Political Correctness and Elite Prestige

BSE Working Paper 1375| December 2022

Esther Hauk, Javier Ortega

bse.eu/research

Political Correctness and Elite Prestige*

Esther Hauk† Javier Ortega‡

December 2022

Abstract

Consider a society where the prestige of orthodox views is linked to the prestige of the elite. Heterodox individuals are less likely to express their views if other peers refrain from doing so and if the elite is prestigious. In turn, corruption by the elite is less easily detected if orthodox views dominate. We characterize equilibrium self-denial and corruption and show that an exogenous increase in the range of orthodox views may result in a decrease in the total number of individuals truthfully expressing their views. Some features of the model are shown to be compatible with U.S. data.

JEL: C72, D7, Z1, Z13

Keywords: political correctness, Overton window, social pressure, conformity, preference falsification.

*A previous version of the paper circulated under the title "Equilibrium Political Correctness". We would like to thank Jonathan Hopkin, Julie Mallet, Simon Susen and seminar participants at Kingston University and at the African Meeting of the Econometric Society (Rabat) for helpful comments and discussions, as well as James Hunter and Carl Bowman from the UVA Institute for Advanced Studies for sharing the data on their 2016 Survey of American Political Culture and the Gallup Organization for fielding the study. Ortega thanks the Department of Government at the LSE for their hospitality while revising this paper. Hauk acknowledges financial support from the Spanish Agencia Estatal de Investigacion, through the Severo Ochoa Programme for Centers of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S); from the Spanish Ministry Science, Innovation and Universities through project PGC2018-097898-B-100 MCIU/AEI/FEDER, EU; from the Government of Catalonia under project 2017 SGR 1571 AGAUR Generalitat de Catalunya and from the Spanish Ministry of Science and Innovation through research project PID2021-126209OB-I00 funded by MCIN-AEI/10.13039/501100011033 and by ERDF A way of making Europe.

†Institut d'Anàlisi Econòmica (IAE-CSIC), Move and Barcelona School of Economics, Campus UAB, Bellaterra (Barcelona); email: esther.hauk@iae.csic.es.

‡Department of Economics, Kingston University. j.ortega@kingston.ac.uk

1 Introduction

Individuals belonging to a group might be reluctant to express their own views if they perceive them to be uncommon or heterodox within that group: uttering them might damage their reputation or even question their commitment to the group. In some cases, this type of (perceived) peer pressure might then result in individuals expressing views which are not actually theirs, but which will be well-received within the group –see Loury (1994), who refers to this as "political correctness", and also Bernheim (1994) and Michaeli and Spiro (2015).

At the same time, self-censoring in the expression of a heterodox viewpoint may also be linked to the prestige of orthodox views, itself likely to be related to the prestige or power of the social group thought to best represent these orthodox views (see Bourdieu and Boltanski, 1976, and Susen, 2013). One would thus expect infringements of the orthodoxy to be more likely the weaker the power of this dominant group. For instance, liberation of speech from the monopoly of experts/economists and criticism of "excessive" political correctness were important themes put forward by respectively defendants of Brexit and supporters of Trump in 2016,[1] and these political movements rose after the prestige of elites was damaged by the Great Recession (Hopkin, 2020).

This paper thus develops a theory of equilibrium political correctness in which the prevalence of orthodox views can in principle depend on the prestige/power of the dominant group. To assess the relevance of this novel mechanism, we start by presenting a more standard political correctness model (which we refer to as the simple model) in which heterodox individuals decide whether or not to express their views purely on the basis of what other heterodox individuals do.

In the simple model, heterodox individuals are heterogeneous in the extent to which their views differ from the orthodoxy, and each of them decides whether to truthfully express her views or instead be politically correct by expressing the closest orthodox view. Truthfully expressing her views generates utility to the individual but less so the larger the number of individuals who instead choose to be politically correct, especially if peer pressure is high. In turn, being politically correct entails for the individual a cost increasing in the distance between her views and the orthodoxy.

When peer pressure is low, we show that nobody chooses self-denial at equilibrium, as being truthful generates a utility and any self-denial by other individuals would be heavily discounted due to the low peer pressure. Instead, when peer pressure becomes larger, the interaction among individuals becomes relevant, and the resulting bandwagon effect generates multiple equilibria: while no self-denial remains the best option for everybody if nobody else self-denials, full self-denial now becomes an equilibrium as self-denial by sufficiently many peers renders the expression of heterodox views more costly than self-denial.

In the main model, we introduce an "elite" (a dominant group) which best represents orthodox views, and for this reason we link the prestige of these views with the prestige of the elite. Specifically, we assume that a high level of self-denial among heterodox

[1] On the Brexit campaign, see for instance Clarke and Newman (2017). Ahead of the 2016 U.S. election, the 2016 American Survey of Political Culture showed that 93% of Trump supporters (resp. 53% of Clinton's) agree that political correctness is a serious problem making it hard for people to say what they really think (see Hunter and Bowman, 2016). See also for instance Conway, Repkea and Houck (2017).

individuals renders orthodox views more prestigious, which in turn enhances the prestige of the elites and makes it easier for them to engage in rent-seeking activities (corruption) without being detected. At the same time, we assume that if elite members choose not to engage in corruption, their prestige is higher, and this translates into a higher prestige of orthodox views, and thus raises the value of self-denial for heterodox individuals.

We first show that, whenever the good behavior of the elite sufficiently raises the value of self-denial, self-denial can arise as an equilibrium outcome even for low levels of peer-pressure. The intuition is as follows: in the simple model, in the presence of low peer-pressure, there is no strong enough bandwagon effect reinforcing incentives for self-denial, and for this reason no self-denial is the unique equilibrium; instead, if heterodox individuals care about the good behavior of the elite and the payoff from corruption is not too high, good behavior generates individual incentives for self-denial, which are then reinforced through the bandwagon effect even for lower peer-pressure levels.[2]

At the same time, we show that if the masses care sufficiently about the elite's behavior and peer pressure is sufficiently high, the presence of the elite results in full self-denial becoming the unique equilibrium in situations where there was multiplicity in the simple model. Interestingly, full self-denial always holds in this case no matter the elite's payoff from corruption, and the corruption payoff only affects the observed level of corruption characterizing the equilibrium. Specifically, if corruption pays little, full corruption is never a best reply for the elite, but at the same time the prestige of orthodox ideas stemming from self-denial lowers the corruption detection probability and generates some corruption. In turn, if corruption pays a lot, the same mechanism leads to a situation in which full adhesion of the masses to the ideas of the elite results in full corruption by the elite.

We also examine how an exogenous increase in the range of socially acceptable views affects the proportion of agents expressing their views in equilibrium. Naturally, those individuals whose views have (exogenously) become orthodox can now freely express them, so initially less individuals self-deny. Consider however the situation of an individual holding a "very heterodox" view: this individual was initially choosing to express her views because self-denial was too costly, as her own view was too different from the less distant orthodox view. However, when more views become socially acceptable, her views are not anymore so heterodox, she then may choose to self-denial, which in turn lowers the number of individuals expressing their views. In the simple model, we show that this second effect is always dominated and that as result a wider set of orthodox views always leads to overall less self-denial. This does not necessarily hold in the model with elite, as the initial fall in the proportion of self-deniers boosts the corruption detection probability, which improves the behavior of the elite, and further reinforces self-denial. When the masses care a lot about elite behavior, we can then have that an increase in the range of socially acceptable views ends up lowering the number of individuals expressing their views.

We empirically illustrate some of the features of the model using the 2016 Survey of American Political Culture (Hunter and Bowman, 2016) conducted in August 2016 two months ahead of Donald Trump's election. Consistently with the model, individuals who

[2]In particular, if the masses care to some extent about the behavior of the elites, multiple equilibria characterised by a non-positive correlation between self-denial and corruption are possible even if no self-denial was the only equilibrium in the simple model.

view themselves as more distant from society are more likely to believe there is a political correctness problem in the U.S. In addition, individuals considering that the elites are self-interested are also more likely to believe political correctness to be a problem. This is also in accordance with our model.

Our paper relates to a growing literature on political correctness. Loury (1994) shows how in the presence of norms on acceptable views within a group, the interaction between sender and receiver becomes strategic and can generate a link between ways of expression and bad qualities of senders, resulting in the avoidance of such ways of expression. Bernheim (1994) shows that if intrinsic preferences are not observable and individuals care sufficiently about status, heterogeneous individuals may all choose the same action in order to avoid being characterized as somebody with uncommon intrinsic preferences. Morris (2001) differs from the above papers and ours in that there is a true state of the world, with the analysis focusing on the extent to which an advisor will be able or not to convey this state of the world to the policy maker when the advisor cares about his own reputation in the eyes of the policy maker. The general message of our simple model is similar to that of Michaeli and Spiro (2015) except that in the latter individuals can choose the extent to which they self-deny, implying that conformity to the social norm does not only depend on the number of self-deniers.[3] Overall, the contribution of our paper to the literature is to link the prestige of orthodox ideas and the behavior of the elites representing them, which can be related to Bourdieu (1979)'s notion of symbolic power arguing that while debates of views tend to be presented as belonging to an autonomous asocial sphere, the power of orthodox ideas is also dependent on the power of the associated dominant groups.

The remainder of the paper is organized as follows. In section 2 we analyze a simple political correctness model which is extended in Section 3 to incorporate an elite. Section 4 describes the equilibria of this full model and emphasizes the new results with respect to the benchmark. Section 5 provides an empirical illustration of the main mechanism in the model. Section 6 concludes. Most technical details are relegated to the Appendix.

2 A simple model

2.1 The model ingredients

Consider a society with a continuum of individuals uniformly distributed on the line $\left[0, \overline{k}\right]$ where $k \in \left[0, \overline{k}\right]$ represents individual k's view. Among the existing views, only those located in $[0, k_R]$ with $k_R < \overline{k}$ are thought of as acceptable by society. We will refer to these views as "orthodox" or "politically correct" views and to $[0, k_R]$ as the Overton

[3] Empirically, Funke (2016) uses Swiss referenda data to show that policy areas subject to political correctness debates show the largest distortions between post vote surveys and actual election results, while Bursztyn, Egorov and Fiorin (2017) argues that elections can induce fast changes in the social acceptability of holding and expressing views. More recently, Braghieri (2021) proposes a method for assessing whether political correctness concerns actually render public discourse less informative, and shows this to be the case in an experiment. Our paper is also related to the literature on self-deny due to disagreement aversion (see in particular Asch's (1955), Golman, Loewenstein and Zarri (2016) or Fatas, Hargreaves Heap and Rojo Arjona, 2018) or to fear of expression of minority views (see in particular Prentice and Miller, 1993, and Domínguez, Taing and Molenberghs (2016).

window.

Individuals holding heterodox views, i.e. individuals with views located in the interval $\left[k_R, \overline{k}\right]$ decide whether to be "politically correct", i.e. to exert self-denial and express an orthodox view, or to express their own view (i.e. express dissent) instead.[4] Self-denial comes at a cost, and we assume this cost to be higher the more the individual's opinion differs from orthodox views. Denoting by d_i the distance between the individual's own view and the closest orthodox view (i.e. $d_i = k_i - k_R$), the value of self-denial is simply given by

$$U_i^{SD} = -d_i = -(k_i - k_R). \tag{1}$$

If, instead, the individual chooses not to hide her own view (i.e. she expresses dissent), her utility is

$$U^T(\mu) = v_T - \alpha\mu. \tag{2}$$

The dissenter enjoys a positive value from being truthful v_T but suffers a utility loss increasing in the number of self-deniers μ. The positive constant α represents the intensity of peer pressure towards political correctness.[5]

Equalizing (1) and (2), there is a cutoff distance $\widetilde{d}$ (for each given μ) from the closest acceptable view k_R below which individuals self-deny and above which they express dissent:

$$\widetilde{d} = \alpha\mu - v_T. \tag{3}$$

Taking into account that $\overline{k} - k_R$ is the total mass of people who could potentially self-deny, the proportion of self-deniers is given by

$$\mu = \begin{cases} 0 \text{ if } \widetilde{d} < 0 \\ \frac{\widetilde{d}}{\overline{k} - k_R} \text{ if } 0 < \widetilde{d} < \overline{k} - k_R \\ 1 \text{ if } \widetilde{d} > \overline{k} - k_R \end{cases}. \tag{4}$$

Introducing (3) into (4), the equilibrium level of self-denial is given by:

$$\mu = \begin{cases} 0 \text{ if } v_T > 0 \\ \frac{v_T}{\alpha - \left(\overline{k} - k_R\right)} \text{ if } 0 < \frac{v_T}{\alpha - \left(\overline{k} - k_R\right)} < 1 \\ 1 \text{ if } \alpha - v_T > \overline{k} - k_R \end{cases} \tag{5}$$

[4] In other words, people with views outside the Overton window have to decide whether to truthfully speak out or falsify their preferences, see in particular Kuran (1987) and Duffy and Lafky (2021).

[5] People with views outside the Overton window also care about the behavior of other people whose views lie outside the Overton window. This is captured mathematically by the fact that if they decide to behave truthfully, their utility decreases with the proportion of non-orthodox who self-deny, which captures how isolated they are. Suresh and Jeffrey (2017) discuss this idea of fear of isolation, sometimes referred to as the "spiral of silence" following Noelle-Neumann (1974)'s analysis of public opinion.

2.2 Equilibria in the simple model

Figure 1 presents the equilibria. Given there is a positive value associated to being truthful (v_T) and that the incentives for self-denial are always inexistent if nobody self-denies, a situation whereby all heterodox individuals truthfully express their views ($\mu = 0$) is always an equilibrium. In addition, if the peer pressure parameter α is large enough to generate the self-denial of the most heterodox agent ($\alpha > \overline{k} - k_R + v_T$), an equilibrium with full self-denial ($\mu = 1$) becomes possible, and thus the area where $\alpha > \overline{k} - k_R + v_T$ holds is characterized by multiple equilibria. As the cost of being truthful is increasing in the number of self-deniers (a bandwagon effect[6]), one equilibrium or the other arises depending on whether the agents coordinate on self-denial or instead on being truthful. In addition, there exists an interior equilibrium proportion of self-deniers so that the marginal mass member is indifferent between self-denying and being truthful.[7] Note that as the span of heterodox views becomes larger (i.e. as $\overline{k} - k_R$ increases), political correctness can arise at equilibrium only for an increasingly stronger peer pressure.

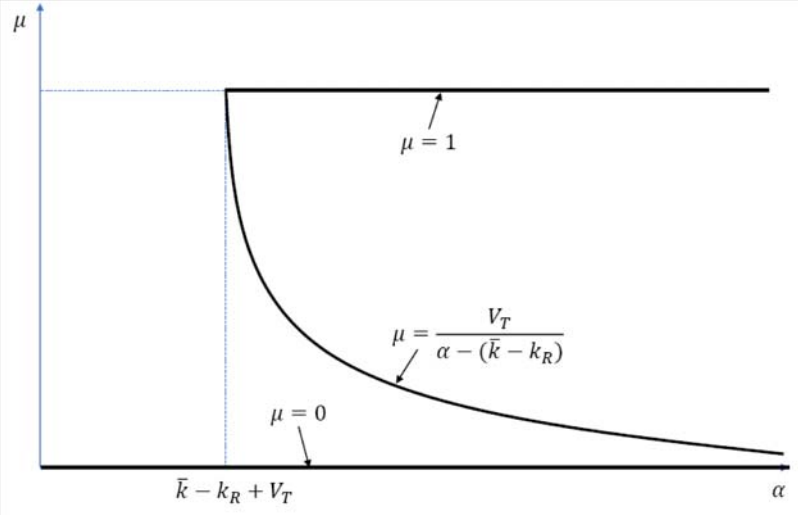


Figure 1: Equilibria in the simple model

2.3 An increase in the orthodoxy range

One important question is to understand whether a society encompassing an increasingly large set of views (i.e. including more views as orthodox as k_R becomes larger) automatically results in more individuals expressing their own views at equilibrium.

In the interior equilibrium, as represented in Figure 2, individuals express their views if they are orthodox –corresponding to the interval $(0, k_R)$– or alternatively if they are sufficiently far away from the orthodoxy –in $\left(k_R + \widetilde{d}, \overline{k}\right)$– as self-denial would in that case entail too large a cost. Instead, the individuals with views closer to the orthodoxy –in $\left(k_R, k_R + \widetilde{d}\right)$– self-deny by expressing the politically correct view k_R.

[6] The bandwagon effect is a phenomenon whereby the rate of uptake of beliefs, ideas, fads or trends by any individual is higher the larger the proportion of individuals who are already uptaking them.

[7] As shown in Figure 1, as the peer pressure parameter α increases, only individuals with views increasingly divergent from the orthodoxy self-deny.

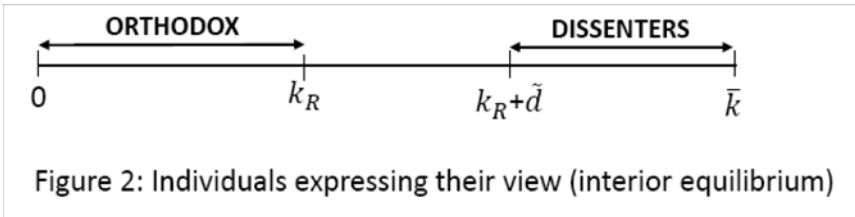


Figure 2: Individuals expressing their view (interior equilibrium)

As the total number of individuals who reveal their true views is given by

$$h(k_R) = k_R + (1-\mu)\left(\overline{k} - k_R\right), \tag{6}$$

we can see that the evolution of this variable depends on both the size of the Overton window (i.e. the value of k_R) and the equilibrium level of self-denial μ^*. Mathematically, it is easy to show that

$$\frac{dh(k_R)}{dk_R} = \mu - \frac{d\mu}{dk_R}\left(\overline{k} - k_R\right) = \frac{v_T\alpha}{\left[\alpha - \left(\overline{k} - k_R\right)\right]^2} > 0 \tag{7}$$

so the number of individuals truthfully expressing their views unambiguously increases when more views become orthodox. Clearly, this partly comes from the individuals with views in (k_R, k_R') stopping to self-deny as their views become orthodox and the cost from expressing them disappears, but the total impact on truthfulness depends on how the change in the Overton window affects those who remain heterodox. Appendix A shows that we unambiguously have that $\frac{d\left(k_R+\tilde{d}\right)}{dk_R} > 0$. Thus, as shown in Figure 3, after the initial increase in the Overton window which renders orthodox the views of individuals in (k_R, k_R'), the cost of self-denial with respect to the new limit for orthodox views k_R' goes down for potential self-deniers, and this results in individuals with views in $(k_R + \widetilde{d}, k_R' + \widetilde{d'})$ starting to self-deny. However, from (7), this effect is dominated and overall more individuals truthfully express their view.

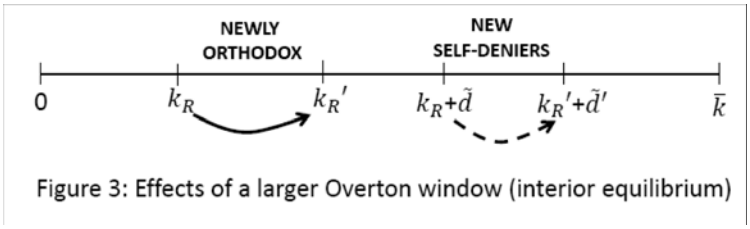


Figure 3: Effects of a larger Overton window (interior equilibrium)

So far, we have only considered how an increase in the orthodoxy range changes the characteristics of the interior equilibrium. However, from the analysis above, we know that the change in the orthodoxy range may also affect the set of equilibria. Assume that initially k_R is quite small: in such a situation, the cost of self-denial is quite high for many individuals because their views are quite different from the closest orthodox view, which implies from Figure 1 that $\mu = 0$ may be the unique equilibrium and thus nobody self-denies. As more views become orthodox, the individuals holding the newly orthodox views will still express their views, and this not as a result of an equilibrium, but simply because their views have become socially acceptable. At the same time, if the change in k_R is such that it results in $\alpha > \overline{k} - k_R + v_T$, equilibria with positive self-denial become possible and

the society potentially switches to a new equilibrium where either all individuals outside of the new orthodoxy range (in the $\mu = 1$ equilibrium) or only individuals close enough to the orthodoxy range start self-denying. This implies that a society accepting more views as orthodox might paradoxically end up at an equilibrium with a smaller total number of individuals speaking up their views if the society coordinates on a different equilibrium when the Overton window increases.

3 A model with elites

We now introduce an elite in the model. The elite are the "guardians" of the orthodoxy, either because orthodox ideas originated with them or because one of their functions in society is to represent/defend these ideas. However, in terms of actions, elite members still have a choice between behaving according to socially acceptable ideas or instead misbehaving –we also refer to this for simplicity as "being corrupt".

3.1 Behavior of the elites

An elite member who chooses to misbehave enjoys a fixed payoff from corruption, but this payoff is lowered by a penalty if the misbehavior is detected. We assume the probability of detection to be smaller the more prestigious the elite is, which is itself tied up to the prestige of orthodox views to the extent to which mass members self-deny. For simplicity, we assume the detection probability to be given by the proportion of dissenters $1-\mu$, i.e. the utility from misbehavior is

$$V_M(\mu) = \beta - c(1-\mu) \tag{8}$$

where β is the payoff from corruption, and c the misbehavior penalty arising with detection probability $1-\mu$.

On the other hand, elite member j derives a fixed utility b_j from good behavior

$$V_j^G = b_j, \tag{9}$$

and this fixed utility is uniformly distributed across the elites between $\underline{b}$ and $\overline{b}$. Putting together (8) and (9), the indifferent elite member is characterized by

$$\widetilde{b}\,(\mu) = \beta - c(1-\mu). \tag{10}$$

From (10), the proportion of honest elite members is determined by

$$e\,(\mu) = \begin{cases} 0 \;\text{ if }\; \overline{b} < \beta - c(1-\mu) \\ \frac{\overline{b}-\beta+c-c\mu}{\overline{b}-\underline{b}} \;\text{ if }\; \overline{b} > \beta - c(1-\mu) > \underline{b} \\ 1 \;\text{ if }\; \beta - c(1-\mu) < \underline{b} \end{cases} \tag{11}$$

i.e. depending on the characteristics of the distribution, on the behavior μ of the masses and on other parameters, we can have an interior value for the threshold (as in the second line of (11)) or a situation in which no elite members or all elite members are honest

(respectively the first and third lines). The bounds in (11) can be rewritten as bounds on the proportion of self-deniers as follows:

$$e(\mu) = \begin{cases} 0 \text{ if } \mu > \frac{\overline{b}-\beta+c}{c} \\ \frac{\overline{b}-\beta+c-c\mu}{\overline{b}-\underline{b}} \text{ if } \frac{\underline{b}-\beta+c}{c} \leq \mu \leq \frac{\overline{b}-\beta+c}{c} \\ 1 \text{ if } \mu < \frac{\underline{b}-\beta+c}{c} \end{cases} \tag{12}$$

which defines the reaction function of the elite to the masses' behavior. To allow the behavior of the masses to influence the elite we assume that $\underline{b} < \beta < \overline{b} + c$.[8]

3.2 Choice of the masses

As in the simple model, we still assume (2) to hold, i.e. expressing dissent entails a cost increasing in the number of self-deniers.

At the same time, and in contrast with (1), we now assume that the value of self-denial is increasing in the prestige of the elite, given that a better behavior of the elites raises the prestige of orthodox views. Specifically, we assume that

$$U_i^{SD}(\mu) = -d_i + ve(\mu) \tag{13}$$

where v is a positive constant measuring the relevance of elite behavior for self-denial and thus $ve(\mu)$ can be interpreted as the intensity of elite-driven self-denial.[9]

Equalizing (2) and (13), there is still a cutoff distance $\widetilde{d}$ from the closest acceptable view k_R below which people self-deny and above which people express their point of view, namely:

$$\widetilde{d} = ve(\mu) + \alpha\mu - v_T \tag{14}$$

Introducing (14) into the proportion of self-deniers derived in (4), the reaction function of heterodox masses to the elite behavior becomes:[10]

$$\mu(e) = \begin{cases} 0 \text{ if } e < \frac{v_T}{v} \\ \frac{v_T - ve}{\alpha - \left(\overline{k} - k_R\right)} \text{ if } 0 < \frac{v_T - ve}{\alpha - \left(\overline{k} - k_R\right)} < 1 \\ 1 \text{ if } e > \frac{v_T}{v} + \frac{\left(\overline{k} - k_R\right) - \alpha}{v} \end{cases} \tag{15}$$

The reaction function (15) illustrates the new forces at play due to the existence of the elite. Without the elite, the only reason for self-denial is the fear of isolation captured by peer pressure α. Now, instead, mass members might choose to self-deny even in the absence of peer pressure if elite's good behavior sufficiently raises the value of self-denial, i.e. if $ve(\mu)$ is sufficiently large.

Indeed, comparing the first branch of (15) with the first branch of (5), it appears that the equilibrium with no self-denial ($\mu = 0$) does not always exist anymore, and a condition

[8] If $\beta < \underline{b}$, the elites are always honest, while if $\beta > \overline{b} + c$ the elites are always dishonest. Observe that the different branches of (12) are connected and never exist simultaneously, which implies that $e(\mu)$ is always unique.

[9] Notice that for $v = 0$ we are back to our model without an elite. The elite makes it more attractive to belong to society, hence the opportunity cost to express dissent is higher.

[10] See Appendix B.1 for details.

for its existence is now that the value of elite-driven self-denial ve for the least heterodox individual (i.e. with a view $k_i \to k_R$) is smaller than the value of being truthful, v_T.

At the other extreme, the third branch of (15) represents a situation of full self-denial ($\mu = 1$), which arises whenever the individual with the lowest utility from self-denial (i.e. the most unorthodox agent) has a payoff $ev - \left(\overline{k} - k_R\right)$ from self-denial larger than the payoff from being truthful when everybody else self-denials ($v_T - \alpha$).

Finally, the intermediate branch of (15) indicates that there will be an interior equilibrium proportion μ of self-deniers if the least unorthodox individual prefers to self-deny as $ve > v_T - \mu\alpha$ for her, while at the same time the most unorthodox agent does not self-deny given that $v_T - \mu\alpha > ve - \left(\overline{k} - k_R\right)$.

For $\alpha < \overline{k} - k_R$, i.e. in a situation where peer pressure for political correctness is mild, the three branches of the reaction function (15) do not overlap and (if they exist) are connected. Thus, in contrast with the simple model where $\mu = 0$ was the unique equilibrium for $\alpha < \overline{k} - k_R$, the presence of the elites makes self-denial possible at equilibrium as their prestige creates an extra cost for expressing dissent.

For $\alpha > \overline{k} - k_R$, in turn, i.e. when political correctness is "high-stakes", the three branches of the reaction function overlap for $v_T + \left(\overline{k} - k_R\right) - \alpha < ve < v_T$ due to the bandwagon effect, giving thus potentially rise to multiple equilibria. This is the same type of outcome that was observed in the simple model for $\alpha > \overline{k} - k_R$ when $\alpha > \overline{k} - k_R + v_T$.

4 Equilibria in the model with elites

An equilibrium is a proportion of self-deniers and a proportion of honest elite members (e, μ) satisfying both the reaction function of the elite and that of the masses, i.e. respectively (12) and (15).

As in the simple model, more self-denial raises the isolation of those that would express dissent and renders self-denial more attractive by lowering the value of being truthful, see (2). At the same time, more self-denial now generates incentives to misbehavior by the elites (lowers $\widetilde{b}(\mu)$ in (10)), which in turn weakens the value of self-denial as shown in (13).

Proposition 1 in Appendix B.3 fully characterizes the equilibrium when the reaction function of the masses is unique (i.e. $\alpha < \overline{k} - k_R$) and shows the equilibrium to be always unique in that case.[11] In turn, Proposition 2 in Appendix B.4 considers the case where several branches of the reaction function of the masses are simultaneously active ($\alpha > \overline{k} - k_R$) and shows that multiple equilibria arise in different parts of the parameter space.[12]

We now graphically present the equilibria as a function of peer pressure (α) and the relevance of elite behavior for self-denial (v) with the equilibrium without elites corresponding to the particular case where $v = 0$. When doing so, the different equilibrium

[11] If we draw this unique reaction function in the $e - \mu$ space, the slope of the interior branch is positive as $\frac{\partial \mu}{\partial e} = \frac{v}{\overline{k} - k_R - \alpha} > 0$.

[12] If $\alpha > \overline{k} - k_R$, the interior branch of (15) is downward sloping in the $e - \mu$ space. As the slope of the elite's interior behavior is also downward sloping in that space, the stability of the interior equilibrium depends on which interior branch is steeper: if there is an intersection and the slope of the masses' reaction function is flatter (resp. steeper), the equilibrium is stable (resp. unstable).

configurations end up depending on the two parameters characterizing the payoff from corruption for the elite, namely the fixed payoff β and the misbehavior penalty c linked to the number of self-deniers. The full characterization of the equilibrium in the (α, v) space is presented in Proposition 3 (see Appendix B.2).

4.1 Case 1: Low corruption payoff

Figure 4 presents the equilibria when the payoff from corruption β is low enough ($\underline{b} < \beta < \min[\underline{b} + c, \overline{b}]$) to discourage full corruption ($e = 0$) as a best response for the elite.

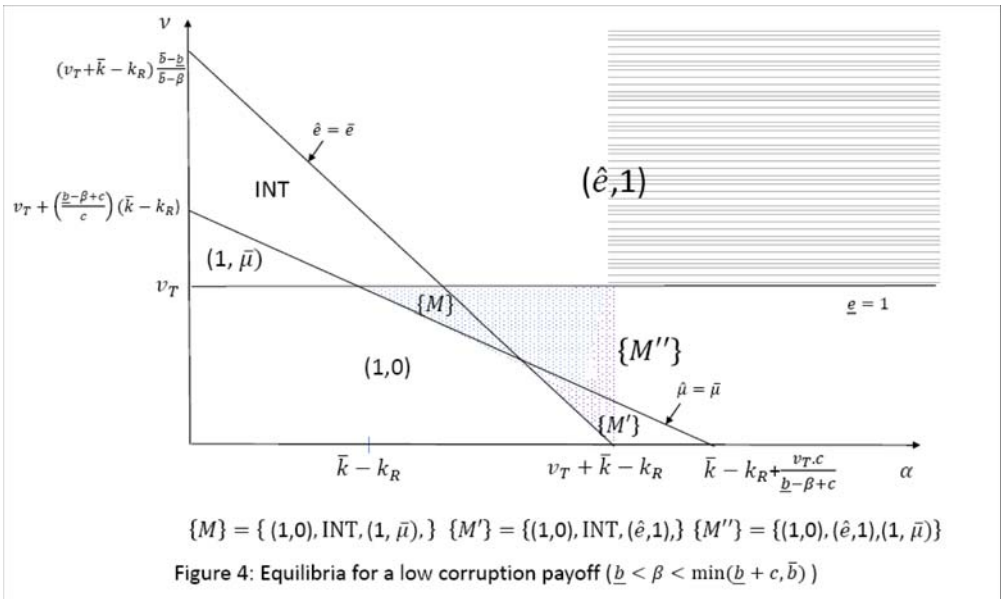


Figure 4: Equilibria for a low corruption payoff ($\underline{b} < \beta < \min(\underline{b} + c, \bar{b})$)

In the model without elites ($v = 0$), multiple equilibria arise when a large peer-pressure ($\alpha > \overline{k} - k_R + v_T$) results in a strong enough bandwagon effect. When peer pressure is low, the equilibrium with no self-denial is unique.

As the behavior of the elite becomes more and more relevant to the masses, i.e. v becomes larger, multiple equilibria start arising also for smaller and smaller values of α, as shown in the shaded area in the South-West of the figure. The intuition for this is simple: as the elite behaves (relatively) well in this area, this reinforces self-denial by the masses, and thus the equilibrium with self-denial remains now feasible despite a relatively low peer-pressure, implying that self-denial is here partly elite-driven.

At some point ($v > v_T$), elite-driven self-denial is strong enough to outweigh any incentives for a truthful expression of views resulting from a low peer pressure among the masses or from a coordination in the no self-denial equilibrium through the bandwagon effect. Thus, relative to the simple model, multiple equilibria vanish in the stripped area in the North-East and full self-denial now becomes the unique equilibrium also in that area. Intuitively, in this area, the prestige of orthodox ideas stemming from full self-denial lowers the corruption detection probability and this systematically generates corruption at equilibrium.

More generally, corruption is more likely to arise whenever self-denial is higher, as higher self-denial makes it less likely for misbehavior to be detected.

4.2 Case 2: Intermediate corruption payoff

Consider next a situation characterized by a larger fixed corruption payoff β, and specifically $\min[\underline{b}+c, \overline{b}] < \beta < \max[\underline{b}+c, \overline{b}]$. We show that two equilibrium configurations arise depending on whether the misbehavior penalty c is small or large.

4.2.1 Case 2-low: Small misbehavior penalty

In the case of a larger corruption payoff β, the elite has clearly stronger misbehavior incentives, and this tendency is reinforced if the misbehavior penalty is small ($\underline{b}+c < \beta < \overline{b}$), see Figure 5. As a result, full honesty ($e = 1$) is no longer an equilibrium outcome and this explains why the equilibrium in the South-West is now $(\widetilde{e}, 0)$ instead of $(1, 0)$ in the preceding case, and why $(1, \overline{\mu})$ is not anymore an equilibrium. Otherwise, the equilibrium configuration is qualitatively similar to the preceding case, with multiple equilibria still located in the South-East in a larger area than for $v = 0$ (see shaded triangle) but still a North-East quadrant (for $v > v_T \frac{\overline{b}-\underline{b}}{\overline{b}-\beta+c}$ and $\alpha > \overline{k} - k_R + v_T$) in which multiplicity vanishes and full self-denial is the unique equilibrium (see stripped area).

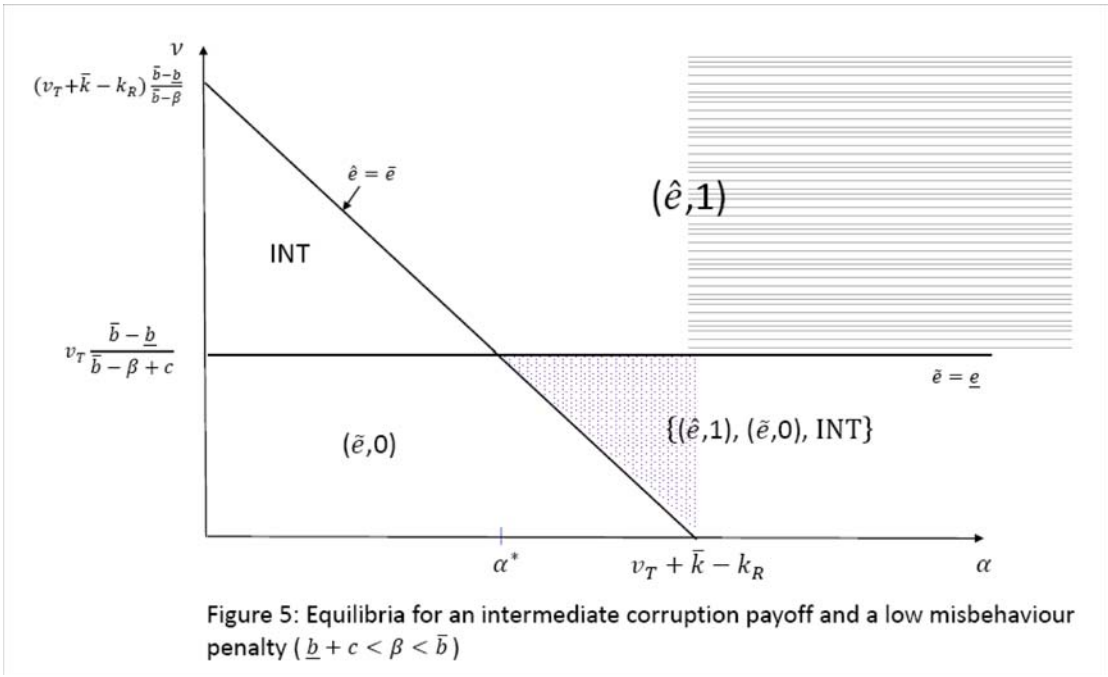
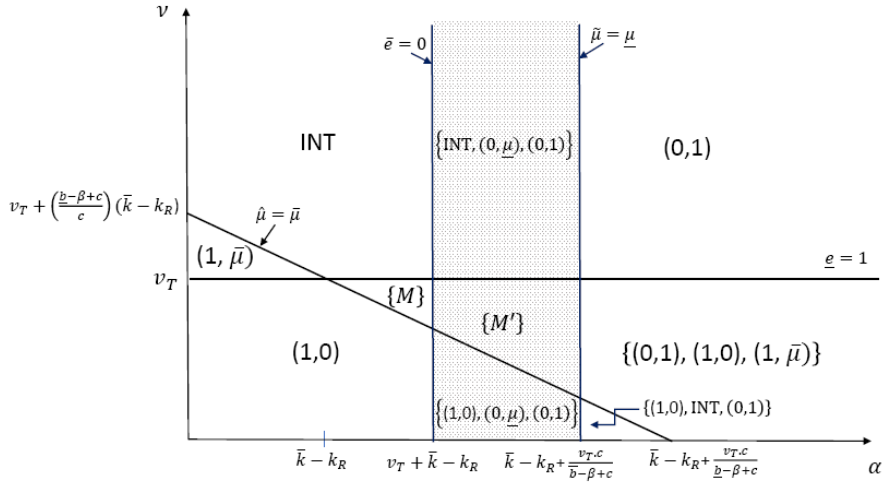


Figure 5: Equilibria for an intermediate corruption payoff and a low misbehaviour penalty ($\underline{b}+c<\beta<\bar{b}$)

4.2.2 Case 2-high: Large misbehavior penalty

A large misbehavior penalty ($\overline{b} < \beta < \underline{b} + c$) instead restores no-corruption as a possible equilibrium outcome. Indeed, if the masses do not fully self-deny, they can detect elite corruption with a positive probability, and the large misbehavior penalty makes corruption a bad choice for the elite in some cases. At the same time, full corruption also becomes a possible equilibrium outcome for the first time, because in the presence of full self-denial by the masses, the probability of detection of elites' corruption is zero, and the high misbehavior penalty becomes in that case irrelevant in a context where the corruption pay-off is non negligeable.

In terms of the presence of multiple equilibria, the main difference with respect to the two preceding cases is that, for intermediate values of α, multiple equilibria now arise for any value of v (see the shaded area in Figure 6). Indeed, while all three cases are characterized by multiple equilibria for intermediate values of v, in Figure 6 multiple equilibria arise also for small and large values of v. For instance, a high v generates self-denial among the masses, reducing the corruption detection probability and pushing the elite towards full corruption, and we can then end up in an equilibrium in which the full adhesion of the masses to the ideas of the elite results in full corruption by the elite. In the two preceding cases, instead, while the high v still generated self-denial by the masses and lowered the detection probability, full corruption was never an equilibrium outcome for the elite.



$\{M\}=\{(1,\bar{\mu}),(1{,}0),\text{INT}\}$ and $\{M'\}=\left\{(1{,}0),(1,\bar{\mu}),\text{INT},(0{,}1),(0,\underline{\mu})\right\}$

Figure 6: Equilibria for an intermediate corruption payoff and a high misbehaviour penalty ($\bar{b}<\beta<\underline{b}+c$)

Finally, in the areas characterized by a unique equilibrium, there is still an equilibrium with no-corruption by the elite and no self denial by the masses in the South-West, and in the North-East an equilibrium with no honest behavior and full self-denial –this time $(0,1)$ instead of $(\hat{e},1)$ because the corruption payoff is larger ($\beta>\overline{b}$ instead of $\beta<\overline{b}$ in the two preceding cases).

4.3 Case 3: Large corruption payoff

Finally, the case of a large corruption payoff ($\max[\underline{b}+c,\overline{b}]<\beta<\overline{b}+c.$) is represented in Figure 7, with outcomes very similar to those of Figure 6 except for the fact that full honesty is not anymore an equilibrium outcome and at most a proportion e_{int} or $\widetilde{e}$ of the elites are honest.

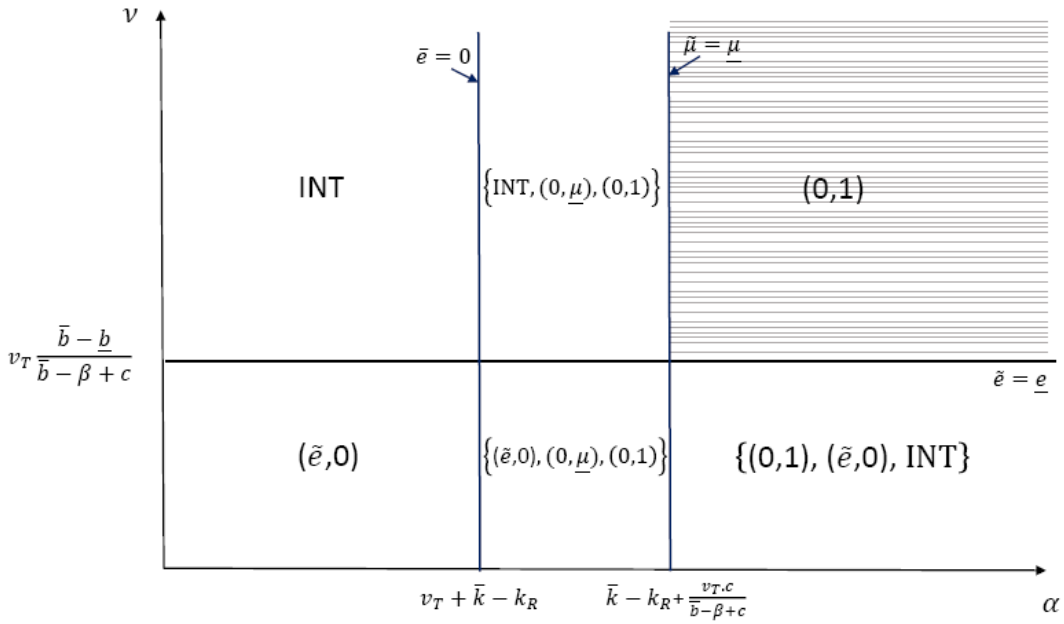


Figure 7: Equilibria for a high corruption payoff ($\max(\underline{b}+c, \bar{b}) < \beta < \bar{b}+c$)

Appendix B.6 presents how the different cases arise as β changes and shows there is a simple pattern explaining how the equilibrium configuration changes with β. The next section summarizes how.

4.4 An increase in the orthodoxy range with elites

We next examine how changes in the size of the Overton window k_R affect the total number of individuals $h(k_R)$ expressing their own views at equilibrium, see (6). To that purpose, we first need to understand which are the relevant equilibria for different values of k_R. This is actually quite straightforward: since both k_R and α appear with a positive sign in the denominator of the reaction function of the masses (15), the figures characterizing the equilibrium in the space (k_R, ν) are actually identical to those presented earlier on in this section in the space (α, ν).

Figure 8 represents a case where the masses care little about the behavior of the elite (low ν) and the payoff from corruption is low (a represented in Figure 4).[13] The horizontal line on top of the figure indicates that in the no self-denial equilibrium (which arises for every value of k_R) everybody expresses their view no matter the value of k_R i.e. independently on whether their view is orthodox or not. For larger values of k_R, an equilibrium with full self-denial always exist: in that case, only the individuals with newly orthodox views join the set of those who express their views as k_R grows, and $h(k_R)$ corresponds for this reason to the 45° line. Most interestingly, we see that in the case of the equilibrium with partial self-denial, $h(k_R)$ is monotonically increasing in k_R, i.e. we get the same qualitative results as in the benchmark model, which is reassuring given that the elite matters little for the masses: a fall in h as k_R goes up can only happen if there is a switch to a different equilibrium. Also, we can see that initial increases in k_R

[13] Specifically, we set $\upsilon = 1$, $\overline{k} = 10$, $\alpha = 9$, $\underline{b} = 2$, $\overline{b} = 6$, $\beta = c = 3$, and $v_T = 2$ in this example.

are reinforced by lower self-denial among the heterodox, and instead further increases in k_R having a smaller impact on h because the fraction of self-deniers starts to grow.

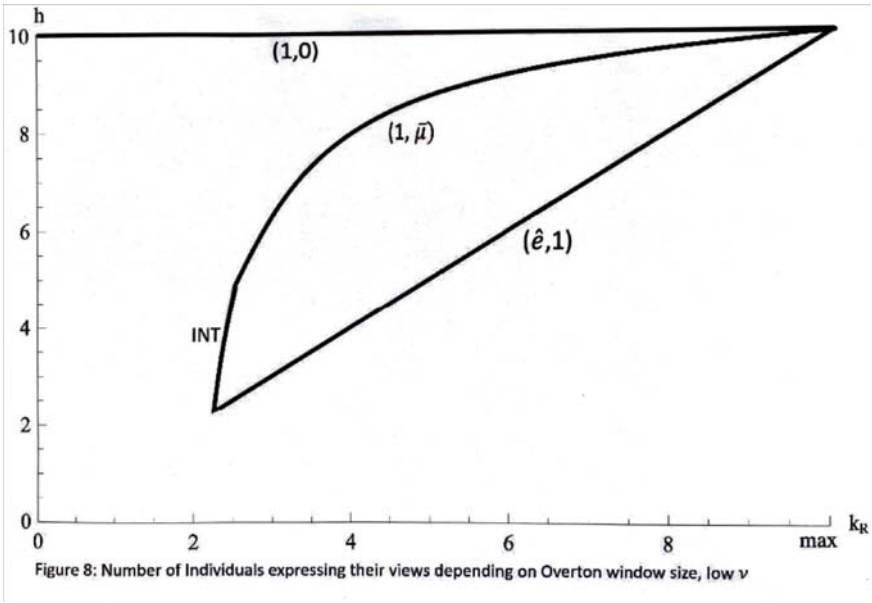


Figure 8: Number of Individuals expressing their views depending on Overton window size, low ν

Figure 9 represents a case with a low corruption payoff (as represented in Figure 4) when the masses care to an important extent about the elite, i.e. $\nu > \nu_T$.[14] The graph represents the three unique equilibria that are relevant for different values of k_R, namely two consecutive interior equilibria for low values of k_R and then a full-self denial equilibrium when k_R becomes larger. In contrast with the benchmark model, we can see that for the two interior equilibria, the initial increase in the Overton range leads actually to an increase in self-denial that more than offsets it in the sense that the overall number of individual expressing their own views at equilibrium ends up falling. So although the group considers a larger set of views as being acceptable, a smaller variety of views is uttered at equilibrium. Intuitively, consider the situation of a "very heterodox" individual. Initially, her views were so far away from the orthodoxy, that the cost of self-denying was prohibitive. Instead, when views closer to hers become socially acceptable because the Overton window becomes larger, self-denial becomes "cheaper". While this type of effect exists also in the benchmark model, this second induced effect is never large enough to outweigh the initial effect. This becomes possible instead in the model with elite because the initial fall in the in the proportion of self-deniers boosts the corruption detection probability, which improves the behavior of the elite and further reinforces self-denial. Figures A1 and A2 in appendix B.7 represent the cases for other areas of the parameter space, with a downward sloping $h(k_R)$ arising in some cases.

[14] Specifically, we keep the same parameters as in Figure 8 except for $\upsilon = 2.5$.

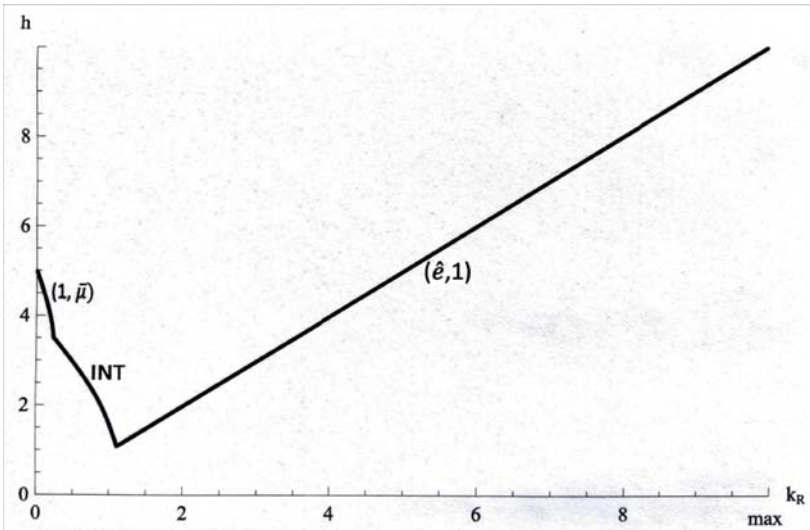


Figure 9: Number of Individuals expressing their views depending on Overton window size, low β and high ν

The next section considers some predictions of the model in light of the Survey of American Political Culture.

A blue line represents any equilibrium where nobody self-denies while a red line represents any equilibrium where all heterodox individuals express politically correct views. All other colors refer to intermediate levels of self-denial.

5 Empirical illustration based on the Survey of American Political Culture

The 2016 Survey of American Political Culture (Hunter and Bowman, 2016) includes a question on the perception of political correctness as a problem. This survey consists of 1,904 telephone interviews by Gallup on a representative sample of American adults and was performed in late August 2016, i.e. two months ahead of the election of Donald Trump. Respondents were questioned about the extent to which they agreed with the statement "political correctness is a serious problem in the U.S., making it hard for people to say what they really think" (Hunter and Bowman, 2016, p. 67). In the sample, 40 percent of the individuals completely agree with the statement, 33 percent mostly agree, 17 percent mostly disagree and 10 percent completely disagree.

In our model, if all views were orthodox, everybody would simply express their own view and get a payoff v_T. When not all views are orthodox, this generates a loss to some individuals through social pressure for political correctness. Figure 10 studies the size of this loss depending on the view held by the individual. Individuals holding an orthodox view –located in $(0, k_R)$– achieve v_T as they can express their view at no cost. Moving towards the right in Figure 10, self-deniers –located in $(k_R, k_R + \widetilde{d})$– get a payoff $-d_i + ve(\mu)$ instead of the v_T they would obtain if their view was orthodox, and thus lose $-d_i + ve(\mu) - v_T$. Political correctness is a cost to them because they end up expressing a view that provides them with a lower utility. Finally, unorthodox individuals who choose to speak up their view –located in $(k_R + \widetilde{d}, \overline{k})$– have a pay-off of $v_T - \alpha\mu$, and thus lose $-\alpha\mu$ from their view not being orthodox. As $d_i + v_T < ve(\mu) + \alpha\mu$ from (14) for $d_i < \widetilde{d}$, the loss of unorthodox individuals truthfully expressing their views is even larger than

that of self-deniers. To them, political correctness is a problem because it stigmatizes the view they are expressing. Putting together these three different segments, our model thus predicts that the individual loss from the social pressure for political correctness is (weakly) increasing in the distance of the individual with respect to the orthodoxy.

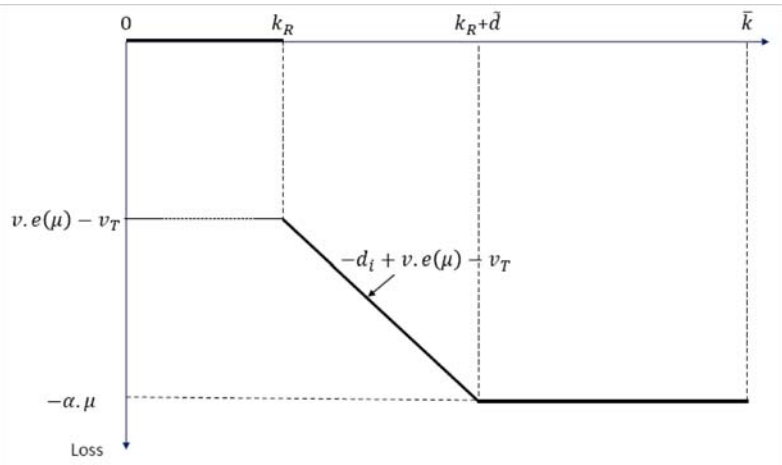


Figure 10: Loss to individual i associated to the existence of orthodoxy range $[0, k_R]$

Our measure of distance of the individual i with respect to the orthodoxy range is constructed from a set of questions asking the respondent whether the beliefs and values of Americans like them are completely different, mostly different, mostly similar, or completely the same as those of different subgroups of the population. These subgroups are African Americans, Hispanic Americans, White Americans, Muslim or Islamic Americans, Conservative Christians, Non-religious people, and gays and lesbians. Specifically, let s_{ig} be the level of similarity with respect to group g, we use $s_{ig} = 1$ whenever the individual says she shares the same values with group g, $s_{ig} = 2/3$, when she shares mostly similar values, $s_{ig} = 1/3$ when the values are mostly different and $s_{ig} = 0$ when they are completely different. Then, aggregating across groups, we construct an average measure of distance for each individual i, namely

$$d_i = \frac{1}{G}\sum_{g=1}^{G} 1 - s_{ig}.$$

The second main variable of interest is the extent to which the respondent believes that "the most educated and successful people in America are more interested in serving themselves than in serving the common good", for which again four possible answers are given (completely agree, mostly agree, mostly disagree, and completely disagree). In the sample, 21 percent of the individuals strongly agree with this statement, 40 percent mostly agree, 30 percent mostly disagree and 8 percent completely disagree. Heterogeneity in the beliefs about the elite can be easily introduced in our model by assuming that the effort produced by the elite $e(\mu)$ is not perfectly observable and that individuals differ on their perception η_i of that effort. In that case, the utility from self-denial (13) becomes $U_{SD}(\mu) = -d_i + v.\eta_i e(\mu)$ and self-denial is less costly for individuals with a more positive view on the elites (i.e. a higher η_i). Intuitively, individuals with a higher η_i value the orthodoxy because they have a higher opinion on the behavior of the guardians of the

orthodoxy, i.e. the elites. This immediately implies that the threshold distance for self-denial now depends on η_i and, as self-denial is less painful the larger η_i, individuals with a higher perception of the elite choose to self-deny even for higher distances to the orthodoxy, i.e. $\frac{\partial \widetilde{d}_i}{\partial \eta_i} > 0$. Figure 11 illustrates how the cost of holding a heterodox view changes depending on whether the individual has a high (η_H) or low (η_L) view on the elite.

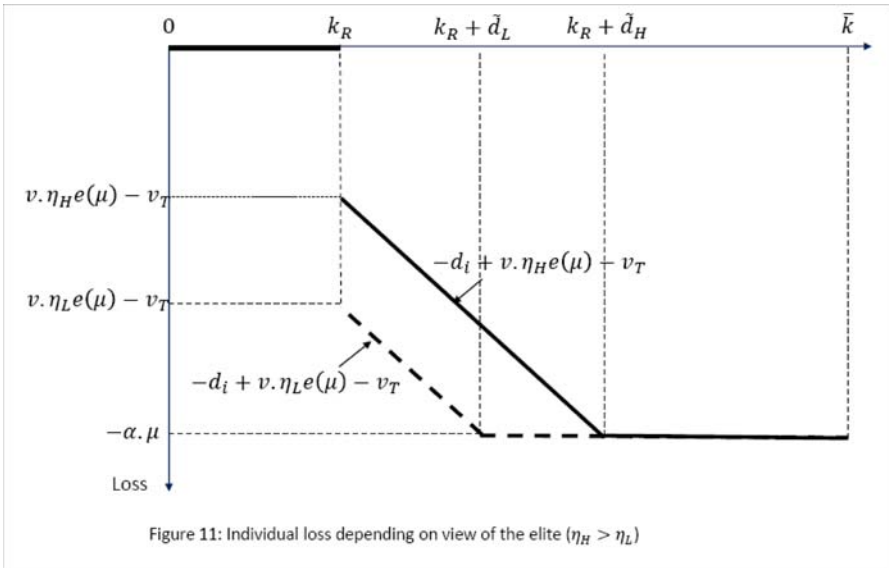


Figure 11: Individual loss depending on view of the elite ($\eta_H > \eta_L$)

Note first that individuals holding a view in $(k_R + \widetilde{d}_H, \widetilde{k})$ choose to speak up their mind no matter their perception of the elite, and thus their loss does not depend on this perception. Instead, for other heterodox individuals the loss from their views not being considered orthodox is increasing in their negative perception of the elite. Specifically, individuals holding views close to the orthodoxy – in $(k_R, k_R + \widetilde{d}_L)$- self-deny no matter their η_i, but the utility of self-denial is higher (and thus the loss lower) when η is high because a positive view of the elite raises the utility of adhering to the orthodoxy. A similar argument holds for individuals in $(k_R + \widetilde{d}_L, \widetilde{k} + \widetilde{d}_H)$, except that different perceptions of the elites for a given view result in different behaviors (self-denial for $\eta = \eta_H$ as opposed to truthfulness for $\eta = \eta_L$).

In order to be able to run a binomial probit, we pull together for the endogenous variable the individuals that completely agree and mostly agree, on the one hand, and those that completely or partly disagree with the statement. For exogenous variables, we keep separate the different existing categories. Column (1) in Table 1 shows that, as predicted by the model, individuals who view themselves as more distant from society are more likely to believe there is a political correctness problem in the U.S. In addition, in accordance too with our model, individuals considering that the elites are self-interested are also more likely to believe there is a political correctness problem.

Interestingly, columns (2) to (6) show these two results to be robust to the introduction of additional variables. Specifically, in column (2) we introduce age, gender, income and the education level. While income and age do not seem to play a role, males and individuals with less than college education[15] are systematically more likely to consider there is a political correctness problem. In turn, race is introduced from column (3) and the

[15]The omitted level of education is having at least college education.

Table 1: Determinants of the perception of the existence of a political correctness problem, Survey of American Political Culture, 2016

	Dependent variable: political correctness problem					
	(1)	(2)	(3)	(4)	(5)	(6)
Distance	.939***	.878***	.923***	.888***	.659***	.537*
	(.250)	(.269)	(.261)	(.266)	(.280)	(.280)
Self-interested elites	.505***	.394***	.426***	.396***	.439***	.433***
	(.084)	(.090)	(.092)	(.093)	(.099)	(.101)
Age		.028	.0003	-.008	-.041	-.030
		(.024)	(.025)	(.025)	(.028)	(.028)
Male		.321***	.306***	.309***	.318***	.234**
		(.086)	(.088)	(.088)	(.092)	(.097)
Income		-.002	-.002	-.002	-.004	-.005
		(.005)	(.005)	(.005)	(.005)	(.005)
High school dropout		.355***	.601***	.590***	.500**	.500**
		(.181)	(.186)	(.192)	(.213)	(.216)
High school graduate		.624***	.726***	.739***	.653***	.650**
		(.117)	(.120)	(.120)	(.125)	(.128)
Some college		.372***	.408***	.391***	.309**	.280**
		(.112)	(.116)	(.117)	(.121)	(.123)
White			.383	.291	.328	.072
			(.248)	(.243)	(.256)	(.268)
Black			-.154	-.266	-.219	-.215
			(.271)	(.264)	(.279)	(.291)
Asian			-.173	-.188	-.030	-.125
			(.346)	(.348)	(.353)	(.368)
Hispanic			-.460***	-.492***	-.616***	-.552***
			(.137)	(.139)	(.149)	(.151)
Density neighborhood				-.002**	-.002***	-.002**
				(.001)	(.001)	(.001)
Midwest				.008	-.108	-.083
				(.147)	(.153)	(.156)
South				.276**	.080	.053
				(.134)	(.142)	(.147)
West				.200	.130	.104
				(.140)	(.151)	(.155)
Christian					.069	-.019
					(.163)	(.165)
No religion					-.049	-.105
					(.180)	(.180)
Catholic					.277**	.324**
					(.133)	(.136)
Evangelical					.203	.187
					(.140)	(.143)
Religion important					.041	.004
					(.205)	(.204)
Religious conservative					1.09***	.735***
					(.182)	(.192)
Republican						.400***
						(.149)
Democrat						-.456***
						(.109)
Observations	1,765	1,692	1,641	1,628	1,608	1,598
Pseudo R^2	.04	.08	.11	.12	.18	.21

Notes: The figures reported are the coefficients obtained from probit estimation. Standard errors in parentheses. *, ** and *** denote significance at 10%, at 5%, and 1% levels, respectively. The data are from the Survey of American Political Culture, 2016.

only significant coefficient is a systematically lower perception of a political correctness problem by Hispanics. From column (4), individuals living in neighborhoods with a higher density are shown to systematically have a lower perception, while no significant regional effect is generally found. In columns (5) and (6) different variables on the religious beliefs/attitudes to religion are introduced, and both Catholics and individuals perceiving themselves as conservatives in terms of religion believe the problem to be more acute. Finally, and unsurprisingly, column (6) shows that individuals declaring to be Republicans (resp. Democrats) are more (resp. less) likely to perceive the existence of a political correctness problem.

6 Conclusion

If the views individuals hold are partly based on their own experiences or interests, and group belonging is correlated with specific experiences or interests, one would expect that an association is made between having certain views and belonging to a certain group. For this reason, this paper argues that the prevalence of political correctness cannot be fully understood if one abstracts from considering that the prestige of orthodox ideas is partly linked to the prestige of the elites perceived as representing them.

Specifically, while in a standard model self-denial is unlikely when peer pressure is low, the link with elite prestige makes self-denial a possible equilibrium outcome. Conversely, full self-denial becomes the unique equilibrium when the masses care sufficiently about the behavior of the elite in situations in which self-denial is just one of the possible outcomes in the standard model. In addition, we show that exogenous increases in the range of socially acceptable views do not necessarily imply anymore that a larger set of views are actually expressed at equilibrium.

Our model considers a linear distribution of views and assumes that heterodox views are only located at one extreme of the distribution. An extension of the model where heterodox views are situated at both ends of the distribution might be useful as in some cases observed changes in the Overton window may imply that formerly 'very' orthodox views become heterodox, as implied for instance by the cultural backlash theory (Norris and Inglehart, 2019).

Appendix

A Increase in the orthodoxy range in the simple model:

It is easy to show that $\frac{d\tilde{d}}{dk_R} = \frac{-\alpha v_T}{\left[\alpha-\left(\overline{k}-k_R\right)\right]^2}$ and thus $\frac{d\left(k_R+\tilde{d}\right)}{dk_R} = \frac{\left[\alpha-\left(\overline{k}-k_R\right)\right]^2-\alpha v_T}{\left[\alpha-\left(\overline{k}-k_R\right)\right]^2}$, which is negative if and only if $v_T > \frac{\left[\alpha-\left(\overline{k}-k_R\right)\right]^2}{\alpha}$. As we are in an interior equilibrium, this needs to be compatible with $\alpha > \left(\overline{k}-k_R\right)+v_T \Leftrightarrow v_T < \alpha-\left(\overline{k}-k_R\right)$. A necessary condition for both conditions to hold simultaneously is $\alpha-\left(\overline{k}-k_R\right) > \frac{\left[\alpha-\left(\overline{k}-k_R\right)\right]^2}{\alpha} \Longleftrightarrow 0 > -\left(\overline{k}-k_R\right)$, which is a contradiction, and thus $\frac{d\left(k_R+\tilde{d}\right)}{dk_R} > 0$ always.

B Model with elites

B.1 Reaction function of the heterodox masses

The proportion of self-deniers is given by

$$\mu(e) = \begin{cases} 0 \text{ if } \widetilde{d} < 0 \\ \frac{\widetilde{d}}{\overline{k}-k_R} \text{ if } 0 < \widetilde{d} < \overline{k} - k_R \\ 1 \text{ if } \widetilde{d} > \overline{k} - k_R \end{cases}$$

Introducing (14) into this expression we get the reaction function of the heterodox masses as:

$$\mu(e) = \begin{cases} 0 \text{ if } ve - v_T < 0 \\ \frac{ve+\alpha\mu-v_T}{\overline{k}-k_R} \text{ if } 0 < ve + \alpha\mu - v_T < \overline{k} - k_R \\ 1 \text{ if } ve + \alpha - v_T > \overline{k} - k_R \end{cases}$$

Simplifying, we get

$$\mu(e) = \begin{cases} 0 \text{ if } e < \frac{v_T}{v} \\ \frac{v_T - ve}{\alpha - \left(\overline{k}-k_R\right)} \text{ if } 0 < \frac{\left(\overline{k}-k_R\right)(v_T-ve)}{\alpha-\left(\overline{k}-k_R\right)} < \overline{k} - k_R \\ 1 \text{ if } e > \frac{v_T}{v} + \frac{\left(\overline{k}-k_R\right)-\alpha}{v} \end{cases}$$

Using $\overline{k} - k_R > 0$, the reaction function of the heterodox masses becomes (15).

B.2 Equilibrium analysis

We first introduce notation that will be useful when jointly considering the best responses of the elites and the masses.

- The proportion of honest elite members when all masses self-deny ($\mu = 1$) is denoted by $\widehat{e}$ and given by:

$$\widehat{e} = e(\mu = 1) = \frac{\overline{b} - \beta}{\overline{b} - \underline{b}} \tag{16}$$

  while from the masses' interior branch, full self denial can only arise if honesty among the elites is above a threshold $\overline{e}$ given by:

$$\overline{e} \equiv \min\left[e^{-1}(\mu = 1)\right] = \frac{v_T - \left(\alpha - \left(\overline{k} - k_R\right)\right)}{v} \tag{17}$$

  Hence, whenever $\widehat{e} > \overline{e}$ full self-denial is part of the equilibrium. Similarly, the proportion of honest elite members when no mass member self-denies ($\mu = 0$) is denoted by $\widetilde{e}$ and given by:

$$\widetilde{e} = e(\mu = 0) = \frac{\overline{b} - \beta + c}{\overline{b} - \underline{b}} \tag{18}$$

while no self-denial can only arise if honesty among the elites is below a threshold $\underline{e}$ given by:

$$\underline{e} \equiv \max[e^{-1}(\mu = 0)] = \frac{v_T}{v}. \tag{19}$$

Hence, whenever $\widetilde{e} < \underline{e}$, no self-denial is part of the equilibrium.

- We next turn to similar notation for the behavior of the masses. The proportion of self-deniers when the entire elite is honest ($e = 1$) is denoted by $\overline{\mu}$ and given by:

$$\overline{\mu} = \mu(e = 1) = \frac{v - v_T}{\left(\overline{k} - k_R\right) - \alpha}. \tag{20}$$

while the largest proportion of self-deniers compatible with fully honest elites is the proportion of self-deniers that makes the most dishonest elite member indifferent between behaving honestly or not i.e. $\widehat{\mu}$ such that $\underline{b} = \beta - c(1 - \widehat{\mu})$ or

$$\widehat{\mu} \equiv \max[\mu^{-1}(e = 1)] = \frac{\underline{b} - \beta + c}{c} \tag{21}$$

Hence, full honesty $e = 1$ is part of the equilibrium whenever $\widehat{\mu} > \overline{\mu}$. In turn, the proportion of self-deniers when nobody in the elite is honest ($e = 0$) is denoted by $\underline{\mu}$ and given by:

$$\underline{\mu} = \mu(e = 0) = \frac{v_T}{\alpha - (\overline{k} - k_R)} \tag{22}$$

At the same time, the lowest proportion of self-deniers that induces full dishonesty among the elites ($e = 0$) is the proportion of self-deniers that makes the most honest elite member indifferent between being honest or not i.e. $\overline{b} = \beta - c(1 - \widetilde{\mu})$ or

$$\widetilde{\mu} = \min[\mu^{-1}(e = 0)] = \frac{\overline{b} - \beta + c}{c} \tag{23}$$

Hence, full dishonesty $e = 0$ is part of the equilibrium whenever $\underline{\mu} > \widetilde{\mu}$.

For the full equilibrium analysis, we need to keep in mind that both e and μ must lie between 0 and 1. Lemma 1 presents first the conditions for e :

Lemma 1 *(i) For $\underline{b} + c < \beta < \overline{b}$, all the values of e are interior (ii) If both $\beta < \overline{b}$ and $\beta < \underline{b} + c$ then both an interior value of e and $e = 1$ are possible (iii) If $\overline{b} < \beta < \underline{b} + c$ then both corner and interior values of e are possible (iv) If both $\underline{b} + c < \beta$ and $\overline{b} < \beta$ hold, then interior values of e and $e = 0$ are possible.*

Proof In the reaction function of the elite (12), if $\overline{b} > \beta$ then the condition for the first branch becomes $\mu > \frac{\overline{b} - \beta + c}{c} > 1$ so $e = 0$ cannot be part of an equilibrium. If $\underline{b} + c < \beta$ then in the third branch $\mu < \frac{\underline{b} - \beta + c}{c} < 0$ so $e = 1$ cannot be part of an equilibrium. If $\beta < \underline{b}$, $\frac{\underline{b} - \beta + c}{c} > 1$ which implies that the condition for the third branch is always satisfied and thus $e = 1$ can always be part of an equilibrium. If $\overline{b} + c < \beta$, $\frac{\overline{b} - \beta + c}{c} < 0$ which implies

that the condition for the first branch is always satisfied and thus $e = 0$ can always be part of an equilibrium. ■

Consider next the reaction function of the heterodox masses (15). As $\mu = \frac{ve - v_T}{\left(\overline{k} - k_R\right) - \alpha}$ in the second branch, the possible values of e compatible with $0 < \mu < 1$ depend on whether the denominator $\overline{k} - k_R - \alpha$ is positive or negative. Intuitively, $\alpha < \overline{k} - k_R$ corresponds to a situation in which political correctness is a low-stakes issue compared to the existing span of heterodox views, while instead for $\alpha > \overline{k} - k_R$ political correctness is a high-stakes issue.

B.3 Low-stakes political correctness ($\alpha < \overline{k} - k_R$)

In this case, $\underline{e} < \overline{e}$ and $\underline{\mu} < 0$, and the reaction function of the masses (15) can be rewritten as

$$\mu = \begin{cases} 0 \text{ if } e < \frac{v_T}{v} \\ \frac{ve - v_T}{\left(\overline{k} - k_R\right) - \alpha} \text{ if } \frac{v_T}{v} < e < \frac{v_T - \left(\alpha - \left(\overline{k} - k_R\right)\right)}{v} \\ 1 \text{ if } e > \frac{v_T}{v} + \frac{\left(\overline{k} - k_R\right) - \alpha}{v} \end{cases} \tag{24}$$

which gives rise to three different possible relations between μ and $0 \leq e \leq 1$ depending on the relevant branch of (24).

1. If $v < v_T$ then $\mu = 0$ always.
2. If $v > v_T$ and $v - v_T < \left(\overline{k} - k_R\right) - \alpha$: for $0 \leq e \leq \underline{e}$, $\mu = 0$ always holds, while μ is linearly increasing in e for $e > \underline{e}$, reaching its maximum at $\overline{\mu}$.
3. If $v_T < v$ and $v - v_T > \left(\overline{k} - k_R\right) - \alpha$: for $0 \leq e \leq \underline{e}$, $\mu = 0$ always holds, while μ is linearly increasing in e for $e > \underline{e}$ reaching its maximum at $\mu = 1$ for $\overline{e} = \frac{v_T - \left(\alpha - \left(\overline{k} - k_R\right)\right)}{v}$ and staying at that value for $\overline{e} < e < 1$.

In this case, the equilibrium is characterized by Proposition 1:

Proposition 1 *For $\alpha < \overline{k} - k_R$, the equilibrium (e, μ) is always unique and given by:*

1. *For $v < v_T$: (1i) $(\min[\widetilde{e}, 1], 0)$ if $\underline{b} + c < \beta$, with $\widetilde{e}$ relevant for $\beta < \overline{b} + c$ (1ii) $(1, 0)$ if $\underline{b} + c > \beta$ (1iii) $(0, 0)$ if $\beta > \overline{b} + c$.*
2. *For $v_T < v < v_T + \left(\overline{k} - k_R\right) - \alpha$: (2i) (e_{int}, μ_{int}) if $\widetilde{e} > \underline{e}$ and $\widehat{\mu} < \overline{\mu}$ where $e_{int} \equiv \frac{\overline{b} - \beta + c - c\left(\frac{v_T - v\frac{\overline{b} - \beta + c}{\overline{b} - \underline{b}}}{\left(\alpha - \left(\overline{k} - k_R\right)\right) - \frac{vc}{\overline{b} - \underline{b}}}\right)}{\overline{b} - \underline{b}}$ and $\mu_{int} \equiv \frac{v_T - v\frac{\overline{b} - \beta + c}{\overline{b} - \underline{b}}}{\left(\alpha - \left(\overline{k} - k_R\right)\right) - \frac{vc}{\overline{b} - \underline{b}}}$ (2ii) $(1, \overline{\mu})$ if $\widetilde{e} > \underline{e}$ and $\widehat{\mu} > \overline{\mu}$. (2iii) $(\max[0, \widetilde{e}], 0)$ if $\widetilde{e} < \underline{e}$.*
3. *For $v > v_T$ and $v - v_T > \left(\overline{k} - k_R\right) - \alpha$: (3i) (e_{int}, μ_{int}) if $\overline{e} > \widehat{e}$ and $\underline{e} < \widetilde{e}$. (3ii) $(\max[0, \widetilde{e}], 0)$ if $\overline{e} > \widehat{e}$ and $\underline{e} > \widetilde{e}$. (3iii) $(\min[\widehat{e}, 1], 1)$ if $\overline{e} < \widehat{e}$.*

Proof This is done by combining (24) with the reaction function of the elites (12). In case 1, the value from being truthful is so high that the masses never self-deny, independently on the behavior of the elites. In case 2, as $v < v_T + \left(\overline{k} - k_R\right) - \alpha$ is equivalent to $v + \alpha - \left(\overline{k} - k_R\right) < v_T$, the most heterodox mass member does not self-deny when the elite is honest even if everybody else self-denies. Finally, in case 3, the most heterodox self-denies if the elite is sufficiently honest. ■

B.4 High-stakes political correctness ($\alpha > \overline{k} - k_R$)

In this case, by simple algebra both $\overline{\mu} < \underline{\mu}$ and $\overline{e} < \underline{e}$ hold. $\overline{e} \equiv \min\left[e^{-1}(\mu = 1)\right] < \underline{e} \equiv \max[e^{-1}(\mu = 0)]$ implies there are values of e for which the effort of the elites is sufficiently high to generate full self-denial and at the same time sufficiently low to generate no self-denial. When this arises, the multiplicity of equilibria is due to a bandwagon effect: the higher (lower) self-denial, the higher (lower) the incentives for everybody to self-denial (be truthful). The reaction function of the masses (15) can be rewritten as

$$\mu = \begin{cases} 0 \text{ if } e < \frac{v_T}{v} \\ \frac{v_T - ve}{\alpha - \left(\overline{k} - k_R\right)} \text{ if } \frac{v_T - (\alpha - \left(\overline{k} - k_R\right))}{v} < e < \frac{v_T}{v} \\ 1 \text{ if } e > \frac{v_T}{v} + \frac{\left(\overline{k} - k_R\right) - \alpha}{v} \end{cases} \tag{25}$$

The resulting graphs when plotting possible values of $0 \leq e \leq 1$ against μ are

1. If $\overline{e} > 1$ then the only possible solution is $\mu = 0$ for all e
2. If $\overline{e} < 1$ we get $\mu = 0$ for $0 \leq e \leq \min\left[\underline{e}, 1\right]$. We get $\mu = 1$ for $\max\left[\overline{e}, 0\right] \leq e \leq 1$ and interior values (straight line from $\overline{e}$ to $\underline{e}$) for $\max\left[\overline{e}, 0\right] \leq e \leq \min\left[\underline{e}, 1\right]$.

Proposition 2 characterizes the equilibrium in this case:

Proposition 2 *For $\alpha > \overline{k} - k_R$, the bandwagon effect might lead to multiple equilibria (μ, e):*

1. If $v_T - v > \alpha - (\overline{k} - k_R)$, $(\min\left[\max[0, \widetilde{e}], 1\right], 0)$ is the unique equilibrium, with $\widetilde{e} < 1 \Longleftrightarrow \beta > \underline{b} + c$ and $\widetilde{e} > 0$ for $\beta < \overline{b} + c$
2. If $v_T - v < \alpha - (\overline{k} - k_R)$, we need to distinguish four subcases:
   (a) If $0 < \overline{e} < \underline{e} < 1$, then (i) If both $\widehat{e} > \overline{e}$ and $\widetilde{e} < \underline{e}$ hold, (e_{int}, μ_{int}), $(\widehat{e}, 1)$ and $(\widetilde{e}, 0)$ are all equilibria, with multiplicity arising only for $\overline{b} > \beta > \underline{b} + c$ and the interior equilibrium being unstable. In the rest of the subcases, the equilibrium is always unique and given by (ii) (e_{int}, μ_{int}) if $\widehat{e} < \overline{e}$ and $\widetilde{e} > \underline{e}$. (iii) $(\max[\widehat{e}, 1], 1)$ if $\widehat{e} > \overline{e}$ and $\widetilde{e} > \underline{e}$. (iv) $(\min[0, \widetilde{e}], 0)$ if $\widehat{e} < \overline{e}$ and $\widetilde{e} < \underline{e}$.
   (b) If $\overline{e} < 0 < \underline{e} < 1$, $(\max\left[0, \min[\widehat{e}, 1]\right], 1)$ is always an equilibrium. Additionally: (i) if $\widetilde{\mu} < \underline{\mu}$ and $\underline{e} < \widetilde{e}$, $(0, \underline{\mu})$ and (e_{int}, μ_{int}) are also equilibria, and (e_{int}, μ_{int}) is stable (ii) if $\widetilde{\mu} < \underline{\mu}$ and $\underline{e} > \widetilde{e}$, $(0, \underline{\mu})$ and $(\max[0, \widetilde{e}], 0)$ are also equilibria. (iii) if $\widetilde{\mu} > \underline{\mu}$ and $\underline{e} > \widetilde{e}$, $(\max[0, \widetilde{e}], 0)$ and (e_{int}, μ_{int}) are also equilibria, and (e_{int}, μ_{int}) is unstable (iv) If $\widetilde{\mu} > \underline{\mu}$ and $\underline{e} < \widetilde{e}$ no further equilibrium exists.

(c) If $\overline{e} < 0 < 1 < \underline{e}$, $(\max[0, \min[\widehat{e}, 1]], 1)$ and $(\min[0, \max[\widetilde{e}, 1]], 0)$ are always equilibria. In addition, (i) if $\widehat{\mu} < \overline{\mu}$ and $\widetilde{\mu} > \underline{\mu}$, (e_{int}, μ_{int}) is also a (stable) equilibrium. (ii) if $\widehat{\mu} < \overline{\mu}$ and $\widetilde{\mu} < \underline{\mu}$, $(0, \underline{\mu})$ is also an equilibrium. (iii) if $\widehat{\mu} > \overline{\mu}$ and $\widetilde{\mu} < \underline{\mu}$, $(1, \overline{\mu})$, $(0, \underline{\mu})$ and (e_{int}, μ_{int}) are also equilibria, and (e_{int}, μ_{int}) is unstable. (iv) if $\widehat{\mu} > \overline{\mu}$ and $\widetilde{\mu} > \underline{\mu}$, $(1, \overline{\mu})$ is also an equilibrium.

(d) If $0 < \overline{e} < 1 < \underline{e}$, $(\min[\max[0, \widetilde{e}], 1], 0)$ is always an equilibrium. The additional equilibria are:

- (e_{int}, μ_{int}) if $\widehat{e} > \overline{e}$ and $\widehat{\mu} < \overline{\mu}$ (it is stable) or if $\widehat{e} < \overline{e}$ and $\widehat{\mu} > \overline{\mu}$ (it is unstable)
- $(\min[\widehat{e},1], 1)$ if $\widehat{e} > \overline{e}$
- $(1, \overline{\mu})$ if $\widehat{\mu} > \overline{\mu}$

(e) if $\widehat{e} < \overline{e}$ and $\widehat{\mu} < \overline{\mu}$ no further equilibrium exists

Proof This is done by combining (25) with the reaction function of the elites (12). Case 1 corresponds to a situation in which the most heterodox person does not self-deny, while in instead in Case 2 a sufficiently high e induces the most heterodox person to self-deny. The stability properties in the case of multiple equilibria are easily seen by drawing the reaction functions of the elite and masses in the $\mu - e$ space. If the slope of the interior reaction function for the heterodox masses is flatter than for the elite, the interior equilibrium is stable. Otherwise it is unstable. ■

B.5 Full characterization of the equilibrium

Proposition 3 fully characterizes the equilibrium in the space (α, v):

Proposition 3 *The equilibrium is given by the following figures: (i) Figure (case1) if (ia) $c < \overline{b} - \underline{b}$ and $\underline{b} < \beta < \underline{b} + c$ or (ib) $c > \overline{b} - \underline{b}$ and $\underline{b} < \beta < \overline{b}$. (ii) Figure (case 2 low) if $c < \overline{b} - \underline{b}$ and $\underline{b} + c < \beta < \overline{b}$. (iii) Figure (case 2high) if $c > \overline{b} - \underline{b}$ and $\overline{b} < \beta < \underline{b} + c$. (iv) Figure (case 3) if (iva) $c < \overline{b} - \underline{b}$ and $\overline{b} < \beta < \overline{b} + c$ or (ivb) $c > \overline{b} - \underline{b}$ and $\underline{b} + c < \beta < \overline{b} + c$.*

Proof Let us first consider the conditions for specific equilibria to arise. $(0, \underline{\mu})$: as $\underline{\mu} = \frac{v_T}{\alpha - (\overline{k} - k_R)}$, the proportion of self-deniers is independent from v and decreasing in α. Let α^* be defined as the value of α such that $\widetilde{\mu} = \underline{\mu}$, i.e. $\alpha^* = \frac{cv_T}{\overline{b} + c - \beta} + \overline{k} - k_R$. As $\widetilde{\mu} > \underline{\mu} \Leftrightarrow \frac{\overline{b} + c - \beta}{c} > \frac{v_T}{\alpha - \left(\overline{k} - k_R\right)}$, we have that $\widetilde{\mu} > \underline{\mu} \Leftrightarrow \alpha > \alpha^*$ if $\alpha > \overline{k} - k_R$ and instead $\widetilde{\mu} > \underline{\mu} \Leftrightarrow \alpha < \alpha^*$ if $\alpha < \overline{k} - k_R$. This implies that this equilibrium can only arise for $\alpha > \overline{k} - k_R$ since otherwise $\underline{\mu} < 0$. As $\frac{\partial \alpha^*}{\partial \beta} > 0$ and $\widetilde{\mu} < \underline{\mu} \Leftrightarrow \alpha < \alpha^*$ for $\alpha > \overline{k} - k_R$, a greater β makes the range of α-values for which $\widetilde{\mu} < \underline{\mu}$ bigger, and this is the condition under which the proportion of self-deniers when the elite is fully dishonest is low enough to indeed induce full dishonesty. As $\underline{\mu} = \frac{v_T}{\alpha - (\overline{k} - k_R)} < 1$ and $\widetilde{\mu} = \frac{\overline{b} - \beta + c}{c}$, $\widetilde{\mu} < \underline{\mu}$ requires $\overline{b} < \beta$, so this equilibrium can only be found in Figure 6 and Figure 7. As we have shown that $\alpha < \alpha^*$ and $\underline{\mu} = \frac{v_T}{\alpha - (\overline{k} - k_R)} < 1 \Leftrightarrow v_T + \left(\overline{k} - k_R\right) < \alpha$,this equilibrium arises for $v_T + \left(\overline{k} - k_R\right) < \alpha < \alpha^*$.
$(1, \overline{\mu})$: as $e(\mu) = 1$ if $\mu < \frac{\underline{b} - \beta + c}{c}$, this equilibrium can only arise for $\beta < \underline{b} + c$ i.e. in

Figure 4 or in Figure 6. Consider first $\alpha < \overline{k} - k_R$. *As* $\overline{\mu} = \frac{v - v_T}{\left(\overline{k} - k_R\right) - \alpha}$, we need $v > v_T$ for $\overline{\mu} > 0$. At the same time, we need $\widehat{\mu} > \overline{\mu}$ in order to have that the proportion of self-deniers $\overline{\mu}$ is compatible with full honesty among the elite. As $\overline{\mu}$ is increasing in α and $\widehat{\mu}$ is independent from α, this happens for points located below $\widehat{\mu} = \overline{\mu}$. In addition, for $\widehat{\mu} = \overline{\mu}$, $\overline{\mu}$ coincides with μ_{int} and for $\widehat{\mu} < \overline{\mu}$, $e = 1$ is no longer sustainable and we get a fully interior equilibrium. For $\alpha > \overline{k} - k_R$, we need $v < v_T$ instead. Again the equilibrium area is limited by $v = v_T$ (but this time from above) and $\widehat{\mu} = \overline{\mu}$ but this time from below. The higher α, the bigger the range of low v-values for which $(1, \overline{\mu})$ is an equilibrium. Once $\alpha > \frac{cv_T}{\underline{b} + c - \beta} + \left(\overline{k} - k_R\right)$ this equilibrium exists for $0 < v < v_T$ but $\frac{\partial \overline{\mu}}{\partial \alpha} < 0 \Leftrightarrow \alpha > \overline{k} - k_R$ and $v < v_T$. Now $\overline{\mu}$ is decreasing in α and decreasing in v and $\overline{\mu} = 0$ when $v = v_T$.
$(\widehat{e},1)$: only occurs for $\beta < \overline{b}$ in the entire area where $\widehat{e} > \overline{e}$ i.e. when in the presence of full denial the proportion of honest elites is sufficient to generate full denial. It is independent of both α and v. However $\overline{e}$ depends on both α and v (decreasing in both) and $\overline{e} = 0$ if $\alpha > v_T + \left(\overline{k} - k_R\right)$. (At $\beta > \overline{b}$, $\widehat{e} < 0$ and the equilibrium turns into (0,1)). This equilibrium arises in Figures 4 and 5.
$(\widetilde{e}, \mathbf{0})$: occurs for $\beta > \underline{b} + c$ in the entire area where $\widetilde{e} < \underline{e}$, i.e. the level of honesty of the elite when nobody self-denies is low enough for all the heterodox masses to self-deny. It is also independent of both α and v. However, $\underline{e}$ is decreasing in v, so for v sufficiently high the condition $\widetilde{e} < \underline{e}$ is violated. $\underline{e}$ is independent of α. This equilibrium arises in Figures 5 and 7.
Multiplicity can only arise when $\alpha > \overline{k} - k_R$ and $\overline{e} < 1$ since in this setup the reaction of the masses is already multivalued (the possible bandwagon effect kicks in). The three different branches become active for some of the e-values only, one branch corresponds to $\mu = 0$, another one to $\mu = 1$ and the last one has an interior μ. To have both corner equilibria $(0, 1)$ and $(1, 0)$ is only a possibility if $\overline{b} < \beta < \underline{b} + c$ since only in this case the reaction function of the elite goes through both corners. In this case the elite responds with total honesty if $\mu \leq \widehat{\mu}$ and with total dishonesty if $\mu \geq \widetilde{\mu}$ and $0 < \widehat{\mu} < \widetilde{\mu} < 1$. Instead, $\widetilde{\mu} > 1$ when $\beta < \overline{b}$ so the $e = 0$ is no longer possible as some elite members are always honest. Similarly, $\widehat{\mu} < 0$ when $\beta > \underline{b} + c$, so $e = 1$ is no longer possible as some elite members are always dishonest.
The equilibrium configuration in each Figure is now discussed in turn. (i) Figure 4: in this case, full dishonesty is not part of the reaction function of the elite. One first relevant relation is in this case $\widehat{\mu} = \overline{\mu} \Leftrightarrow v = v_T + \frac{\underline{b} + c - \beta}{c}\left(\left(\overline{k} - k_R\right) - \alpha\right)$. For $\alpha = 0$, $v_T + \frac{\underline{b} + c - \beta}{c}\left(\overline{k} - k_R\right)$, while for $v = 0$ we have that $\alpha = \frac{cv_T}{\underline{b} + c - \beta} + \left(\overline{k} - k_R\right)$. In addition, it is easy to show that:

$$\widehat{\mu} > \overline{\mu} \Longleftrightarrow \begin{array}{l} v < v_T + \frac{\underline{b} + c - \beta}{c}\left(\left(\overline{k} - k_R\right) - \alpha\right) \text{ if } \alpha < \overline{k} - k_R \\ v > v_T + \frac{\underline{b} + c - \beta}{c}\left(\left(\overline{k} - k_R\right) - \alpha\right) \text{ if } \alpha > \overline{k} - k_R \end{array}.$$

A second relevant relation in this case is $\widehat{e} = \overline{e} \Leftrightarrow v = \left(v_T + \left(\overline{k} - k_R\right) - \alpha\right)\frac{\overline{b} - \underline{b}}{\overline{b} - \beta}$, for which it is easy to show that $\widehat{e} > \overline{e} \Leftrightarrow \begin{array}{l} v > \left(v_T + \left(\overline{k} - k_R\right) - \alpha\right)\frac{\overline{b} - \underline{b}}{\overline{b} - \beta} \text{ if } \overline{b} > \beta \\ v < \left(v_T + \left(\overline{k} - k_R\right) - \alpha\right)\frac{\overline{b} - \underline{b}}{\overline{b} - \beta} \text{ if } \overline{b} < \beta \end{array}$. The last relevant relation in this case is $\underline{e} = 1 \Leftrightarrow v = v_T$.
(ii) Figure 5: in this case, corruption is sufficiently attractive to make full dishonesty part of the reaction function of the elite but full honesty no longer belongs to the reaction function. The first relevant relation in this case is $\widetilde{e} = \underline{e} \Leftrightarrow v = v_T \frac{\overline{b} - \underline{b}}{\overline{b} - \beta + c}$where it is easy

to show that $\widetilde{e} > \underline{e} \Longleftrightarrow v > v_T \frac{\overline{b}-\underline{b}}{\overline{b}-\beta+c}$. The second relevant relation is $\widehat{e} = \overline{e}$, as in the preceding case.
(iii) Figure 6: in this case, we get both full and no dishonesty as part of the reaction function of the elite. In addition to $\underline{e} = 1$ and $\widehat{\mu} = \overline{\mu}$, $\overline{e} = 0 \Leftrightarrow \alpha = v_T + \left(\overline{k} - k_R\right)$ and $\widetilde{\mu} = \underline{\mu} \Leftrightarrow \alpha^* = \frac{cv_T}{\overline{b}+c-\beta} + \overline{k} - k_R$ are also relevant. It is easy to show that $\widetilde{\mu} > \underline{\mu} \Longleftrightarrow \frac{\overline{b}+c-\beta}{c} > \frac{v_T}{\alpha-\left(\overline{k}-k_R\right)}$ or equivalently $\alpha > \alpha^*$ if $\alpha > \overline{k} - k_R$ and $\alpha < \alpha^*$ if $\alpha < \overline{k} - k_R$.
(iv) Figure 7: In this subcase, corruption is again sufficiently attractive to make full dishonesty part of the reaction function of the elite but full honesty no longer belongs to the reaction function. The relevant relations are in this case $\widetilde{e} = \underline{e}$, $\overline{e} = 0$ and $\widetilde{\mu} = \underline{\mu}$. ■

B.6 Corruption payoff and equilibria

In this appendix we explain how the different cases are linked, i.e. how the equilibria change with the payoff from corruption (β). The exact cutoff levels are presented in Table A1. For low β, we are always in Case 1, then for intermediate values we move to case 2 (Case 2high or Case 2low depending on the variable cost of corruption) and finally for large enough β,we are in case 3. Although at a first sight the different cases seem to lead to different equilibria, we now show that the change in equilibria is actually very well behaved.

	$\underline{b} < \beta < \min(\overline{b}, \underline{b}+c)$	$\min(\overline{b}, \underline{b}+c) < \beta < \max(\overline{b}, \underline{b}+c)$	$\beta > \max(\overline{b}, \underline{b}+c)$
$c > \overline{b} - \underline{b}$	Case 1	Case 2 high	Case 3
$c < \overline{b} - \underline{b}$	Case 1	Case 2 low	Case 3

Table A1: Cases depending on the payoff from corruption β for high and low penalties c

Consider first a high penalty ($c > \overline{b} - \underline{b}$). When β increases to $\overline{b}$ the line that ends in $v_T + \overline{k} - k_R$ rotates till it is fully vertical when $\beta = \overline{b}$ (note its intercept is $\left(v_T + \overline{k} - k_R\right)\left(\frac{\overline{b}-\underline{b}}{\overline{b}-\beta}\right) = \frac{\left(v_T+\overline{k}-k_R\right)}{\widehat{e}}$ and $\widehat{e} = 0$ when $\beta = \overline{b}$) at which point $\alpha^* = v_T + \overline{k} - k_R$, hence the two vertical lines in Figure 6 coincide. Hence the equilibria nicely connect. Once we are in case 2-high and β grows above $\overline{b}$, a second vertical line appears at α^* where now $\alpha^* > v_T + \overline{k} - k_R$ with multiplicity of equilibria for all values of v. When β grows further and reaches $\underline{b} + c$ the line with the intercept $v_T + \left(\frac{\underline{b}-\beta+c}{c}\right)\left(\overline{k} - k_R\right) = v_T + \widehat{\mu}\left(\overline{k} - k_R\right)$ becomes v_T since $\widehat{\mu} = 0$ at this point. Also $v_T\left(\frac{\overline{b}-\underline{b}}{\overline{b}-\beta+c}\right) = \frac{v_T}{\widehat{e}} = v_T$ at this point and now becomes the relevant horizontal line. When $\beta = \underline{b} + c$ the value of $\widetilde{e} = 1$ but decreases when $\beta > \underline{b} + c$.

With a low penalty, . When $\beta = \underline{b} + c_{low}$ the intercept $v_T + \left(\frac{\underline{b}-\beta+c}{c}\right)\left(\overline{k} - k_R\right) = v_T = v_T\left(\frac{\overline{b}-\underline{b}}{\overline{b}-\beta+c}\right) = \frac{v_T}{\widehat{e}}$ which now becomes the relevant horizontal line and $\widetilde{e} = 1$ but decreases when $\beta > \underline{b} + c$ (transformation of M' equilibria of Case 1). Now, when β grows to $\beta = \overline{b}$ the line that ends in $v_T + \overline{k} - k_R$ rotates till it is fully vertical when $\beta = \overline{b}$ at which point $\alpha^* = v_T + \overline{k} - k_R$ hence the two vertical lines in Figure 7 coincide and $\widehat{e} = 0$ at this point. When β grows further a second vertical line appears for α^* since now $\alpha^* > v_T + \overline{k} - k_R$ and an area with multiple equilibria between these two lines appears.

B.7 Increase in the orthodoxy range: additional cases

For high corruption benefits for the elite (Figure A1)[16], when full self-denial becomes an equilibrium, it is not unique and the level of truthfulness in preference revelation is decreasing in one and increasing in another of the remaining equilibria.

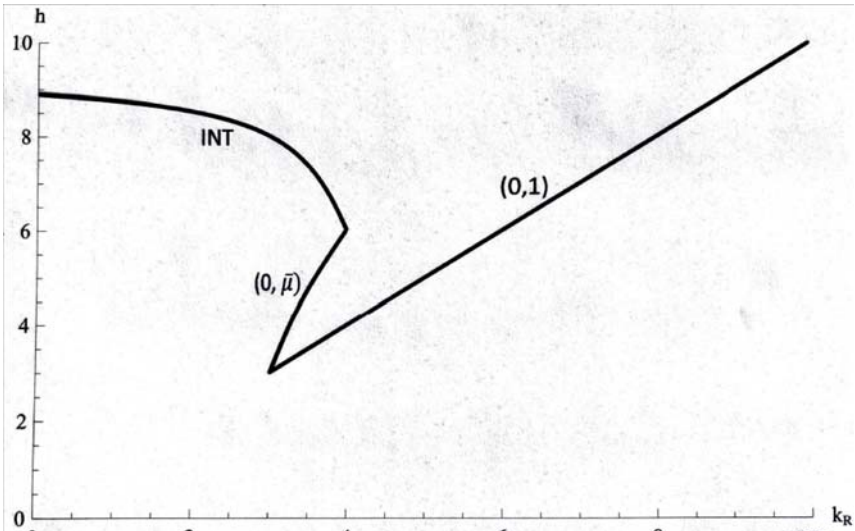


Figure A1: Number of Individuals expressing their views depending on Overton window size, high β and high v

An example of what happens in societies that just fall short of being high v is given by Figure A2.[17] Results are similar as with no elites but additional equilibria occur for intermediate sizes of the Overton window where either 2 or 4 extra equilibria occur: in half of them the degree of truthfully expressing one's own view increases with Overton window while in the other half it falls. This is another new result due to the existence of elites.

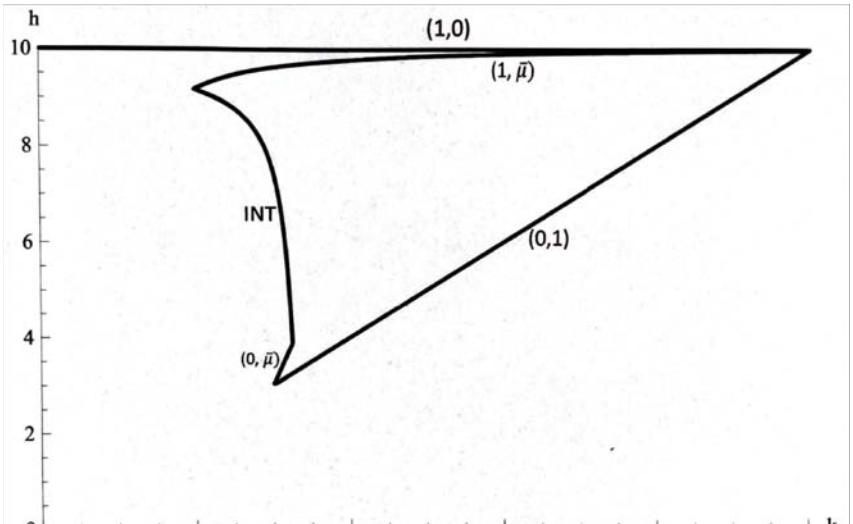


Figure A2: Number of individuals expressing their views depending on Overton window, intermediate β and v

[16] Specifically, we set $v = 5$, $\overline{k} = 10$, $\alpha = 9$,
$\underline{b} = 2$, $\overline{b} = 6$, $\beta = 7$, $c = 3$, and $v_T = 2$ in this example.

[17] The parameters are the same as in Figure A1 except for $v = 1.9$.

References

[1] Asch, Solomon E. (1955). "Opinions and social pressure". *Scientific American* vol 193 No 5, pp. 31-35.

[2] Bernheim, B. Douglas. (1994). "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841-877.

[3] Bourdieu, Pierre (1979) "Symbolic Power". *Critique of Anthropology*, 4, pp. 77-85.

[4] Bourdieu, Pierre and Luc Boltanski (1976) "La production de l'idéologie dominante". *Actes de la Recherche en Sciences Sociales*, 2-3, pp. 3-73.

[5] Braghieri, Luca (2021) "Political Correctness, Social Image, and Information Transmission". Mimeo.

[6] Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin (2020). "From extreme to mainstream: The erosion of social norms" *American Economic Review*, vol. 110 No. 11; pp. 3522-48.

[7] Clarke, John and Janet Newman (2017) ""People in this country have had of experts': Brexit and the paradoxes of populism", *Critical Policy Studies*, 11:1, pp. 101-16.

[8] Conway, Lucian G., Meredith A. Repkea, Shannon C. Houck (2017) "Donald Trump as a Cultural Revolt Against Perceived Communication Restriction: Priming Political Correctness Norms Causes More Trump Support", *Journal of Social and Political Psychology* Vol. 5(1), 244–259

[9] Domínguez, Juan F., Sreyneth A Taing and Pascal Molenberghs (2016) "Why Do Some Find it Hard to Disagree? An fMRI Study". *Frontiers in Human Neuroscience*, Volume 9, Article 718

[10] Duffy John and Jonathan Lafky (2021) "Social Conformity under Evolving Private Preferences". *Games and Economic Behavior* 128 (2021), pp. 104-124.

[11] Fatas, Enrique, Shaun P. Hargreaves Heap and David Rojo Arjona (2018) "Preference conformism: An experiment", *European Economic Review* 105, 71–82.

[12] Funke, Patricia (2016) "How Accurate Are Surveyed Preferences for Public Policies? Evidence from a Unique Institutional Setup", *Review of Economics and Statistics* Volume 98 Issue 3 p.442-454

[13] Golman,Russell, George Loewenstein, Karl Ove Moene and Luca Zarri (2016) "The Preference for Belief Consonance" *The Journal of Economic Perspectives*, Vol. 30, No. 3 (Summer 2016), pp. 165-187

[14] Hopkin, Jonathan (2020) *Anti-System Politics*: *the Crisis of Market Liberalism in Rich Democracies.* Oxford University Press.

[15] Hunter, James Davison and Carl Desportes Bowman (2016) "The 2016 Survey of American Political Culture, Initial Report of Findings", The Vanishing Center of American Democracy, Institute for Advanced Studies in Culture, University of Virginia

[16] Kuran, Timur (1987) "Preference Falsification, Policy Continuity and Collective Conservatism", *The Economic Journal*, Vol. 97, No. 387 , pp. 642-665

[17] Loury, Glenn C. (1994) "Selfcensorship in public discourse: a theory of "political correctness" and related phenomena." *Rationality and Society* 6, no. 4, 428-461

[18] Michaeli, Moti, and Daniel Spiro (2015) "Norm Conformity Across Societies." *Journal of Public Economics*, 132, 51-65.

[19] Morris, Stephen (2001) "Political correctness". *Journal of Political Economy* 109(2), 231–265.

[20] Noelle-Neumann Elisabeth (1974) "The spiral of silence: a theory of public opinion". *Journal of Communication* 24(2):43-51

[21] Norris, Pippa and Ronald Inglehart (2019) *Cultural Backlash. Trump, Brexit, and Authoritarian Populism.* Cambridge University Press

[22] Prentice, Deborah A. and Dale T. Miller (1993) "Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm". *Journal of Personality and Social Psychology* 64(2), 243.

[23] Suresh. Shyam Gouri and Scott Jeffrey (2017) "The Consequences of Social Pressures on Partisan Opinion Dynamics" *Eastern Economic Journal*, March 2017, Volume 43, Issue 2, pp 242–259

[24] Susen, Simon (2013) "Bourdieusian Reflections on Language:Unavoidable Conditions of the Real Speech Situation" *Social Epistemology* vol. 27, Nos. 3-4, pp. 199-246.