Gårn Hansen, Lars

Working Paper

The power of woke and other forms of disproportionate punishment

IFRO Working Paper, No. 2023/01

Provided in Cooperation with:
Department of Food and Resource Economics (IFRO), University of Copenhagen

Suggested Citation: Gårn Hansen, Lars (2023) : The power of woke and other forms of disproportionate punishment, IFRO Working Paper, No. 2023/01, University of Copenhagen, Department of Food and Resource Economics (IFRO), Copenhagen

This Version is available at:
https://hdl.handle.net/10419/269211

DEPARTMENT OF FOOD AND RESOURCE ECONOMICS
UNIVERSITY OF COPENHAGEN

IFRO Working Paper

The Power of Woke and Other Forms of Disproportionate Punishment

Lars Gårn Hansen

2023 / 01

IFRO Working Paper 2023 / 01

The Power of Woke and Other Forms of Disproportionate Punishment

Authors: Lars Gårn Hansen

JEL-classification: D0, D7, H0, P4, L51

Published: January 2023

See the full IFRO Working Paper series here:
https://ifro.ku.dk/english/publications/ifro_series/working_papers/

Department of Food and Resource Economics (IFRO)
University of Copenhagen
Rolighedsvej 23
DK 1958 Frederiksberg DENMARK
https://ifro.ku.dk/english/

The Power of Woke
and Other Forms of Disproportionate Punishment

Lars Gårn Hansen

January 2023

Department of Food and Resource Economics (IFRO),
University of Copenhagen

Rolighedsvej 25,
DK-1958 Frederiksberg C,
Denmark
E-mail: lgh@ifro.ku.dk

Keywords: woke activism, private enforcement, lobbying

JEL codes: D0 D7 H0 P4 L51

Abstract: I suggest that simple enforcement and lobbying models may explain how a small minority of dedicated believers (religious fundamentalists, animal rights activists, woke activists, etc.) are able to impose changes in behavior on the majority in a society who do not believe. With this, a society typically has two stable states: one with and one without the majority changing behavior. I show how activists can facilitate transition to their preferred state by focusing punishment on subsets of the behavior they want to change and on subgroups of the majority one at a time, as well as by exploiting inherent advantages they have in lobbying the leadership of subgroups with power hierarchies (corporations, universities, organizations, etc.). The willingness of dedicated believers to inflict highly disproportionate punishment on members of the majority turns out to be critical for their ability to facilitate transition. I show that transition to the state in which the majority changes behavior may substantially reduce social welfare. I conclude with a discussion of strategies for avoiding transition, which the majority may consider.

1. Introduction

In 2005, Muslim fundamentalists threatened a Danish newspaper and its illustrators for publishing a number of cartoon drawings of the Prophet Muhammad. After a number of unsuccessful attacks in Denmark, a devastating attack on the offices of the French magazine, Charlie Hebdo, in Paris in 2015 succeeded in killing a number of employees of the media outlet. At the time of the attack, Charlie Hebdo was one of the few media outlets in the world who were persisting in printing Muhammad cartoons periodically. In 2020, a French school teacher who had presented some of the Muhammad cartoons in his class was killed by Muslim activists. Today there is (virtually) no public dissemination of Muhammad cartoons through media outlets, museums, exhibitions, or other public fora in Europe. In effect, a few dedicated fundamentalists willing to kill and be killed to enforce the ban they have proclaimed on the public dissemination of Muhammad cartoons have succeeded in changing the behavior in all of Europe (and probably also in most of the rest of the world).

This is a dramatic case of a small minority inducing a change in behavior through disproportionate punishment but much more benign examples abound. In many western countries, animal rights activists have been able to effectively ban the use of fur clothes and products, woke activists have been able to ban or alter words, songs, books, illustrations, etc. Typically, a popular word or product is called out as offensive and often very quickly ends up being disused (or altered) and so effectively banned, and it seems that most of the members of the public affected by the ban only feel slightly inconvenienced by the change. However, the small group of activists successfully promoting the change feel very strongly about it, so strongly that they are willing to protest, organize boycotts, social media campaigns, etc., aimed at those using the words or products that they have called out. Activist punishment typically utilizes public shaming of violators and by association their employers and social network with the aim of getting violators fired and socially excluded. Though the methods of punishment in these cases are substantially less dramatic and typically legal, they can nevertheless be devastating for the careers and social life of violators. After a (typically short) period of time during which violators are shamed, fired or in other ways excluded, we often see that the small group of dedicated believers has been successful in changing the behavior of the majority.

What distinguishes activists in these situations is not their willingness to incur costs for a cause they believe in. Many people contribute to charities and volunteer at events whose cause they believe in. Rather, I suggest that what distinguishes them is their willingness to use these resources to *hurt* other people and to hurt them disproportionately in the eyes of the majority. Killing someone for printing a cartoon is considered extreme and disproportionate by most. Most people would also

consider firing a teacher for using a commonly used word that an activist has called out as extreme and disproportionate. Of course, the activists' perspective is different. They not only feel personally offended by the behavior they call out, they often also feel that it offends fundamental (religious, intersectional, etc.) principals that must be defended at all costs. They are defending suppressed groups and/or a true and higher cause, which makes the punishment they administer reasonable in their eyes, and the act of inflicting punishment not only necessary but also honorable. Irrespective of the reasons why, a distinguishing characteristic of these social situations is that activists are willing to inflict a degree of punishment on violators which the majority considers (highly) disproportionate and (far) outweighs the benefits derived from undertaking the behavior in question for most individual members of the majority. This turns out to be important for understanding why activists often succeed despite only being a small minority, which is the aim of my paper.

I propose a simple enforcement model which, assuming that activists are willing to inflict disproportionate punishment, emits two stable social equilibria: one with the majority changing behavior, which I call the *woke* equilibrium, and one without the majority changing behavior, which I call the *initial* equilibrium. Then I point to ways that activists with agency can facilitate a transition to the woke equilibrium. By focusing all their punishment resources on small subsets of behavior and/or small subgroups, they are able to transition these one at a time. After transitioning a subset or subgroup, the punishment resources needed to stabilize the new woke equilibrium are very limited because the punishment for violations is so disproportionate that virtually no violations occur in this equilibrium. This allows activists to refocus almost all their resources on the next subset/subgroup to be transitioned. I also propose a simple lobbying model that under plausible assumptions explains how activists can exploit power hierarchies in subgroups (universities, corporations, sports associations, etc.) by lobbying the subgroup leadership for policies that increase the effectiveness of the resources they use to punish members of the majority. Activists turn out to have an inherent advantage in this lobbying competition with the majority because their interests are more concentrated than those of the members of the majority. After showing that transition to the woke equilibria can substantially reduce social welfare (as defined in a standard CB-analysis), I conclude by discussing strategies that counter-activists from the majority might consider.

The paper (only) contributes by suggesting that simple versions of well-known regulation-enforcement and lobbying models can help us understand the basic dynamics of social situations in which a few dedicated activists succeed in imposing behavioral changes on the majority. As such, the models I suggest are well understood from the regulation-enforcement and lobbying

literature (see Becker, 1968 and 1983 for foundations, and Baron, 2016 and Hayes and Dijkstra, 2001, for my specific inspiration). Further, the social situations that I consider are (much) more complicated than allowed for by the simple models I propose. Thus, the proposed models can only (potentially) explain some of the dynamics at play. What is (potentially) explained nevertheless seems to be fundamental and important for understanding these social situations, which have become increasingly common in western societies. Though many papers suggest similar dynamic explanations for other settings (see, e.g., Baron, 2016, and Hayes and Dijkstra, 2001), I do not know of other papers that have done so for this setting.

In the next section, I propose a model of the social situation in which the majority derives utility from undertaking behavior that a minority wants to ban, and I find that it has two stable equilibria: one in which the majority has changed behavior and one in which it has not. In the following section, I consider how activists can proactively facilitate transition to the equilibrium in which majority behavior has changed by focusing their punishment on subgroups or subsets of behavior one at a time. In section 4, I consider how activists can facilitate transition of subgroups with power hierarchies by exploiting inherent advantages in lobbying subgroup leaderships for favorable policies. In section 5, I consider social welfare while section 6 concludes the paper with a discussion of majority strategies for avoiding transition. In the paper, I impose a number of simplifying assumptions to bring the intuition of the models and results to the forefront, but these can be relaxed without significantly changing the results.

2. A simple model of activist enforcement

In this section, I propose a model of a social situation in which a majority derives utility from undertaking behavior that a minority wants to ban and is willing to punish. The model applies theory from the standard economics of crime and punishment approach (originally laid out by Becker, 1968) to this setting. Baron (2016) is perhaps the paper that comes closest to the approach I apply. Baron shows how environmental activists with the ability to undertake campaigns that illuminate the environmental damage that firms cause can induce them to self-regulate in order to forestall or mitigate campaigns. Baron develops a comprehensive model with moderate and radical activists and a concerned public contributing to their campaigns that illuminate the environmental damage done by firms. Baron finds a unique equilibrium and is able to characterize it richly because of the comprehensive model. The model I develop utilizes this basic setup but is much cruder with the drawbacks this entails, e.g., encompassing only one type of activist and no explicit contributions

market. An important difference is that the model I develop nevertheless emits two stable equilibria. This is an advantage in our setting since what often seems to be happening – and needs to be explained - in the social situations I want to investigate is a shift between two apparently stable states: one without any change in behavior and one with almost universal change in behavior.

The model has activists who want to change the behavior of the majority who derive utility from undertaking this behavior.

Members of the majority

First consider members of the majority who engage in behavior chosen from a set of possible behavior alternatives, X, where x is the subset of these behaviors that activists want to ban. Assume that initially, before activists become active, there are M situations (during a given time period in the given social situation) in which members of the majority exhibit x-behavior and let $U_m(x)$ be the utility loss experienced by the member of the majority who initially exhibits x-behavior in situation m if prevented from doing so. Thus $U_m(x)$ is the utility loss experienced by the member of the majority in situation m in which a behavior from x would initially be chosen if he were forced to choose a behavior from $X \setminus x$ instead. The value of this utility loss, $U_m(x)$, differs across the M situations depending on the preferences of the specific member of the majority in question and on the situational context. As an example, X could be the set of words that teachers at a university use during class room teaching and x a subset of these words that activists wish to ban while $U_m(x)$ is the utility loss resulting from the inconvenience that the relevant teacher experiences when not allowed to use an x-word in a teaching situation in which he would otherwise do so.

In figure 1, the vertical axis measures utility loss of not undertaking x-behavior in any given situation m, and we have arranged the M situations in which x-behavior is initially exhibited out the horizontal axis ranked by falling utility loss. For any given situation, m, on the horizontal axis, the D-curve indicates the utility loss derived from not undertaking x-behavior:

$$u = D(m) = U_m(x) \tag{1}$$

The initial point (m=1) on the loss curve indicates the utility loss for the relevant member of the majority in the situation in which the greatest utility loss ($U_1(x)$) is experienced while the last point

on the curve (*m*=*M*) depicts the situation with the smallest utility loss ($U_M(x)$). For technical reasons, we assume that the function is strictly decreasing and that utility loss in situation *M* is exactly zero ($D(M) = U_M(x) = 0$).

Figure 1

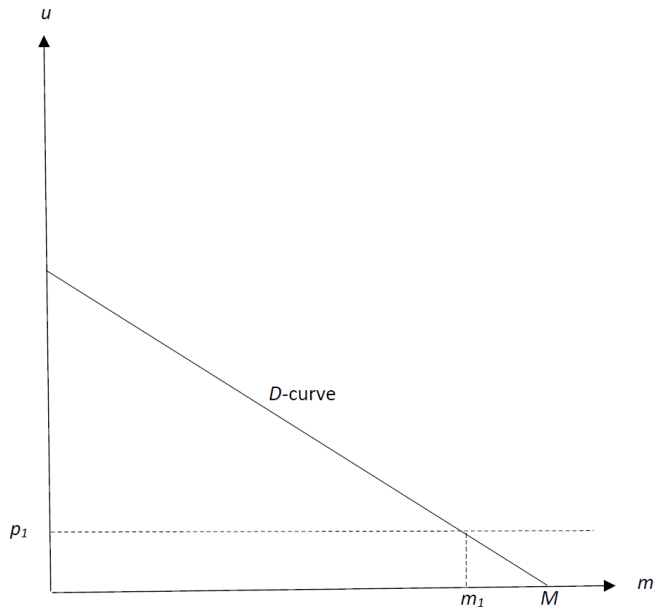
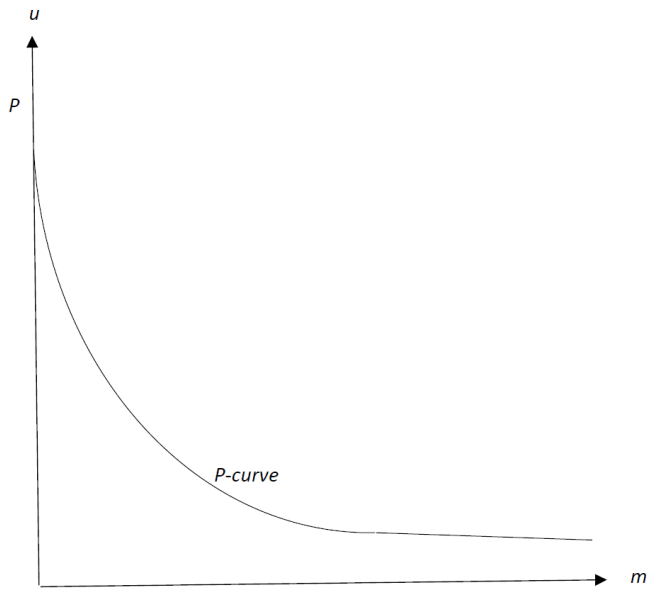


Figure 2



Now suppose that exhibiting *x*-behavior entails a risk of being punished by activists and that the utility value of the expected punishment is *p*. The majority member contemplating an *x*-behavior in situation *m* will undertake it if $U_m(x) \geq p$ and will not undertake it if $U_m(x) < p$. In either case, the majority member expects to incur a utility loss (of *p* and $U_m(x)$ respectively) but choosing the behavior to exhibit in this way minimizes the utility loss. Thus, the loss curve also indicates the maximal utility value of punishment for undertaking the preferred *x*-behavior that the relevant majority member is willing to risk before he chooses *not* to exhibit the *x*-behavior in the given situation that he would otherwise undertake. Essentially, the *D*-curve describes the majority's demand for *x*-behavior as a function of its price in the form of expected punishment inflicted by activists. In figure 1, if expected punishment is p_1, then members of the majority in situations *1* to m_1 will undertake *x*-behavior because $U_m(x) \geq p_1$, while *x*-behavior is not undertaken in the remaining situation in which $U_m(x) < p_1$ resulting in m_1 instances of *x*-behavior being exhibited.

Activists

Turning to activists, I assume they can coordinate their behavior and are willing and able to allocate resources amounting to a total of *C* to implement a ban on *x*-behavior. If all these resources are used to punish members of the majority who exhibit the *x*-behavior activists want to ban, then activists have a total capacity of *P* for inflicting utility loss on members of the majority through punishment given by:

$$P = \alpha C \tag{2}$$

where α characterizes the effectiveness of the punishment technology available to activists. In some settings, substantial punishment may be inflicted using few resources if say it is university policy that using a banned word during classroom teaching is sufficient grounds for being fired. In this case, activists can inflict substantial damage on violators just by reporting violations to the university administration, and so α is large and activists have a large punishment capacity *P* given the resources they are willing to use, *C*. In other settings, α is smaller if say a university has no such policy and activists have to disrupt classroom teaching and protest for an extended period of time before the administration caves in and the teacher is fired.

The total punishment capacity, *P*, is allocated across the situations in which *x*-behavior is exhibited. Thus, if there are *m* situations (out of the *M* possible) in which *x*-behavior is actually undertaken by a member of the majority, the expected punishment for doing so is *P/m*. In figure 2 (in which axes in figure 1 are replicated), we have introduced the *P*-curve describing the expected punishment for a violation inflicted by activists as a function of the number of violations. For any given situation, *m*, on the horizontal axis, the *P*-curve indicates the expected utility loss from punishment that activists generate assuming that *x*-behavior is undertaken in all *m* situations *1* to *m* (in which utility loss is greater than or equal to $U_m(x)$) but not in any of the situations *m+1* to *M* (in which utility loss is less than $U_m(x)$):

$$p = P(m) = P / m \tag{3}$$

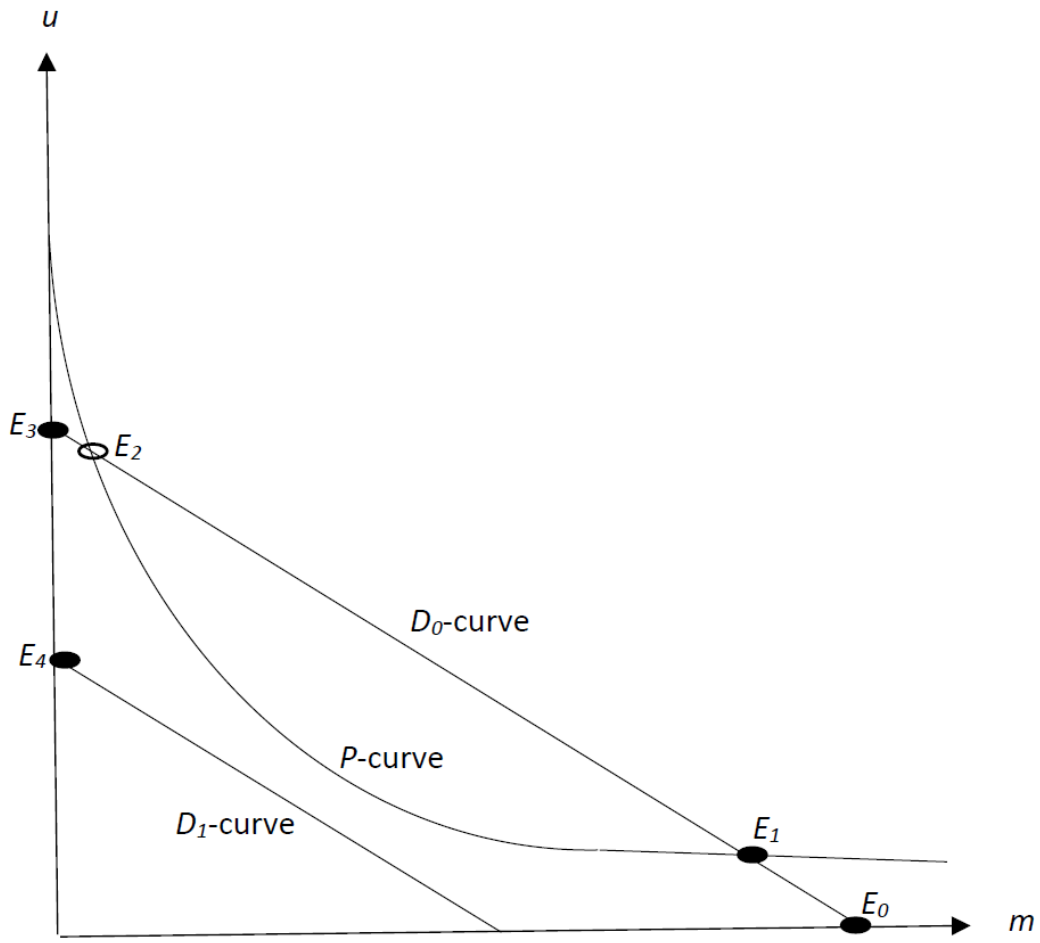
The *P*-curve describes activist 'pricing' of *x*-behavior as a function of the realized demand for *x*-behavior by members of the majority. The *P*-curve starts at the high expected utility value of punishment of *P* since with one instance of *x*-behavior, all punishment is concentrated on one violator

(i.e. $P(1) = P$) while it terminates at a substantially lower but strictly positive utility value of punishment P / M if punishment capacity must be allocated across a large number of situations in which *x*-behavior is initially undertaken (i.e. $P(M) = P / M$). Also note that $P'(m) = -P / m^2$. Thus, the slope of the *P*-curve initially is -1 (since $P'(1) / P(1) = -P / P$) with the expected punishment falling by 50% when the number of instances of punishment increases from 1 to 2. In contrast, the *P*-curve is almost horizontal at the end as long as *M* is large with a slope of $-1/M$ (since $P'(M) / P(M) = (-P / M^2) / (P / M) = -1 / M$) so that expected punishment, for example, falls by only 1% if the number of instances of punishment increases from 99 to 100. These characteristics are depicted in the *P*-curve in figure 2.

Model equilibria

Let us now consider the models equilibria and their stability. In figure 3, we depict both the *D*- and *P*-curves to investigate possible equilibria in this social situation. Since we assume that the greatest utility loss from not using *x*-behavior ($D(1) = U_1(x)$) is small compared to the total punishment capacity of activists (P), the *P*-curve is always above the *D*-curves at this initial point. This is also the case at the end, in situation *M*, since the *P*-curve is strictly positive, and the *D*-curves cross the vertical axis here attain the value of zero. Therefore, one possibility, depicted by the D_1-curve, is that the *P*-curve is above the *D*-curve everywhere. Since expected punishment for any level of *x*-demand is greater than the utility of *x*-behavior in the marginal situation, demand starting at any point on the *D*-curve is reduced until no *x*-behavior is demanded. In this case, E_4 is the only stable equilibrium where no *x*-behavior is demanded and no punishment is effectuated, but members of the majority are kept in check by the threat of the substantial expected punishment of *P* for the marginal transgression were anyone to try. In the following, we call this the woke equilibrium since attaining this equilibrium is the goal of activists.

Figure 3



The other possibility is depicted by the D_0-curve, which crosses the *P*-curve from below at E_1 and again from above at E_2. Here E_2 is an unstable equilibrium. For any *x*-demand below E_2, the effective penalty is higher than the utility loss in the marginal situation, thereby inducing reduced *x*-demand until the stable equilibrium E_3 is reached where no *x*-behavior is demanded. For any *x*-demand above E_2, the expected penalty is lower than the utility loss in the marginal situation, thereby inducing higher *x*-demand until the stable equilibrium E_1 is reached. In this case, there are two stable equilibria with active activists: a woke equilibrium, located at E_3, which like E_4 has no *x*-behavior, no punishment but a threat of substantial punishment *P* for the marginal transgression and one with lots of *x*-behavior, lots of punishment, but where the threat of punishment for the marginal transgression is small (E_1).

In equilibrium E_1, however, activists can discontinue punishment in the current social setting and induce a move to equilibrium E_0 without punishment. This would save activists the cost of inflicting punishment in the current social setting and give them the opportunity to refocus their

activism elsewhere. In fact, equilibrium E_1 is not likely to be attractive to activists compared to equilibrium E_0. Though some instances of *x*-behavior with low utility value for those engaging in the activity are avoided, the benefits activists reap from this are small compared to their goal of banning *x*-behavior altogether and likely also compared to their costs of inflicting punishment. Thus, if there is no hope of inducing a shift from E_1 to the woke equilibrium, E_3, activists may prefer to stop punishment in the current social setting and focus their activism elsewhere. Thus, the woke equilibrium, E_3, with no punishment and no *x*-behavior and, E_0, which we will call the initial equilibrium, with no punishment and lots of *x*-behavior, are the likely long-term stable equilibria in most social situations. Equilibrium E_1 may be a stable equilibrium, but activists would likely only find this attractive as a short term steppingstone for a transition to the woke equilibrium E_3.

The model seems to replicate what we see in social situations in which activists are successful at banning behavior as a transition from equilibrium E_0 or E_1 with lots of *x*-behavior to equilibrium E_3 with no *x*-behavior. But how is this done?

3. Transitioning to woke equilibria by focusing punishment

Shocks that temporarily reduce majority benefits from undertaking *x*-behavior may be sufficient to facilitate such a transition. An example of this might be the use of the confederate flag in the USA in various social situations. Use of the confederate flag had for a long time been called out and punished by activists periodically, presumably with some success, but this did not result in anything resembling a universal ban before 2015. This is illustrated in figure 4 in which majority benefits of flag use are indicated by the D_0-curve and punishment is indicated by the *P*-curve where punishment has moved flag use from equilibrium E_0 to E_1 on the D_0-curve. The confederate flag was still used in many social situations and continued to be flown and used officially by several southern US states. However, this changed after the Charleston church shooting in 2015, when a perpetrator draped in the confederate flag killed a number of churchgoers attending services. Presumably, benefits from using the flag were (dramatically) reduced because of its association with the shooting (shifting the *D*-curve down to D_0), resulting in a new equilibrium (E_4) in which use of the flag was stopped. This reduction in benefit may have been permanent, but even if it has waned over time, the equilibrium has transitioned. Anyone considering using the confederate flag assuming that benefits have increased again would no longer be one among many but would probably face the full wrath of activists almost alone. This is illustrated in figure 4. As the *D*-curve shifts back to D_0, expected punishment at *D(1)* continues to

be greater than the increasing benefit as *D(1)* shifts up from E_4 and so eventually equilibrium E_3 on the D_0-curve is reached and the transition is completed.

Figure 4

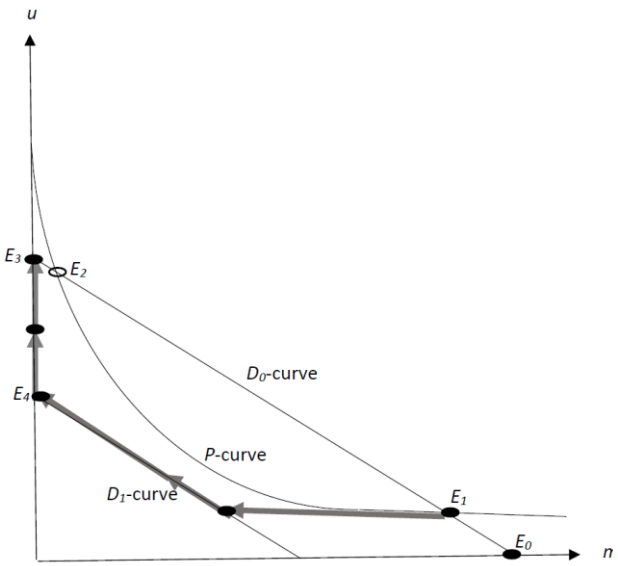
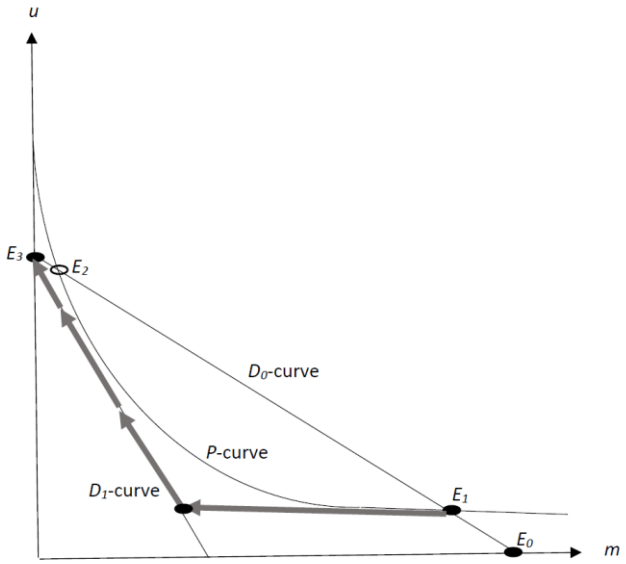


Figure 5



Thus, shocks to the *D*-curve may cause a transition. However, activists may also be able to pro-actively facilitate the transition. If activists are able to coordinate between themselves and act in coalition, this potentially gives them agency in the social situation; in effect, they could become a first mover able to predict how the majority will react and able to act strategically. Using this knowledge could make it possible for activists to proactively facilitate the transition (similar to how a Stackelberg market leader behaves).

One way activists who are able to act strategically might be able to facilitate transition is by dividing the set of *x*-behaviors that they want to ban into subsets that they are able to transition one at a time. If activists can focus all punishment on a small subset of the *x*-behaviors they want to ban, there will be fewer situations in which this behavior is initially used, and the utility cost of disuse by members of the majority will be smaller since more alternatives will be available. Focusing their total punishment, *P*, on a smaller set of behaviors essentially causes a downward shift of the *D*-curve that activists are up against in the same way as illustrated in figure 4. This makes it possible to transition this smaller subset of *x*-behavior to the woke equilibrium with no *x*-behavior and, therefore, no actual punishment needed to enforce the ban. Once the new equilibrium has been implemented for

this smaller subset of *x*-behaviors (i.e., the new equilibrium has been stabilized with all members of the majority having updated their priors on expected punishment to *P* if they undertake an instance of this subset of *x*-behaviors), activists no longer expend any punishment resources because no member of the majority finds it attractive to undertake this behavior when expected punishment is *P*. Then activists can refocus their resources on another subset of *x*-behaviors that can be transitioned in the same way. Using this piecemeal approach, all of the original set of *x*-behaviors can ultimately be transitioned and the E_3 equilibrium implemented for all of the original set of *x*-behaviors.

In a similar way, activists can focus punishment by dividing the majority into subgroups which are transitioned one at a time. If activists can focus enforcement on a subgroup, there will be fewer situations in which *x*-behavior is initially undertaken, which causes a shift to the right of the *D*-curve that activists are up against, as illustrated in figure 5. This again can make it possible to transition one subgroup at a time to the woke equilibrium, where no *x*-behavior and no actual punishment is undertaken freeing up punishment resources for transitioning the next subgroup. Ultimately, all members of the majority have once again been transitioned and the woke E_3 equilibrium implemented for all members of the majority.

Looking more closely at the banning of illustrations and words we have experienced in recent years, there does seem to be this kind of progression. Trademarks using illustrations of minorities have been called out and effectively banned in succession. Specific words such as 'Negro' have been called out and initially disused in media and public discourse and later in current literature. Now even past literature, using the word is being banned or it is changed to avoid being banned. What activists exploit is that the woke equilibrium, once transition is completed, is effectively enforced by the threat of massive punishment but that projecting this threat does not entail actually having to inflict punishment as long as punishment is sufficiently disproportionate so as to make any instance of *x*-behavior unattractive. This is where the activists' willingness to inflict what in the eyes of the majority is extreme and disproportionate punishment is critical. Being able to credibly threaten punishment that is disproportionate in comparison to the utility gain from transgressing ensures that the behavior is effectively banned and does not occur. Activists do not need to expend resources on punishing actual violations (or they only have to do this once in a while to bolster the credibility of their threat). This allows the activists to refocus (most of) their resources on new subsets of behavior for which actual punishment is needed in order to induce members of the majority to update their priors on expected punishment.

4. Transitioning to woke equilibria by exploiting power hierarchies

Another way that activists can facilitate transition is by exploiting power hierarchies within subgroups. Corporations, universities, sports clubs, etc., are groups of people with distinct hierarchies in which a leadership may have some power to impose policy constraining the behavior of group members (employees, students and faculty, club members, etc.). When such hierarchies exist in subgroups, it becomes possible for group members to lobby the leadership for favorable policies, especially if these policies do not conflict with the goals and purpose of the subgroup's leadership. If, for example, a corporation's main focus is profits and market shares, then its leadership may have little vested interest in policing certain forms of speech in the workplace. In this case, they may be open to lobbying efforts for such policies from activist employees. Examples of such policy changes being implemented by leaderships of universities, corporations, associations, etc., abound. Of course, members of the majority also have an incentive to lobby against such policy changes that will affect them negatively. So why do activists often seem to be winning this lobbying contest? To investigate this, I suggest a simple model of lobbying incorporating a basic insight from the literature on interest groups and lobbying (see Becker, 1983, for an original contribution) which suggests that concentrated interests tend to be more successful. My specific inspiration is Hayes and Dijkstra (2001), but the basic insight is reflected in much of the lobbying literature (see Oates and Porteny, 2003, for an older overview and Catola and D'Allessandro, 2020, for a recent example).

We consider an organization of people with a power hierarchy and a leadership who, to some extent, can police its members' behavior. Assuming that the main goals of the organization and its leadership are orthogonal to those of activists, there may initially be no policies addressing the *x*-behavior that activists want to ban. In this case, the activist's opportunities to punish other members of the organization who undertake *x*-behavior (i.e., the initial size of the punishment effectiveness parameter α in equation (2), which we denote α_0) may be determined by their position in the power hierarchy, other organizational policies and the structure of activities undertaken in the organization. For example, a university may have a policy of academic and teaching methodology freedom but not an explicit policy on promoting a safe and inclusive classroom environment for students. In this case, an activist calling out the use of a word in classroom teaching might find little help from the university administration, who may just ask him to take it up with the teacher. The activist has to confront the teacher directly, convince fellow students and perhaps organize protests implying a relatively small α_0. If, however, the university leadership were to enact a policy of ensuring a safe and inclusive

teaching environment for all, then it might be possible for the activist to anonymously initiate an investigation of the teacher by the administration simply by reporting the incident. Enactment of a safe and inclusive teaching policy could, therefore, substantially increase the punishment effectiveness parameter α above its original value of α_0. Thus, using some of their resources to lobby the organization's leadership for policy changes that increase the effectiveness of their punishment efforts within the organization may be an attractive option for activists.

Let us, therefore, assume that activists can allocate any part of their total resources, *C*, to lobbying for policies that increase α and let *c* be the number of resources they decide to allocate to lobbying. Further, let *k* be the total resources allocated by members of the majority to lobbying against these policies and let:

$$\alpha = \alpha_0 \frac{1+c}{1+k} \tag{4}$$

define the resulting α as a function of the resources allocated to lobbying for and against policies that increase α. The functional form is chosen for simplicity, but it allows the situation where no lobbying occurs in which case $\alpha = \alpha_0$ and generates the effects of lobbying one would expect: an increase in α if activists undertake more lobbying than members of the majority (i.e. $\alpha > \alpha_0$ if $\frac{c}{k} > 1$) and a decrease in α if members of the majority undertake more lobbying than activists (i.e. $\alpha < \alpha_0$ if $\frac{c}{k} < 1$). Activists want to expend the amount of lobbying effort, *c*, that maximizes total punishment, which by (2) becomes:

$$P = \alpha\left[C - c\right] \tag{5}$$

and so, they solve the following problem:

$$\underset{c}{Max}\, P = \alpha\left[C - c\right] \tag{6}$$

Inserting (4) into (6), the first order condition for a maximum is[1]:

$$\alpha_0 \frac{C-c}{1+k} = \alpha \tag{7}$$

Condition (7) ensures that the activist lobbying effort that maximizes punishment conditional on the given majority total lobbying effort *k.* Thus, in any state where condition (7) is satisfied, activists have no incentive to change their lobbying effort if they do not expect members of the majority to react to such a change (i.e., if activists assume that $\frac{dk}{dc} = 0$).

Members of the majority have an incentive to lobby their organization's leadership against policy changes that increase α since they will increase the punishment they can expect to suffer at the hands of activists. Assuming that members of the majority maximize their own utility, member *i* wants to minimize the total utility loss he incurs from punishment and lobbying costs. If there are *N* members of the majority, and we for simplicity assume that they all have the same number of violations of the behavior activists punish, then the punishment expected by any member of the majority *i* is *P/N*. Given this, member of the majority *i* sets k_i (where $k = \sum_{j=1}^{N} k_j$) to minimize $P / N + k_i$, i.e., he solves the following problem:

$$\underset{k_i}{Min}\ P / N + k_i \tag{8}$$

Inserting (5) and (4) into (8), the first order condition for a minimum is[2]:

$$-\alpha_0 \frac{1+c}{(1+k)^2}\left[C-c\right] / N + 1 = 0 \tag{9}$$

Using (4) this reduces to:

$$\alpha \frac{C-c}{1+k} = N \tag{10}$$

---

[1] Note that $\alpha'' = 0$ ensures that the second order condition for a maximum is satisfied since $\frac{d^2P}{d^2c} = -\frac{\alpha'}{k} < 0$.

[2] Note that $\alpha'' = 0$ ensures that the second order condition for a minimum is satisfied since $\frac{d^2W}{d^2k} = \alpha' \frac{c}{k^2} / N > 0$.

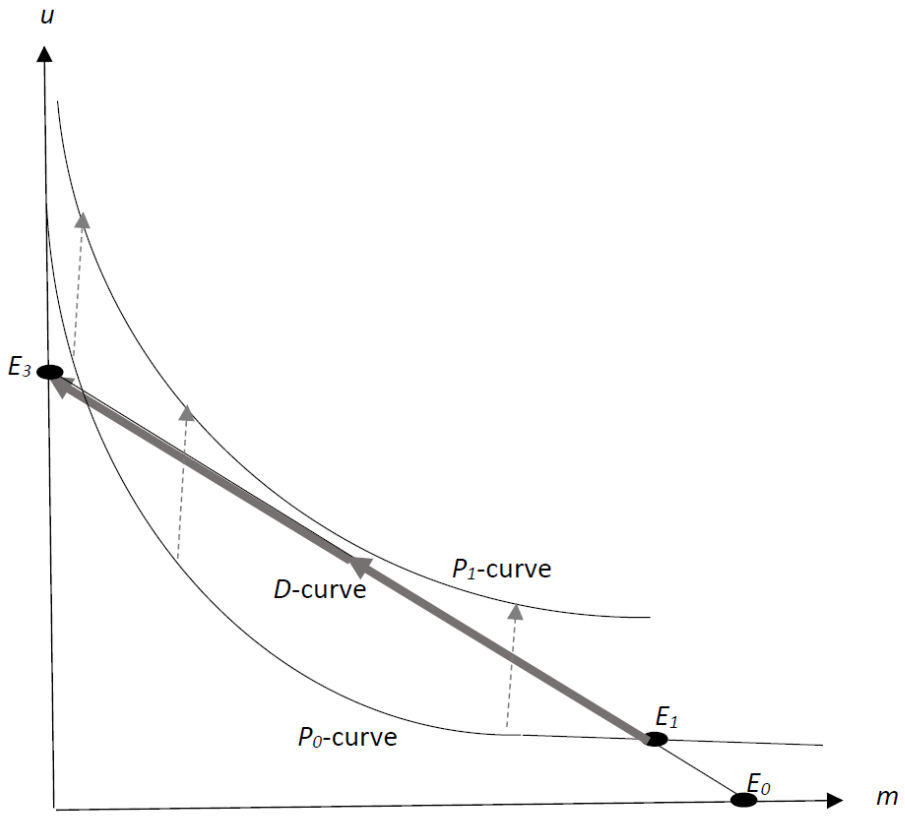
Condition (10) ensures that the lobbying effort of member of the majority *i* minimizes his utility loss conditional on the activist lobbying effort *c* and the lobbying effort of other members of the majority. Thus, in any state in which condition (10) is satisfied, members of the majority have no incentive to change their lobbying effort if they assume that $\frac{dc}{dk_i} = 0$ and that $\frac{dk_j}{dk_i} = 0$ for all other members of the majority.

Conditions (7) and (10) define a Nash equilibrium of the lobbying game we have specified. Inserting $\frac{C-c}{1+k} = \frac{\alpha}{\alpha_0}$ (implied by (7)) into (10) and rearranging, we get the following characterization of the Nash equilibrium:

$$\alpha = \sqrt{\alpha_0 N} \tag{11}$$

Thus, the resulting punishment effectiveness parameter α increases with the number of members of the majority. All other things equal, the larger the majority is, the worse it fares in the lobbying contest with activists. The model illustrates the implications of lobbing being a club good enjoyed by all on the same side of the issue. Thus, when a member of the majority lobbies, he not only reduces his own expected punishment but also the expected punishment of all other members of the majority. However, if, as we have assumed, members of the majority only consider their own benefits from lobbying and disregard those bestowed on other members of the majority when deciding their lobbying effort, they end up freeriding on the lobbying effort of other like-minded individuals. This freeriding problem results in members of the majority systematically lobbying less than their collective benefits from lobbying would imply, and this distortion grows with the number of like-minded individuals on the same side. This gives dedicated activists each with a substantial interest at stake an inherent advantage when competing in a lobbying game with a large majority each of which have little at stake. An old quote from Buckanan and Tullock (1975) which encapsulates this core result from the lobbying literature seems well suited for the social setting we are considering: Though unconventional 'this political choice setting is, however, the familiar one in which a small, concentrated, identifiable, and intensely interested pressure group may exert more influence on political choice making than the much larger majority of persons, each of whom might expect to secure benefits in the second order of smalls'.

Figure 6



The upward shift in the *P*-curve (from the original P_0-curve to the new P_1-curve) resulting from lobbying is illustrated in figure 6. The dynamics imply that once the woke equilibrium (E_3) is achieved, reversing policy causing the *P*-curves to shift back again will not cause a shift back to equilibrium E_1. Thus, just one policy shift that increases α sufficiently may be enough to transition the organization's equilibrium even if it is short lived. Policies returning to normal would not induce a transition back to the original equilibrium. For this to happen, policies would need to go further and reduce α substantially below the original level α_0.

This seems consistent with what we see in many universities, corporations and organizations. At first glance, it seems an alliance has been forged between the leadership and activists for policies that increase the effectiveness of punishment. This may, of course, reflect the actual convictions of the leadership, but as we have seen above, this could also just be because of the well understood asymmetry in lobbying efforts underlying the leadership's policy choices. When activists are (highly) successful at lobbying because of their concentrated interests, leadership policy will reflect activists' preferences to a large degree.

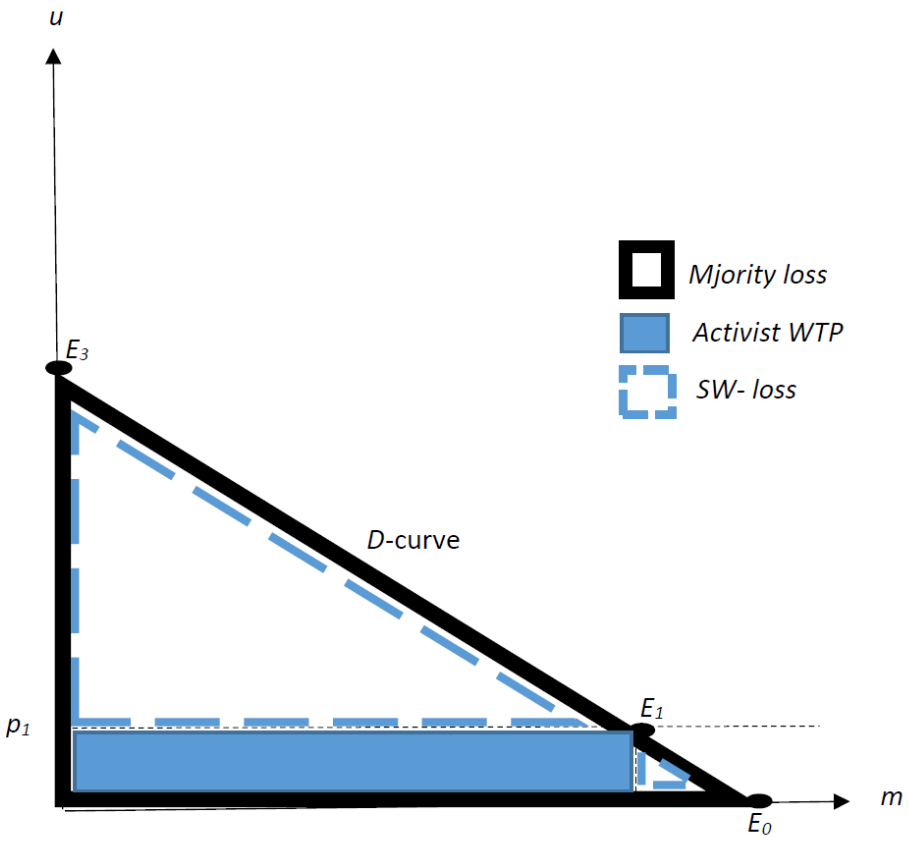
5. Social welfare

Compared to the initial equilibrium E_0 in figure 3, the woke equilibrium E_3 is beneficial to activists who succeed in implementing their ban. The majority, however, are less well-off incurring the cost of abiding by the ban on the behavior. Aa shift from the initial to the woke equilibrium does not result in a pareto improvement and so would strictly speaking require a value judgement allowing interpersonal utility comparison of winners and losers. However, traditional cost-benefit analysis (see, e.g., Mishan and Quah, 2020) suggests that projects with winners and losers can be evaluated by their potential Pareto improvement. Following this widely applied tradition, a project implies a potential Pareto improvement on the current state of the economy if the winners *in theory* could fully compensate the losers and still find the project attractive, and the social welfare gain of the project can be estimated by the winners' benefits minus the cost of fully compensating the losers.

In our case, the question to ask is whether activists would find the ban attractive if they had to fully compensate the majority for the costs inflicted on them by a ban on *x*-behavior?

Figure 7 replicates the *D*-curve in figure 3 and the two long term stable equilibria. The initial equilibrium E_0 with lots of *x*-behavior and no threat of activist punishment, and the woke equilibrium E_3, with no *x*-behavior, no actual punishment but a substantial punishment threat, as well as the short-term stable equilibrium E_1, with a small threat of activist punishment resulting in lots of *x*-behavior and actual punishment. The horizontal line from E_1 is expected punishment for violations incurred by members of the majority in E_1, where the area under the line: p_1E_1 is total incurred punishment, P, in equilibrium E_1. The majority's loss when moving from equilibrium E_0 without activists to the woke equilibrium E_3 is the area under the D-curve indicated as 'majority loss' in the figure.

Figure 7



Turning to activists, they are clearly willing to incur a cost of *C* for the time period in question in order to implement a ban on *x*-behavior and, apparently, they are not willing to incur greater costs to achieve this. However, their willingness to do so is likely not indefinite since they realize that once the ban has become effective, their costs of enforcement will be reduced substantially. Thus, if we assume that activists' willingness to pay for a ban on *x*-behavior is *C* per time period indefinitely this is very likely an (substantial) overestimate. Further, it is likely that the utility loss from receiving punishment is (substantially) greater than the cost to activists of inflicting punishment, so that *C* <*P*. Based on this alone, activists' punishment capacity *P* is likely to be a substantial overestimate of activists' permanent willingness to pay for an *x*-ban.

However, the allocation of *C* to the cause of banning *x*-behavior is the result of (costly) collective action designed to overcome freeriding among activists and their supporters and, therefore, *C* is likely an underestimate of the sum of activists and their supporters' individual willingness to pay for an *x*-ban. All in all, *P* may, therefore, be both an overestimate and an underestimate of the activists and their supporters' permanent willingness to pay for an *x*-ban.

If we, nevertheless, take *P* as an estimate of total willingness to pay for an *x*-ban in the current social setting, we see in figure 7 that this is (substantially) smaller than the total loss incurred by the majority. A classical cost-benefit analysis based on this WTP estimate would, therefore, recommend against implementing the woke equilibrium and conclude that it would generate a reduction in social welfare equal to the two indicated triangles in figure 7 labelled social welfare loss (the area under the *D*-curve minus P).

In conclusion, we cannot rule out the possibility that a shift to the woke equilibrium results in a social welfare increase (if *P* substantially underestimates activists and their supporters' WTP for an *x*-ban). However, it is also possible and perhaps more likely that such a shift will reduce social welfare (if *P* overestimates or only slightly underestimates activists and their supporters' WTP for an *x*-ban).

6. Conclusion

In this paper, I have proposed simple models of enforcement and lobbying that may explain how a small minority of dedicated believers are able to impose changes in behavior on the majority in a society who do not believe. Under plausible assumptions, the enforcement model emits two stable equilibria: an initial equilibrium in which activists are not active, and a woke equilibrium in which the majority has changed behavior. I show how activists can facilitate transition to their preferred equilibrium by focusing punishment on subsets of behavior and/or subgroups one at a time. The willingness of dedicated believers to inflict highly disproportionate punishment on members of the majority is critical for their ability to facilitate transition through this piecemeal approach and for the stability of the woke equilibrium after transition.

I also show that transition can be facilitated by lobbying leaderships for policies that increase the effectiveness of activists' punishment in subgroups with power hierarchies. This is because activists have an inherent lobbying advantage in such subgroups because their interests are more concentrated than those of the majority. Again the willingness of dedicated believers to inflict highly disproportionate punishment on members of the majority ensures that once equilibrium has been transitioned, it is stable even if the organization's policies return to those before the transition. Thus, the analysis in the paper suggests that the majority's prospect of avoiding transition when facing dedicated activists is bleak, and that the prospect of transitioning back is even bleaker as long as no collective counteraction is organized.

A coalition of counter-activists willing to expend resources may be able to match the advantage that activists initially have in lobbying leaderships of organizations for favorable policies. However, if activists are first movers in the lobbying game and can initially transition to the woke equilibrium, counter-activists face a greater challenge than the one initially overcome by activists because a return to the initial equilibrium requires more than just returning to the initial policies of the organization. Thus, the most promising lobbying strategy for counter-activists is to try to organize early enough to prevent the initial round of woke policy changes from being implemented in the first place.

Another counter-activist strategy for combating punishment effectiveness is to directly help or compensate those being punished. However, if the utility value of inflicted punishment is greater than the resources activists expend, which seems likely, this is a challenging strategy requiring counter activists to allocate (substantially) more resources than do activists. At any rate, if activists are initially able to transition to the woke equilibrium, counter-activists again face a greater challenge than the one initially overcome by activists, and so the most promising strategy for counter-activists is once again to try to organize early enough to prevent the initial transition from taking place.

A third potentially promising strategy for counter activists is to degrade the punishment technology utilized by activists. Activists' punishment often relies on the public shaming of violators and by association their employers and the organizations of which they are members who in turn are induced to fire and exclude violators. When undertaking *x*-behavior is shameful in the eyes of most people, this can be a devastatingly effective punishment. However, this requires that most people agree or at least passively accept the premise of the argument that undertaking *x*-behavior is cruel and hurtful and thereby shameful. Taking and winning the debate about this premise could in theory reduce and even neutralize the effectiveness of punishment based on public shaming. Further, broad public rejection of the woke premise may open for the wider public perception that it is cruel and hurtful to fire or exclude someone for undertaking behavior that is harmless. In this environment, firms, organizations and activists that punish violators would themselves run the risk of being publicly shammed, which further reduces the effectiveness of the activists' punishment technology currently used.

The goal of this paper is to understand why (woke and other) activists have had success in changing majority behavior in numerous social situations across the western world in recent years. Given that the transition to woke equilibria could be (and likely is) welfare reducing in many cases, investigating the welfare effects of such transitions empirically seems an important area for future

research. Further, the concluding suggestions for how activists' success might be counteracted are speculative and are, therefore, also an obvious subject for future research.

Literature

Baron, D.P. (2016) 'Self-Regulation and the Market for Activism' *Journal of Economics & Management Strategy* 25 (3), pp 584–607.

Becker, G.S. (1968), 'Crime and punishment: an economic approach'
Journal of Political Economy LXXVI (1968), pp. 169-217.

Becker, G.S. (1983), 'A Theory of competition among pressure groups for political influence'
Journal of Political Economy vol. XCVM (August 1983) No. 3, pp. 371-400.

Buchanan, J.M. and Tullock, G. (1975) 'Polluters' Profits and Political Response: Direct Controls versus Taxes' *The American Economic Review* Vol. 65, No. 1 (Mar., 1975), pp. 139-147.

Catola, M., S. D'Alessandro (2020) 'Market competition, lobbying influence and environmental externalities' *European Journal of Political Economy* Vol 63.

Hayes, A.G. and B. Dijkstra (2001) ' Interest Groups and the Demand for Environmental Policy' in The International Yearbook of Environmental and Resource Economics 2001/2002. H. Folmer and T. Tietenberg (Eds.). Edward Elgar, Massachusetts, USA.

Mishan,E.J and Quah, E (2020). Cost-Benefit Analysis, 6th Edition. Imprint Routledge, London. eBook ISBN9781351029780.

Oates, W.E, P.R. Portney (2003) 'Chapter 8 - The Political Economy of Environmental Policy' in Handbook of Environmental Economics: Environmental**,** K-G. Maler, J.R. Vincent (Eds.). Elsevier.