in white-message-pooling equilibria.

D.3 The class of white-message-separating equilibria is empty under "HES and LER"

Suppose there exists a white-message-separating equilibrium. This means that therein, for some $\boldsymbol{v}^* \in \{0,1\}^n$ with $n_{\mathcal{W}} \geq 2$, racist whites send a different message than non-racist whites.

To be an equilibrium, sequential rationality and weak consistency of beliefs have to be satisfied. Therefore the sub-strategy of player i, $a_i(\bar{\theta}_i, \bar{\boldsymbol{v}}, \bar{\boldsymbol{m}})$, has to satisfy

$$a_i(\bar{\theta}_i, \boldsymbol{v}^*, \bar{\boldsymbol{m}}) = a(\bar{\theta}_i, \boldsymbol{v}^*, \bar{\boldsymbol{m}}) \in \begin{cases} \mathcal{N}_{BN} & \text{if } \bar{\theta}_i = WR \text{ and } n_{BN} > 0 \text{ and } \widetilde{F_{WR}} \geq 1 - \frac{\beta\delta}{n_{BN}}; \\ \{p\} & \text{otherwise.} \end{cases}$$

That racists attack blacks, if and only if $n_{BN} > 0$ and $\widetilde{F_{WR}} \geq 1 - \frac{\beta\delta}{n_{BN}}$, should not be a surprise since a similar result was shown in the proposition.

What might have to be explained is that nobody attacks whites. This is because for *any* equilibrium and realized state we have that, relative to the peaceful action, the expected benefit from attacking a white player is zero, but the expected cost is strictly positive. And this is true for everyone.

Now, there are two ways for having a white-message-separation: either racists or non-racists send the politically incorrect message (M_1), but not both. However, the white-message separation with non-racists sending the politically incorrect message is not sequentially rational *for any* $c > 0$: Relative to the politically correct message, M_2, non-racists get zero expected benefit and a strictly positive expected cost for sending the politically incorrect message, M_1.

So for $c > 0$, the only white-message-separating equilibria that may exist are those where racists send the politically incorrect message, M_1.

Therefore, in particular, our hypothetical equilibrium must satisfy the following sequential rationality condition for $\theta_i = WR$:

$$E[u^{pi} - u^{pc} \mid \mathcal{I}_i(\bar{\theta}_i, \boldsymbol{v}^*), \boldsymbol{m}_{-i}, \boldsymbol{a}_{-i}] \geq 0\,,$$

where u^{pi} is the payoff from sending the politically incorrect message and attacking like the other racists do, u_{pc} is the payoff from an alternative strategy to be presented below, and the expectation is conditional on the information set $\mathcal{I}_i(\bar{\theta}_i, \boldsymbol{v}^*)$ and takes everyone else's strategies into account (here represented by $\boldsymbol{m}_{-i}, \boldsymbol{a}_{-i}$).

Let i's alternative strategy, or pc strategy, satisfy

$$\begin{aligned} m_i^{pc}(WR, \boldsymbol{v}^*) &= M_2 \\ a_i^{pc}(WR, \boldsymbol{v}^*, \boldsymbol{m}_{-i}, m_i^{pc}) &\in \begin{cases} \mathcal{N}^{BN} & \text{if } \widetilde{F_{WR}}'(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}}; \\ \{p\} & \text{otherwise,} \end{cases} \end{aligned}$$

where $\widetilde{F_{WR}}'(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \equiv \frac{\#\{j \in \mathcal{N} \mid \bar{m}_j = M_1 \text{ and } \bar{v}_j = 0\} + 1}{n}$.

We therefore have

$$E[u^{pi} - u^{pc} \mid \mathcal{I}_i(\bar{\theta}_i, \boldsymbol{v}^*), \boldsymbol{m}_{-i}, \boldsymbol{a}_{-i}] \geq 0\,,$$

(we drop the expectation's conditionals in what follows, for the sake of space)

$$\begin{aligned}
&\Leftrightarrow E\left[-c\frac{n_{WN}}{n_{WR}} + (n_{WR}-1)\frac{\beta\delta}{n_{BN}} \cdot I\left(\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}} > \widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}')\right)\right] \geq 0\,,\\
&\qquad\Leftrightarrow E\left[(n_{WR}-1)\frac{\beta\delta}{n_{BN}} \cdot I\left(\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}} > \widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}')\right)\right] \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,\\
&\qquad\Leftrightarrow \Pr\left[\widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}) \geq 1 - \frac{\beta\delta}{n_{BN}} > \widetilde{F_{WR}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{m}}')\right] E\left[(n_{WR}-1)\frac{\beta\delta}{n_{BN}}\right] \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,
\end{aligned}$$

where the expectation on the left side is conditional on the statement of the probability,

$$\Leftrightarrow \Pr\left[n_{WR} \geq n\left(1 - \frac{\beta\delta}{n_{BN}}\right) > n_{WR} - 1\right] E\left[(n_{WR}-1)\frac{\beta\delta}{n_{BN}}\right] \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,$$

(in what follows we use the so-called ceiling function; this function maps any real number x to the smallest integer that is weakly greater than x; it is written as $\lceil x \rceil$)

$$\Leftrightarrow \Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right] E\left[(n_{WR}-1)\frac{\beta\delta}{n_{BN}}\right] \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,$$

since the expectation on the left side is conditional on the statement of the probability, we have

$$\begin{aligned}
&\Leftrightarrow \Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right] E\left[(n_{WR}-1)\frac{\beta\delta}{n_{BN}}\middle| n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right] \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,\\
&\qquad\Leftrightarrow \Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]\left(\left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil - 1\right)\frac{\beta\delta}{n_{BN}} \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,\\
&\qquad\Leftrightarrow \Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]\left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1\right\rceil\frac{\beta\delta}{n_{BN}} \geq E\left[c\frac{n_{WN}}{n_{WR}}\right]\,,
\end{aligned}$$

We pause here to notice that the probabilities and expectations are conditional on $n_{WR} \geq 1$ since i knows his type, thus having

$$\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil \middle| n_{WR} \geq 1\right] = \frac{\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]}{\Pr[n_{WR} \neq 0]}\,.$$

And also having

$$\begin{aligned} E\left[c\frac{n_{WN}}{n_{WR}} \mid n_{WR} \geq 1\right] &= c \cdot E\left[\frac{n_{\mathcal{W}} - n_{WR}}{n_{WR}}\middle| n_{WR} \geq 1\right] \\ &= c \cdot \sum_{h=1}^{n_{\mathcal{W}}} \Pr[n_{WR} = h \mid n_{WR} \geq 1] \cdot \frac{n_{\mathcal{W}} - h}{h} \\ &= c \cdot \frac{\sum_{h=1}^{n_{\mathcal{W}}} \Pr[n_{WR} = h] \cdot \frac{n_{\mathcal{W}} - h}{h}}{\Pr[n_{WR} \neq 0]} \end{aligned}$$

Therefore the inequality becomes

$$\begin{aligned} &\frac{\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]}{\Pr[n_{WR} \neq 0]} \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1\right\rceil \frac{\beta\delta}{n_{BN}} \geq c \cdot \frac{\sum_{h=1}^{n_{\mathcal{W}}} \Pr[n_{WR} = h] \cdot \frac{n_{\mathcal{W}} - h}{h}}{\Pr[n_{WR} \neq 0]}, \\ \Leftrightarrow\ &\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right] \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1\right\rceil \frac{\beta\delta}{n_{BN}} \geq c \cdot \sum_{h=1}^{n_{\mathcal{W}}} \Pr[n_{WR} = h] \cdot \frac{n_{\mathcal{W}} - h}{h}, \end{aligned}$$

which is equivalent to

$$\left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1\right\rceil \frac{\beta\delta}{n_{BN}} \geq c \cdot \sum_{h=1}^{n_{\mathcal{W}}} \frac{\Pr[n_{WR} = h]}{\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]} \cdot \frac{n_{\mathcal{W}} - h}{h}.$$

It can be shown that due to the binomial nature of n_{WR},

$$\frac{\Pr[n_{WR} = h]}{\Pr\left[n_{WR} = \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil\right]} = \left(\frac{F_{WR}}{F_{WN}}\right)^{h - \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil}.$$

Therefore the inequality becomes

$$\left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1\right\rceil \frac{\beta\delta}{n_{BN}} \geq c \cdot \sum_{h=1}^{n_{\mathcal{W}}} \left(\frac{F_{WR}}{F_{WN}}\right)^{h - \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right)\right\rceil} \cdot \frac{n_{\mathcal{W}} - h}{h}.$$

But this inequality contradicts "HES and LER" since

$$
\begin{aligned}
& c > n \cdot \beta\delta \left(\frac{F_{WR}}{F_{WN}}\right) \text{ by "HES",} \\
\Rightarrow\ & c > (n-1)\beta\delta \left(\frac{F_{WR}}{F_{WN}}\right) \text{ since } \beta\delta \left(\frac{F_{WR}}{F_{WN}}\right) \text{ is positive,} \\
\Rightarrow\ & c > \lceil n-1 \rceil \beta\delta \left(\frac{F_{WR}}{F_{WN}}\right)^{\lceil \frac{n}{2}-1 \rceil} \text{ since } \frac{F_{WR}}{F_{WN}} \leq 1 \text{ by "LER", and } n \geq 3, \\
\Rightarrow\ & c > \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \beta\delta \left(\frac{F_{WR}}{F_{WN}}\right)^{\lceil (1-\beta\delta)n-1 \rceil} \text{ since } \beta\delta \leq 1/2 \text{ by "LER", and } n_{BN} \geq 1, \\
\Rightarrow\ & c > \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \frac{\beta\delta}{n_{BN}} \left(\frac{F_{WR}}{F_{WN}}\right)^{\lceil (1-\beta\delta/n_{BN})n-1 \rceil} \text{ by the reasons above,} \\
\Leftrightarrow\ & c > \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \frac{\beta\delta}{n_{BN}} \left(\frac{F_{WR}}{F_{WN}}\right)^{\lceil (1-\beta\delta/n_{BN})n \rceil - 1};
\end{aligned}
$$

this is the same as

$$
\begin{aligned}
c \left(\frac{F_{WR}}{F_{WN}}\right)^{1-\lceil (1-\beta\delta/n_{BN})n \rceil} &> \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \frac{\beta\delta}{n_{BN}}, \\
\Leftrightarrow c \sum_{h=1}^{2} \left(\frac{F_{WR}}{F_{WN}}\right)^{h-\lceil (1-\beta\delta/n_{BN})n \rceil} \frac{2-h}{h} &> \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \frac{\beta\delta}{n_{BN}},
\end{aligned}
$$

$n_{\mathcal{W}} > 2$ and the summation is increasing in the strictly positive summands, so

$$
\Rightarrow c \sum_{h=1}^{n_{\mathcal{W}}} \left(\frac{F_{WR}}{F_{WN}}\right)^{h-\lceil (1-\beta\delta/n_{BN})n \rceil} \frac{n_{\mathcal{W}}-h}{h} > \left\lceil \left(1 - \frac{\beta\delta}{n_{BN}}\right) n - 1 \right\rceil \frac{\beta\delta}{n_{BN}},
$$

or, rearranging terms,

$$
\Leftrightarrow c \cdot \sum_{h=1}^{n_{\mathcal{W}}} \left(\frac{F_{WR}}{F_{WN}}\right)^{h-\left\lceil n\left(1-\frac{\beta\delta}{n_{BN}}\right) \right\rceil} \cdot \frac{n_{\mathcal{W}}-h}{h} > \left\lceil n\left(1 - \frac{\beta\delta}{n_{BN}}\right) - 1 \right\rceil \frac{\beta\delta}{n_{BN}}.
$$

E Proof of Claim 2

To prove the claim, we derive a sufficient condition for PC to be harmful and show that this condition is met by one particular choice of parameters. However, the reader should

be aware that the sufficient condition herein derived is far from necessary, and that a much greater set of parameters are likely to meet the weaker sufficient conditions.

From subsection D.1 we know that, for there to be discrimination in some state, there have to be at least two white workers, i.e., $n_{\mathcal{W}} \geq 2$. And, by necessity, there has to be at least one black worker, i.e., $n_{BN} \geq 1$.

Now, from subsection D.2 we know that, under pooling in messages, racist white player i chooses to attack iff

$$\begin{aligned} & \frac{\beta\delta}{n_{BN}} \geq 1 - \frac{1}{n}\left[1 + (n_{\mathcal{W}} - 1)\frac{F_{WR}}{F_{\mathcal{W}}}\right], \\ \Leftrightarrow & \frac{F_{WR}}{F_{\mathcal{W}}} \geq \frac{n(1 - \beta\delta/n_{BN}) - 1}{n_{\mathcal{W}} - 1}. \end{aligned}$$

Clearly, if this last condition was met for every possibly-discriminatory state (every state with $n_{\mathcal{W}} \geq 2$ and $n_{BN} \geq 1$), then a message-pooling equilibrium—like the one effected by a Political Correctness regime with high enough social sanction—would have discrimination in every possibly-discriminatory state, as opposed to only in those states where there are enough white racist players. Because the type distribution has full support, we know that there are indeed states where the white racists are not enough, and that these states have a positive probability.

Hence, all we need to do is find the value that $\frac{F_{WR}}{F_{\mathcal{W}}}$ would have to be to in order to be weakly greater than $\frac{n(1-\beta\delta/n_{BN})-1}{n_{\mathcal{W}}-1}$ for every state such that $n_{\mathcal{W}} \geq 2$ and $n_{BN} \geq 1$.

It is straight forward to show that the expression $\frac{n(1-\beta\delta/n_{BN})-1}{n_{\mathcal{W}}-1}$ is maximized for $n_{BN} = n - 2$ and $n_{\mathcal{W}} = 2$. The resulting—sufficient—condition is

$$\frac{F_{WR}}{F_{\mathcal{W}}} \geq (n-1) - \frac{n}{n-2}\beta\delta.$$

It is easy to see that this condition is satisfied—with $n = 3$—for many parameters $\boldsymbol{F}, \beta, \delta$. The condition is extremely strict though and, as the reader may appreciate, is useless for $n \geq 4$. Again, though, we would like to emphasize that a much greater set of parameters are likely to meet the weaker sufficient conditions.