ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
MSc Thesis Economics of Markets and Organisations

JOINING THE DARK SIDE:
WELFARE EFFECTS OF SORTING BY MORALITY IN THE LABOUR MARKET

Julian Tait
483670

Supervisor: Prof. Dr. Robert Dur
Second Assessor: Dr. Josse Delfgaauw
Date Final Version: April 8, 2022

ABSTRACT. Workers sort in the labour market by their education, interests and motivation which provides efficiency gains. Sorting by morality however, could lead to bunching of immoral types in industries at risk of imposing great externalities on society. Leaving these peoples' decisions unchecked could be detrimental for welfare. This paper develops a model in which the welfare effects of such bunching can be assessed and discusses the effectiveness of policy and regulation. It finds that minimum wages, direct limits on externalities and campaigns targeted at increasing awareness of, and interest in social consequences of firms' actions can all be effective at improving social welfare. It concludes that the responsibility to act in the interest of society lies beyond those in government.

Contents

1 Introduction

Climate change's impacts are contested, but widely supported predictions about where we are headed are severe. Extreme weather is becoming common place and we are in a defining decade in which our climate is crossing irreversible tipping points (IPCC, 2021). Though it is not just our climate we are damaging. Human rights groups have long sought to bring attention to the exploitative nature of many supply chains the west heavily relies on.[1] Accusations that tobacco producing organisations employ aggressive marketing strategies riddled with misleading information and target minors are not new (Bates and Rowell, 2004; Heath, 2016). Each of these poses a threat to the livelihood of many. Yet for each of these issues organisations are slow to change their -as of yet- successful practices and questions on their accountability are often sidestepped, and the responsibility is shifted to others.

On climate change an increasing majority think their governments should do more (Laville, 2019; Reuters, 2019), yet this begs the question on where the responsibility lies. Is it the government, the consumers or the organisations that should be first to stand by their values (or develop them to begin with)? In the lead up to COP26 the BBC World Service found that 58% of 18 countries polled expected the government to play a leadership role, up from 43% before COP21 in Paris (GlobeScan, 2021). One problem in dealing with many externalities is that impacts on society are hard to measure, go unaccounted for, and are therefore beyond the control or effective supervision of institutions with authority. Where lack of regulation is paired with enticing private benefit, organisations are free to act as they please and society relies on those in the organisations to act considerately. When

[1] Organisations have run large campaigns on issues from clothing to food. For example, 18 percent of the world's coco supplied for confectionery was found to be produced by minors (US Department of Labor, 2017)

assessing welfare, the composition of the those in the organisations becomes important. If the workforce is responsible, are jobs and industries inherently immoral? Or do they just have the potential to be bad and this potential is abused by misguided individuals in power. If so, the ultimate cause of immoral business practices may come down to the action of employees. So how do these workers make employment decisions?

For employment across the economy, an immorality premium in the labour market has been empirically documented (Frank, 1996). Immorality in those papers take a broader definition as work generally considered immoral, whereas this paper uses it more specifically to refer to actions that have negative impacts on individuals or society as a whole. All else equal, a study in Switzerland finds individuals working at organisations considered immoral are paid more than those in less immoral jobs, essentially rewarding those that care the least (Schneider et al., 2020). The authors suggest a sorting hypothesis to explain this. Sorting here occurs because 'caring' individuals must be compensated for doing bad things (as acting immorally comes at a personal cost) and are undercut on the labour market by those needing no or less compensation. The authors derive an equilibrium in which only the most immoral individuals sort into these immoral jobs. This bunching has important implications for society. By leaving organisations with immoral potential under the influence of immoral types, their careless actions risk exacerbating the already negative potential of operations. This occurs because they have little interest in finding better alternatives or improving best practices and instead risk unleashing detrimental consequences on society.

The concern is not just one of fairness. From an efficiency perspective, uncontrolled externalities pose massive financial burdens too. Canadian tax payers are already contributing 41 million USD towards permafrost-related damages which are quickly being exacerbated by global warming (Hjort et al., 2022). The authors also find that in Finland, maintenance

costs related to climate change could exceed 35 billion USD a year by 2060. Linked to the sorting by financial remuneration is the observation that less trustworthy types sort into the financial industry (Gill et al., 2020). Understanding the implications of these immoral hubs is important to help guide policy and regulation aimed at maintaining social welfare. Hart and Zingales (2017) suggest that instead of maximising market value, firms be managed to maximise the value to the shareholders at the firm. However, in the case of far reaching externalities voting of shareholders may not be an adequate prevention of damaging welfare effects.

This paper develops a theoretical explanation on sorting by morality into immoral jobs. Although the theoretical and empirical literature on sorting in various other domains[2]) has been well documented, this phenomenon in the context of immorality remains underdeveloped. This paper thus adds to the theoretical understanding of the labour markets concerning immoral organisations. It adds to existing theory by allowing individual decisions to be linked to societal welfare. Concretely, including altruism in the preference of workers allows an altruist to gain positive utility by helping others and avoiding harmful outcomes. The willingness to do so however, depends on the value they place in the cause. Their morality on a cause is judged by the extent to which an individual cares about a certain cause. Generally, this type of moral conduct and social preference are developed throughout one's life. Schneider et al. (2020) summarise this moral compass as an abstract term as the "cost of acting immorally". This paper digs deeper into social preferences and under pure altruism finds there to be both a cost and a potential benefit to individuals in working at immoral organisations.

Expanding on this literature is socially relevant because the preferences

[2]In the context of mission-oriented organisations, for example, the benefits and effects of sorting on organisations and society have been theoretically and empirically developed. (See Buurman et al. (2012), Delfgaauw and Dur (2008), Ghatak and Mueller (2011), Besley and Ghatak (2005)), Dur and Zoutenbier (2014)

of the organisations and those of society are at odds. A profit maximising organisation aims to enact personnel policies to increase the number of those acting poorly (yet profitably). However, society would like to see as much as possible of the negative impact curtailed. To help achieve this, this paper assesses three tools an institution could employ. Those discussed are: the impact of minimum wages, maximum bonuses, direct limits on externalities and providing subsidies to encourage externality reduction practices.

By allowing for heterogeneity across individuals on how much they value welfare, the immorality premium is maintained, but, in contrast to Schneider et al. (2020), the potential of a bifurcated sorting equilibrium emerges. In this, individuals with both very low and very high levels of altruism join the labour force, with a middle section refusing to work at the immoral job at all. This occurs because as previously, those with a low moral compass can be compensated enough to act poorly by the organisation. In contrast, those with high values care so much about saving society from another worker acting poorly, that they take the job even at a potential personal cost. This finding crucially hangs on the potential to act in a better way to those they are replacing and thus mediating the costs on society[3]. Those in the middle care enough to not act poorly, but not enough to give up their outside option to save society.

All policies are found to be effective at reducing the social costs. This shows that a government can and should take responsibility in dealing with externalities. Combining policies must be done with caution, as direct limitation on externalities becomes ineffective if it is paired with a minimum wage. Despite the decrease in the externality of each employee acting de-

[3]Another interpretation of moral actions is a worker working-to-give. In this setting a worker may work hard for financial reward, but channel this into a charity or another other good cause. This motivation is reported in the real world but differs from the model considered in this paper as the mitigation decision of the employee is no longer directly linked to the profitability of the firm.

structively, this is offset by moral employees leaving the organisation. They leave because each moral employee is no longer saving society as much as before the limitation.

The level of altruism within the society also plays an important role. In particular, the morality of the most altruistic and that of the marginal immoral worker employed play key roles in the ability of an organisation to entirely deter moral types. Social policies and organisational policies aimed at raising awareness of damaging externalities can have a big impact on how much bad an organisation can do, in two ways. On the one hand it can shift the distribution of workers such that more individuals care about the cause. On the other hand, it can decrease the possible gains an organisation can make by continuing to act immorally. The findings are relevant when discussing policy aimed at increasing societal welfare when there are organisations whose actions are not inline with that of society.

The remainder of this paper is structured as follows. Section 2 begins by discussing the related literature. Section 3 develops the model before analysing it and its equilibria in Section 4. Section 5 assesses potential regulatory solutions before Section 6 presents the results and discussion.

2 Related Literature

This paper is concerned with the morality of individuals and organisations they decide to work for. The morality of an action in this paper will follow the idea of virtue ethics. The morality will be judged by the extent to which it helps achieve a specified end goal of the cause, the telos. Instead of judging actions directly, it builds a framework to assess behaviour around a cause determined at the point of application to each case. Immorality from here on means actions that are harmful to the cause of interest.

Empirical studies have shown that work considered immoral is associated

with higher wages (see: Frank, 1996; Moffatt and Peters, 2004; Arunachalam and Shah, 2008). However, relying on correlational evidence, this wage premium may be explained by variations in a number of unobservables.[4] Where other studies fail to identify a mechanism for the observed wage premium, Schneider et al. (2020) build a model and find causal empirical evidence in the lab and in the Swiss labour market to support and explain their sorting hypothesis. The authors use immorality aversion -a personal disutility of acting immorally[5]- as the driving force for an immorality premium in the labour market. They assume this value to be heterogeneous across individuals and exogenously determined. The sorting occurs as individuals with lower levels of immorality aversion must be compensated less for doing immoral work, are cheaper to hire, and consequently are more desirable employees for immoral organisations. All workers are compensated to satisfy the participation constraint of the marginal worker (returning a positive utility on the outside option for those less moral than the marginal worker), while discouraging more immorality averse workers whose participation constraints are not met. Those that care least are thus rewarded for not caring. A further implication of this model is that as more people care and the marginal worker has a higher immorality aversion, the worst types benefit from this shift.

Schneider et al. (2020)'s theory uses an abstract term of immorality aversion and they do not have a framework to assess the impact of this sorting on social welfare. They assume the aversion to be exogenously determined and do not explore its origins in human preferences.

Nonetheless, workers report their social motivation for joining certain organisations, even ones with potential to be bad, or for turning to activism (for which there is little financial reward and sometimes a large personal

[4]Frank (1996) relies on students' judgement of a sector's morality for his positive correlation.

[5]Or even merely working at an organisation perceived to be immoral.

risk or cost). Schneider et al. (2020)'s model explaining an equilibrium of sorting cannot explain why those that care would also consider working for the dark side.

This paper can explain the decision of working for an organisation considered to be immoral. To do so, it adopts social preferences as discussed by Ashraf and Bandiera (2018) to assess morality in a framework of altruism. This framework also allows to go beyond the individual decision maker to the impact each decision has on societal welfare. A key divergence from Schneider et al. (2020) is the extension of altruism from impure to pure. The individuals in their model care only about *their* involvement and the consequent harm caused by their actions. In contrast, this paper develops a model in which individuals care about a cause regardless of their personal contribution to it.

This paper relates to the well-studied field of intrinsic and pro-social motivation. The public sector and mission-oriented organisations have received a lot of attention for their personnel policies aimed at benefiting from hiring workers whose interests are naturally aligned with theirs (Gregg et al., 2011; Buurman et al., 2012: Friebel et al., 2019; Prendergast, 2007). As a consequence, sorting into the mission-oriented sectors occurs. For example, in the public sector high altruism is a positive attribute which Dur and Zoutenbier (2015) show is associated with a higher likelihood of working in the public sector. Likewise, Gregg et al. (2011) find a positive correlation between British workers self-selecting into the non-profit sector and their propensity to work unpaid overtime. Workers differing on their willingness to exert positive effort for the 'good sectors' without extrinsic incentives comes as an asset to any organisation with that goal. The same altruistic preferences allow workers in this paper to derive personal gain from helpful actions and decisions. Intrinsic motivation does not however, always have to be positively aligned with the organisation's preferences. Auriol and Brilon (2014) discuss the implications of malicious employees; 'bad'

workers working at 'good' organisations. They find that in the presence of potential abuse by spiteful employees, personnel policy must adapt. Depending on how severe the impact of abuse is on society, full-deterrence or partial-deterrence should be sought after to discourage abusive types from working at a mission-oriented organisation. This leads to a trade-off between the efficiency within the organisation and the costs and benefits to society. This is especially necessary because the tools for monitoring and motivating employees are used less in the not-for-profit sector. This paper looks beyond the scenario in which the organisation seeks to keep out such workers. Instead, in this setting such actions that harm society are inline with the interests of the firm.

In contrast to the analysis of intrinsic motivation in the good sectors, less is known about the flip side; the characteristics, and motivations of workers at potentially bad organisations. The decision of an altruist to work for a bad organisation is not immediately clear. To explain why this is nonetheless plausible, a distinction must be made of the altruistic attribute. Acts of altruism can be motivated by a care for somebody else's absolute well-being (pure altruism). However, altruism can also be impure. For example, acts of charitable giving are empirically not subject to complete crowding out as the theory of pure altruism would suggest (see literature presented in Andreoni, 1990). Andreoni (1989) explains this through a warm glow that impure altruists receive from helping others. Unlike pure altruism, impure altruism is derived from a personal gain, rather than the value of another's utility for the sake of their own. Impure altruists would not consider working for a bad organisation, as their involvement is not positive and is, in fact, 'hurting' others. On top of this, such actions can reflect negatively on their image. In contrast, pure altruists could still consider working for a bad organisation if doing so means that society is better off than under the alternative.

Empirical work by Gill et al. (2020) documents a worrying selection into

the financial industry by untrustworthy types. They establish this using experimental and survey data of business and economics students and following up on their career paths 6 years later. This is related to this paper for two reasons. First, it highlights more empirical evidence of sorting by type. Second, misguided and irresponsible actions in the financial industry have already caused harm to society, and continue to have this potential. Egan et al. (2019) find that misconduct is common and is insufficiently sanctioned. Documentation of such sorting and its potential harm thus further motivates this paper's theoretical contributions.

A further stem of related literature is that concerned directly with moral actions in market settings. Particularly worrying for society is the research on the erosion of morals through markets. Falk and Szech (2013) provide experimental evidence that the context of the market stimulates previously concerned individuals to act more immorally. This is relevant to the concept of sorting by morality as the initial social preferences and intended actions of moral types may be warped over time through the market. Despite this adverse impact of markets, Bartling et al. (2015) find that there is significant persistent preference for low-social impact alternative in a laboratory market for both firms and consumers. Ockenfels et al. (2020) find further experimental evidence that a significant proportion of producers forgo their $CO2$ emissions, an action that harms their profitability. The authors further compare two policy designs and find institutional design has an impact on the voluntary morality of producers. These empirical findings of producer heterogeneity in morality spurs the interest in understanding the micro-foundations of firm preferences. This paper develops a model in which the employee heterogeneity in pro-social preferences is the starting point of firm level heterogeneity.

Finally, given the effect the population distribution of altruism has, this paper also relates to the origins of altruism and how this is influenced by policy. Ashraf and Bandiera (2018) suggest that altruism is not innate or

exogenous but following Aristotle's virtue ethics can grow and diminish over time. They call this altruistic capital and propose that it is under the influence of targeted policy. This endogenous altruism falls beyond the scope of the paper, but the factors and policy influencing the distribution of altruism among the population are important and considered in the discussion.

3 The Model

The model built in this paper reflects the interaction between individuals and organisations in the labour market, their preferences and relates the relevant decision to their societal impact. I will begin with the utility maximising individual, move on to the profit maximising organisation and finally the realised welfare of a specific cause. An important factor to be reiterated here is that of the *cause*. In the context of the telos and virtues discussed in the previous section, the cause is the telos. This is something that individuals' and organisations' actions impact either positively or negatively and can be different depending on the application of the model. For example, a desire for a flourishing whale population in the ocean, an interest in lush natural forests or the health of others. Hence, this paper will use cause to refer to any telos the model may be applied to. The morality of actions are determined by their impact on the cause.

First, there are N individuals in the population looking to join the labour force. They are modelled as rational, risk neutral and utility maximising agents. This is comprised of their personal consumption as well as a term representing an interest in the cause. Their personal utility depends on their wage and the cost of effort. The value placed on the external interest varies by individual, and it is this that is referred to as altruism.

$$U_i = w - c + \gamma_i W$$

where w is wage, c is cost of effort, γ_i is i's altruism level and W is total

welfare of their cause. The pure altruism is modelled by the cause's welfare appearing as a separate element in the utility function. The distribution of individuals' altruism is defined in the range $[0, \overline{\gamma}]$ and described by its cumulative distribution function (CDF) $F(\gamma)$. A value of 0 refers to someone who does not care about the cause at all and $\overline{\gamma}$ is the value of the individual who cares the most in the population.

3.1 Destructive Effort

A profit maximising organisation with immoral potential operates in society. It hires workers to fill its n vacancies by offering contracts comprising of a wage and a bonus. The production process offers an opportunity for private gain to the organisation at the expense of inflicting a negative externality on the social cause W. This externality is viewed relative to the domain of interest of the individual. In the examples previously mentioned this could be cutting down long established forests or hunting close to extinct whale species. The morality of the organisation is judged on its impact on the cause of interest.

I speak of immoral *potential* , because the externality I_D is imposed on society only when employees exert the necessary destructive effort; the externality can thus be avoided. However, without internalising this cost it is -when profitable- in the interest of the organisation to encourage this effort of its employees. The choice of effort is a discrete one. They can choose between exerting destructive effort (e_D) or normal effort (e_N). We consider a situation in which every employee exerts normal effort[6], which returns a fixed revenue for the firm. On top of this, pay-for-performance offers a financial tool to incentivise workers. When this is offered, a flat wage w is paid in combination with a bonus b_D conditional on the observed effort. Ex-

[6]To ensure this, the cost of normal effort can be seen as 0 or that it can be contracted on.

erting destructive effort is thus profitable for the organisation if the benefit of each unit B is lower than the bonus paid to workers. The profit earned on each worker is then:

$$\Pi_i = r - w + e_{D,i}(B - b_D)$$

where r is the base revenue of each worker, w is the flat wage paid to the worker and $e_{D,i}$ is the effort selected by individual i and equals 1 if destructive effort is chosen, and 0 otherwise. B is the private benefit of destructive effort to the organisation and b_D is the bonus paid to the worker upon observing destructive effort. The total profit of the organisation is consequently:

$$\Pi = \sum_{i=1}^{n} \Pi_i = n * (r - w + \varnothing e_D * (B - b_D\ )) \tag{1}$$

where n is the number of employees at the firm and $\varnothing$ refers to the share of employees exerting destructive effort.

Finally, depending on the decisions of both the individuals and the organisation, the impact on society is realised. It is a function of the baseline welfare $\overline{W}$[7] and each unit of destructive effort exerted by employees imposing the externality.

$$W = \overline{W} - I_D \sum_{i=1}^{n}\ e_{D,i} \tag{2}$$

where W is total societal welfare, $\overline{W}$ is the baseline societal welfare, $e_{D,i}$ is the dummy variable for the effort level chosen by individual i and I_D is the impact on society, such that $I_D \geqslant 0$.[8] The individual's utility function is

[7] A baseline welfare is used, because when deciding between one or more alternatives, the baseline value becomes irrelevant and thus need not be further determined. The important aspect is how welfare changes in response to the actions taken.

[8] I_D is a cost to society. This itself is positive but features in the welfare function negatively.

updated to reflect this decision:

$$U_i = w + e_{D,i} * (b_D - c_D) + \gamma_i \, W$$

3.2 Cost of Altruism

If exerting destructive effort is profitable, organisations may find it profitable to encourage workers to exert such effort, meanwhile discouraging workers who do not exert destructive effort from joining in the first place. The contract of wages and bonuses can be set such that these goals are maximised. Sections 5.1 and 5.2 develop regulatory limitation on the values these wages and bonuses may take on.

Proposition 1. The bonus required to exert destructive effort increases with the level of altruism.

Proof: This is evident from the fact that altruists internalise the negative cost of destructive effort on society, and thus need compensation to exert high effort.

$$U_D \geqslant \; U_N$$

$$b_D \geqslant \; c_D + \gamma_i \, I_D \tag{3}$$

The bonus must be larger than the personal and the societal cost of the action. Hence, the required compensation to exert destructive effort increases with the extent to which they care about the cause (their morality). This reiterates the immorality premium Schneider et al. (2020) find using immorality aversion. However, the mechanism here is a concrete social motivation of individuals instead of a more abstract personal cost of acting immorally. The pure altruism is an extension on their conceptualisation of immorality aversion and develops further interesting findings.

4 Equilibrium Analysis

To be an equilibrium, each decision maker must have no incentive to deviate from their current decision and subsequent action. If it exists, the stable equilibrium (equilibria) can be identified through backward induction. Hence, the order of the game is important and goes as follows.

First, the organisation sets a wage and bonus package and presents it to the labour market. Second, the workers willing to work for the specified package apply.[9] This decision is made on the expectations of the labour market outcome. For these to be rational expectations, they are correct in equilibrium. Upon receipt of applications, the organisation offers contracts and vacancies to workers. Importantly, the organisation cannot ex-ante distinguish between types of employees or the effort level they will ultimately exert. They thus select a random subgroup of those who applied to fill the vacancies. Next, the workers exert their desired effort level. Finally, societal welfare is realised. Following backward induction, I will start with the final decision to be made and work backward toward the optimal contract setting of the organisation.

Effort Selection

The last decision to be made and the first to be assessed is the worker's decision to exert destructive or normal effort. e_D is only exerted iff the incentive compatibility constraint (IC) of the worker in *Eq.3* is met.

Entering the Labour Force

Depending on their desired effort choice, an individual decides if they are willing to work for the organisation given the salary package. In general

[9] To ensure only those willing to work apply I assume there to be a cost of rejection. Only those ready to accept apply in the first place.

terms, an individual is willing to work for the organisation if their utility through working is higher than not working. This renders the following participation constraint (PC).

$$w + e_D * (b_D - c_D) + \gamma_i W \geqslant \overline{u} + \gamma_i W \tag{4}$$

The left hand side (LHS) represents the income from work, while on the right hand side (RHS) $\overline{u}$ is the utility of their outside option. The total welfare of the cause features on both sides because pure altruists care about welfare in its own right as opposed to exclusively their contribution to welfare. Importantly, they consider the counterfactual of their decision. Knowledge and expectations of the counterfactual are thus important for the decision maker.

Perfect Information

Before working on a more applicable model, I will work in a simple framework of perfect information to highlight the impact of decisions. In this, the individual knows exactly who they are replacing; they know the counterfactual of their action. For each of their decisions to exert destructive or normal effort, 2 cases exist. The first is if they replace a worker who **is exerting** destructive effort. The second is if they replace a worker who is **not exerting** destructive effort. By *Eq.4*, the PC is met if their utility in working at the organisation is higher than their net outside utility and the total welfare were they not to work there.

In each case the PC looks slightly different. These will be assessed one by one before considering if they are a stable equilibrium, where both the decision on effort choice and participation hold simultaneously. For now, we assume they know which worker they replace.

Case 1

Individual **does exert** destructive effort

Case 1.1: they replace a worker **exerting** destructive effort

$$w - \bar{u} + (b_D - c_D) \geqslant 0$$

Case 1.2: they replace a worker **not exerting** destructive effort

$$w - \bar{u} + (b_D - c_D) - \gamma_i I_D \geqslant 0$$

Case 2

Case 2.1: they replace a worker **exerting** destructive effort

$$w - \bar{u} + \gamma_i I_D \geqslant 0$$

Case 2.2: they replace a worker **not exerting** destructive effort

$$w - \bar{u} \geqslant 0$$

To determine the workers willing to work, the condition on effort and the participation constraint must hold simultaneously.

4.1 Equilibrium Considerations - Perfect Information

Case 1.1: exert e_D & replace e_D

$$b_D - c_D \geqslant \gamma_i I_D \qquad \& \qquad w - \bar{u} + b_D - c_D \geqslant 0$$

In this instance the counterfactual does not influence the decision. The externality will be imposed regardless, so the consideration is purely one of self interest. The first equation is the IC and the second is the PC, which both must be met for the individual to be willing to work.

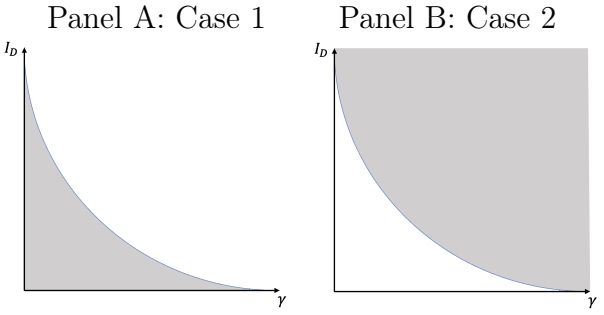

Figure 1: Cases in Equilibrium

Notes: These panels graphically represent the relationship between a destructive impact society on the y-axis and the level of altruism on the x-axis. In both cases, the grey area highlighted represents the combination for which individuals are willing to join an organisation.

Case 1.2: exert e_D & replace e_N

$$b_D - c_D \geqslant \gamma_i I_D \qquad \& \qquad w - \bar{u} + b_D - c_D \geqslant \gamma_i I_D$$

For a worker to be willing to enter the organisation, replacing a worker exerting normal effort and exerting destructive effort himself, the value $\gamma_i I_D$ must be smaller than the smallest condition. Hence:

$$\gamma_i I_D \leqslant \min\{b_D - c_D, w - \bar{u} + b_D - c_D\}$$

The graphic in Panel A of *Fig.1* depicts this decision graphically. Those on the line are indifferent between entering or not. Those in the shaded area are willing to enter the job. The white area are all the individuals who are not willing to do so. All else equal, as the bonus and outside options increase the border line shifts north-east. As the cost of destructive effort or the wage increases, this line shifts downward and left.

Consequently, the maximum level of altruism of an individual willing to exert destructive effort and join the organisation given an impact on society $(\gamma(I_D))$ is:

$$\breve{\gamma}(I_D) = \frac{\min\{b_D - c_D, w - \bar{u} + (b_D - c)_D\}}{I_D}$$

All else equal, this threshold increases with the bonus and the wage, decreases with the cost of destructive effort, and decreases as the magnitude of the negative impact on society increases.

Case 2.1: exert e_N & replace e_D

$$b_D - c_D \leqslant \gamma_i I_D \qquad \& \qquad \bar{u} - w \leqslant \gamma_i I_D$$

For both conditions to hold simultaneously, the value $\gamma_i I_D$ must be larger than the largest condition. Hence:

$$\gamma_i I_D \geqslant \max\{b_D - c_D, \bar{u} - w\}$$

Panel B of *Fig.1* depicts this decision graphically. The shaded area is the combinations of societal impact and altruism in which individuals join the firm to 'save' the impact on society while foregoing the bonus. Those on the line are indifferent and those in the white area are unwilling to join the organisation. All else equal, as the bonus and outside options increase, this line shifts right and up. As the cost of destructive effort or the wage increases, this line shifts down and left. Consequently, the minimum altruism given an impact on society ($\gamma(I_D)$) required for workers to join the workforce and not exert destructive effort is:

$$\gamma(I_D) = \frac{\max\{b_D - c_D, \bar{u} - w\}}{I_D}$$

All else equal, this threshold increases with the bonus and the outside option and decreases with the wage or as the magnitude of the negative impact on society increases.

Case 2.2: exert e_N & replace e_N

$$b_D \leqslant c_D - \gamma_i I_D \qquad\qquad w \geqslant \bar{u}$$

Once again, the counterfactual does not influence the decision here. The externality will not be imposed regardless, so the consideration is purely self motivated. The first equation is the IC and the second is the PC and both must be met for the individual to be willing to work.

4.2 Equilibrium Considerations – Imperfect Information

A worker in the real world does not know exactly whom they replace. In fact, a worker is unlikely to simply replace another. Instead, they change the composition of those applying and -given the random draw from those applying- the composition of the hired workforce. Particularly, this means that the expected 'saved' impact on society is rarely as extreme as saving the entirety of I_D as was assumed in the case of perfect information in the previous section. Instead, the proportions of different types change with the marginal worker at the thresholds $\underline{\gamma}$ and $\check{\gamma}$. In this case, it is the *expectation* of the workforce that is crucial to the decision making. In equilibrium, these expectations must be correct.

To highlight the decision process graphically, the population's altruism levels are presented in Figure *Fig.2*. The continuum of workers runs from the smallest level of altruism 0 to that of the most altruistic in the population $\bar{\gamma}$. $\widetilde{\gamma}$ represents the marginal worker who would be incentivised to exert destructive effort if employed at the organisation. Those to the left would exert destructive effort and those to the right would not. However, not all of each type are willing to join the organisation. Given their effort choice and organisational policies not all PCs are met. Only to the left of $\check{\gamma}$ are workers exerting e_D willing to join the organisation (in green). On the other side of the spectrum, only to the right of $\underline{\gamma}$ are workers exerting e_N willing to join the organisation (in blue). The proportion of the population applying with the intention of exerting destructive effort is then $\frac{\check{\gamma}}{\bar{\gamma}}$ and is the fraction highlighted in green in *Fig.2*. The proportion of the distribution applying for the job who want to exert normal effort is $\frac{(\bar{\gamma}-\underline{\gamma})}{\bar{\gamma}}$ and is highlighted blue in *Fig.2*. Those in between the two thresholds do not join the workforce as their PC is not met for their desired choice of effort. As the employed are a random subgroup of those who apply, the distribution in the applicants pool

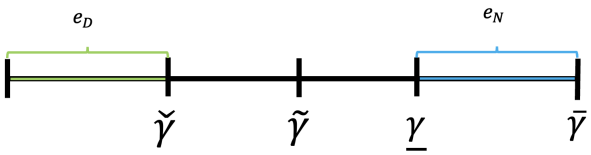

Figure 2: Distribution of Workers

Notes: This figure represents the population distributed by their altruism level. This ranges from 0 on the left to $\bar{\gamma}$ on the right. The two thresholds determine the edges of the ranges for the levels of altruism in which individuals join or do not join the labour force and their effort choice.

matches the distribution in the workforce. Consequently, the proportion of the workforce exerting destructive effort is $\frac{\check{\gamma}}{\check{\gamma}+(\overline{\gamma}-\underline{\gamma})}$. The total workforce is $\frac{\check{\gamma}+\left(\overline{\gamma}-\underline{\gamma}\right)}{\overline{\gamma}}$.

To determine the equilibrium in society, the marginal worker, the one indifferent between working and not working, must be identified at both thresholds.

Identity of the Indifferent Worker

Last **Moral** Worker to Join the Force

The decision to apply depends on the expected net benefit of working. This consideration includes comparing oneself to the characteristics of the worker you would replace. It is not possible to view which the exact type is, but it is possible to observe the expected or average worker one would replace. The saving to society is thus the difference in action between the average worker employed at the organisation and those considering the decision. From *Eq.4* we can derive the PC of this indifferent worker in terms of the expected gain to society.

$$w - \bar{u} = \gamma_i I_D \left(\frac{\check{\gamma}}{\check{\gamma} + \left(\overline{\gamma} - \underline{\gamma}\right)} \right)$$

In equilibrium, the marginal worker must be indifferent and their PC must be binding. Plugging in the identity of the marginal worker $\underline{\gamma}$ and solving for them, we arrive at the following altruism level for this worker.

$$\underline{\gamma} = \frac{\check{\gamma}\,(w - \bar{u}) + \overline{\gamma}\,(w - \bar{u})}{w - \bar{u} - I_D\check{\gamma}}$$

Last **Immoral** Worker to Join the Force

The same steps can be used to derive the identity of the marginal immoral applicant. In this case, *Eq.4* can be rearranged to:

$$w - \bar{u} + e_D * (b_D - c_D) = \ \gamma_i I_D \left(1 - \frac{\check{\gamma}}{\check{\gamma} + (\overline{\gamma} - \underline{\gamma})}\right)$$

Solving this equality for $\check{\gamma}$ the identity of the last immoral worker is characterised by this equation:

$$\check{\gamma} = \ -\frac{(w - \bar{u} + b_D - c_D)\left(\overline{\gamma} - \underline{\gamma}\right)}{(w - \bar{u} + b_D - c_D) - I_D\left(\overline{\gamma} - \underline{\gamma}\right)}$$

Both thresholds depend on the identity of the other threshold. Combining these, the identities can be determined with only exogenous variables.

$$\check{\gamma} = \frac{\overline{\gamma}\,(w - \bar{u} + b_D - c_D)}{\overline{\gamma} I_D - b_D + c_D} \tag{5}$$

$$\underline{\gamma} = \frac{\overline{\gamma}(w - \bar{u})\,(w - \bar{u} + \overline{\gamma} I_D)}{((-b_D + c_D)\,(w - \bar{u}) + I_D(-b_D + c_D))} \tag{6}$$

Given these identities, the share of the workforce exerting destructive effort s is

$$s = \frac{\check{\gamma}}{\check{\gamma} + (\overline{\gamma} - \underline{\gamma})} = \frac{b_D - c_D}{I_D\overline{\gamma}} \tag{7}$$

All else equal, this share is increasing in the bonus paid to workers and decreasing in the cost of destructive effort. This occurs because it directly

affects the willingness of workers to exert destructive effort. This willingness increases in the bonus and decreases in the cost of effort. It also decreases in the size of the externality and the altruism level of the most moral individual. As either increases, more moral individuals are willing to work at the organisation as the benefits of the perceived benefit of their actions increase.

4.3 Organisational Preferences

As described previously, the organisation faces a profit maximisation problem. It has n vacancies that need to be filled and sets its contract to attract the desired workers. It has 2 tools at its disposal to use in the contract; the wage and the bonus for destructive effort.

Maximising revenue is achieved by maximising the share of the employees exerting destructive effort s. However, increasing this share also comes at the cost of b_D, so it may not be profit maximising to completely deter the moral types. This means there is a trade-off between cost and attracting immoral workers in the optimal bonus setting.

As a reference point, we will first assume there are no limitations on any decisions made by the firm, such that complete deterrence is possible. After examining the conditions under which this can occur we will discuss, one-by-one the impact of regulation on wages and bonuses.

Full Deterrence - Reference Equilibrium

To ensure that only immoral workers apply, the contract must be made such that all vacancies can be filled, while the most moral type is unwilling to join the organisation. From *Eq.3* we know that those with the lowest altruism are the first to join the organisation and the cheapest to hire. Therefore, for all vacancies to be filled we require that $F(\check{\gamma}) = n$. To ensure a closed

form solution for the remainder of the analysis $F(\gamma)$ is required explicitly. For computational reasons a uniform distribution of types is assumed across the altruism spectrum. This is a simplification, but the results do not hinge solely on this assumption. This condition then becomes

$$\check{\gamma} = \overline{\gamma} * \frac{n}{N}$$

where

$$\check{\gamma} = \frac{\overline{\gamma}(w - \bar{u} + b_D - c_D)}{\overline{\gamma} I_D - b_D + c_D}$$

Consequently, to ensure that all vacancies can be filled under full deterrence the total number of vacancies cannot exceed:

$$n = N \frac{(w - \bar{u} + b_D - c_D)}{\overline{\gamma} I_D - b_D + c_D} \tag{8}$$

This provides a benchmark for assessing the likelihood of full deterrence being a possible outcome. However, under no restrictions to the contract setting of the organisation, Eq.8 is trivially met. Because both the wage and the bonus paid can be freely adjusted, they can be set to ensure it always holds.[10] To discourage the most moral type, the wage must be set such that the PC of *Eq.4* is not met for $\bar{\gamma}$. To ensure this,

$$w \leqslant \bar{u} - \overline{\gamma} I_D \tag{9}$$

This wage setting condition will be binding in the optimum. Intuitively, the wage must be so low that even for the most altruistic individual, the wage and the perceived benefit of joining the organisation is weakly worse than their outside option. If the outside option is normalised to 0, the wage becomes negative. Individuals will have to pay to be allowed to work at the

[10]Plugging in the optimal bonus and wage of *Eq.9* and *Eq.10* the limit of n becomes N. The restriction on n thus only becomes relevant under restrictions on the optimal contract. These calculations are provided in the appendix.

firm. Section 5.1 discusses the implications this negative wage setting has in the real world.

After the wage has been determined, the other necessary condition on the contract is that the marginal immoral worker *is* willing to work. Given the optimal wage and that their PC from *Eq.5* must be met, the optimal bonus becomes:

$$b_D^* = \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D \tag{10}$$

Intuitively, the first part of the RHS compensates this marginal worker for their personal cost of harming society. The second part is the compensation required to make up for the lower wage setting to keep the moral workers out. The third part is then the compensation for the direct personal cost of destructive effort.

However, the firm is only going to pay out the bonus when it is profitable to encourage destructive effort. As such, the part of the bonus that is not compensating the worker for lower wages must be smaller than the marginal benefit of destructive effort B. This is the case as long as $B \geqslant \frac{\overline{\gamma} n I_D}{N} + c_D$. The optimal bonus does not increase beyond this threshold. This means that the ability of an organisation to employ a full-deterrence contract is increasing in the private benefit of employees acting immorally. In contrast, it is decreasing in the altruism of the most caring individual, the number of employees needed, the size of the externality and the cost of destructive effort.

The optimal contract for full deterrence is the combination $w^* = \bar{u} - \overline{\gamma} I_D$ and $b_D^* = \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D$ so long as $B \geqslant \frac{\overline{\gamma} n I_D}{N} + c_D$. Otherwise it is defined as $w^* = \bar{u} - \overline{\gamma} I_D$ and $b_D^* = B + \overline{\gamma} I_D$. Like in Schneider et al. (2020), this model results in an immorality premium for those acting immorally and similarly, it is the marginal worker that determines how big this immorality premium is. It is important to note that in this equilibrium there is no trade-off between the incentivising effect of the bonus and the cost of b_D. This is

the case because there are no restrictions on the wage, such that all rent is extracted from the marginal employees and the organisation has reached the second-best outcome. Total welfare of the cause under full deterrence is $\bar{W} - nI_D$. This outcome can be influenced by institutions such that more socially efficient outcomes can be reached.

5 Regulation and Policy

A regulatory authority or government agency has multiple tools at their disposal to affect the equilibrium that we observed under full-deterrence. The effects of these tools will vary. This section discusses three tools available as well as the resulting impact on welfare with regard to the cause. The tools discussed are the implementation of a minimum wage, a limit to the size of the bonus and direct limitations on the externalities. A further contributing factor to the equilibrium that is outside the remit of the government directly is the societal perception of immoral behaviour.

5.1 Minimum Wage

Setting a minimum wage is realistic tool implemented to reach many other policy objectives. The ability to fully deter workers relies on the ability to satisfy the inequality in *Eq.9*. This ensures the PC of no moral types is met. Let us first consider a stark example of requiring a negative wage to satisfy *Eq.9*. In this instance individuals would be forced to pay to enter the firm and would only be compensated with their bonus. If negative wages were banned, every employee would need to be paid at least 0 (i.e. not paying to enter employment). At this point the employer can no longer discourage all immoral types from applying. Those for whom *Eq.9* is not met now apply and we reach a bifurcated equilibrium. This section will generalise the imposition of a minimum wage and confirms that under certain situations a

bifurcated equilibrium emerges. It is important to note that the minimum wage only has an effect if the full-deterrence wage is below the threshold to begin with. Thus, two cases exist after a minimum wage has been set.

In the first case, the wage is above the minimum wage ($\bar{w} \leqslant w^* = \bar{u} - \overline{\gamma} I_D$) such that the optimal contract does not break any regulations and the optimal contract offered by the firm remains what it was before.

$$\begin{aligned} w^* =& \bar{u} - \overline{\gamma} I_D \\ b_D^* =& \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D \end{aligned}$$

In the second case, the wage in the optimal contract is below the regulatory minimum set by the authorities such that $\bar{u} - \overline{\gamma} I_D = w^* \leqslant \bar{w}$. Out of necessity, the wage offered must increase to above or equal to that threshold. As the organisation still wants to have the wage as close to the optimal full deterrence value, the wage in the contract becomes the minimum wage. After being compelled to deviate from the optimal contract for the wage, the only tool left for the organisation to determine is the bonus they offer. To determine the optimal bonus, the profit maximisation problem must be solved under the new restriction on wages. *Eq.1* must be maximised w.r.t. b_D.

$$\Pi = n * ((1 - \varnothing e_{D,i})(r - \bar{w}) + \varnothing e_{D,i} * (B - \bar{w} - b_D))$$

where $\varnothing e_{D,i} = s = \frac{b_D - c_D}{I_D \overline{\gamma}}$ such that

$$\Pi = n * \left(\left(1 - \frac{b_D - c_D}{I_D \overline{\gamma}} \right) (r - \bar{w}) + \left(\frac{b_D - c_D}{I_D \overline{\gamma}} \right) (r + B - \bar{w} - b_D) \right) \quad (11)$$

Maximising *Eq.11* w.r.t. the bonus, the optimal bonus becomes:

$$b_D^* = \frac{B - \bar{w} + c_D}{2} \quad (12)$$

Paying this bonus to workers is only profitable for the organisation if the return on destructive effort is higher than the cost of incentivising it. The

requirement for this is that $B - b_D^* \geqslant 0$ and is met if $B \geqslant c_D - \bar{w}$. Otherwise destructive effort is not profitable for the organisation and the organisation is indifferent between hiring moral or immoral workers. As such, the bonus becomes zero. The profitability of destructive effort is negatively related to the effort costs of destructive effort, yet positively to the minimum wage. The optimal contract is then given by $(\bar{w},\ b_D^*)$ if $B \geqslant c_D - \bar{w}$ or by $(\bar{w}, 0)$ otherwise.

The resulting welfare change of the policy is given by the change in the composition of the workforce. The calculations are provided in the appendix, but the expression simplifies to:

$$\Delta W = W_2 - W_1 = n(I_D - \frac{(B - c_D) - \bar{w}}{2\bar{\gamma}})$$

Whether this policy increases or decreases the total welfare depends on the situation at hand. If the benefit to an organisation is high, then it is able and willing to increase bonuses considerably to maintain a high share of immoral workers. Thus, the higher this private return, the lower is the benefit of the policy to society. The effectiveness of the policy however is greater with a higher level of maximum altruism. Furthermore, its effectiveness increases in the size of the externality as well as the share of the population employed at the organisation.

5.2 Maximum Bonus

An alternative target of policy is the bonuses that are paid out to employees. As with the policy on wages, if this affects the optimal contract offered by the organisation, the equilibrium set of workers can be affected. Implementation strategies of such a policy can be direct caps or can target the taxation of above-salary compensation, such that the effective bonus rate is decreased. As in the minimum wage case, there are two cases to consider.

1. It does not limit anything

In the case that the organisation is already paying a bonus below the maximum, their personnel policies do not change and there is no impact on the welfare of the cause.

$$\bar{b} \geqslant b^* = \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D$$

Consequently the optimal contract remains:

$$w^* = \bar{u} - \overline{\gamma} I_D$$
$$b_D^* = \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D$$

2. It limits full deterrence

$$\bar{b} \leqslant b^* = \frac{\overline{\gamma} n I_D}{N} + \overline{\gamma} I_D + c_D$$

In this instance, the bonus is above the threshold and must be decreased. It will be set as high as possible -at $\bar{b}$- to ensure maximal deterrence. The wages are consequently set to maximise profits ensuring all vacancies are filled. Altering the wage has a direct effect on the marginal worker attracted at *both* ends of the altruism spectrum. The profit equation to maximise w.r.t. w in *Eq.1* then becomes:

$$\Pi = n * \left(r - w + \frac{\bar{b} - c_D}{I_D \overline{\gamma}} * \left(B - \bar{b} \right) \right)$$

As profit is strictly decreasing in the wage decision, the organisation maximises profits at the point of minimising the wage costs such that the n vacancies are filled. The change in wage does not change the composition of the workforce, as s is independent of the wages. The change in the cost to society is thus determined through the impact that a decrease in the bonus has on the share of immoral workers at the organisation. The derivations provided in the appendix simplify to:

$$\Delta W = W_2 - W_1 = n(I_D - \bar{b} + c_D) \tag{13}$$

The effects of the policy are again increasing in the damaging effect they have on society as well as the share of the population employed at the organisation. Further, it is increasing in the personal cost of destructive effort, as less workers can then be paid enough to incentivise the effort. The total effect saved on society is also decreasing in the magnitude of the reduction in the bonus. The lower the bonus cap, the greater the positive welfare effects. In the extreme case, in which no bonus can be used to incentivise destructive effort, none would be exerted. As with the minimum wage, the policy is more effective the higher the altruism of the most altruistic.

Implementing a cap on bonuses could prove tricky in practice. Despite the simplified set-up here, a pragmatic solution of implementing something to the same effect is to introduce a tax on bonuses (salary beyond that contracted upon). In this interpretation of the limit on the bonuses it would be the net bonus that the worker cares about. The tax would drive a wedge between what the organisation pays and what the worker receives, in effect restricting the ability to fully-deter individuals.

5.3 Limit Externality

Another tool institutions can implement is to impose direct limits on the size of externalities. This could take the shape of an outright restriction or simply to increase the cost of acting poorly. Doing so makes encouraging destructive effort less (or entirely un-) profitable and desirable for organisations. As regulation increases, the difference in impact between those who do and do not exert destructive effort decreases, such that the 'saved' impact on society of altruists shrinks. Decreasing the externality has two effects on the configuration of the workforce and societal welfare. The first is that for a given number of employees working destructively, the total amount of externality decreases.

However, there is also an indirect effect. As the counterfactual of a moral

worker's decision to join the workforce decreases, the pull to join the organisation also decreases. Therefore, all else equal, less altruists would be willing to work for the immoral organisation. These two effects work in opposite direction, so the overall impact on society is not immediately obvious.

The overall change in equilibrium impact on welfare is calculated using *Eq.2*, where the bonus setting is an endogenous decision the firm faces. In a full-deterrence equilibrium, where there are no restrictions on the bonus and wage combination, these will be set such that the share of employees exerting destructive effort at the organisation is 1. If this policy is implemented alone, the share is 1 both before and after the intervention. Plugging this in, we find the total effect on welfare to be given by:

$$\Delta W = W_2 - W_1 = n\left(I_{D1} - I_{D2}\right) \tag{14}$$

The resulting benefit to the cause in society is thus exactly the difference in the externality of each worker multiplied with the number of employees exerting destructive effort. This makes sense, as the policy is thwarting their freedom on externalities, but not restricting the optimal contract in a way that forces the composition of the workforce the change. An important question that falls beyond the scope of this paper is the cost at which this reduction in externalities on the cause can be implemented by the firm. Depending on the cost function, increasing the welfare of the cause may not be efficient on a broader spectrum.

Under the minimum wage policy considered in section 5.1, the optimal bonus is independent of the size of the externality. Personnel contracts offered both before and after the policy implementation are independent of the the size of the externality, as determined by *Eq.12*. Consequently, the bonus set in both periods is the same. As a consequence, there is *no change* to overall welfare and a limit on the size of the externality becomes an ineffective tool to wield. This happens because the avoided social cost from each of the workers previously employed is one-for-one cancelled out by moral workers

leaving the workforce[11]. The altruism level of the marginal moral worker increases because the amount of welfare saved through their actions drops, so less moral workers remain employed[12]. It is thus important to consider the other limits imposed on the contract before blindly implementing such an ineffective policy. As the calculations in the appendix show, it is only effective when combined with a policy that changes the optimal bonus of the organisation simultaneously. For example, by directly decreasing the bonus cap or indirectly by increasing the minimum wage.

6 Results and Discussion

Modelling pure altruism in labour market decisions of individuals leads sorting in equilibrium. The extent of this sorting depends on various factors of the labour market. It is optimal for the organisation to completely deter moral workers from joining the firm when possible and profitable. When this is feasible there are considerable costs of sorting by morality and the cause's welfare loss is greatest. However, full-deterrence is not always possible. Unlike in Schneider et al. (2020), a bifurcated equilibrium can be sustained in which both moral and immoral types work at the organisation. In this, individuals from both ends of the altruism spectrum join the organisation. Those on the immoral end with the aim to act immorally for monetary compensation, but those at the other end join and don't exert destructive effort. The moral workers are encouraged to join the workforce through the ability to save society some of the externality that would be imposed if an immoral worker were to take their place. The policies dis-

[11]This is a consequence of the uniform distribution of workers across the altruism spectrum. Another CDF could see differing marginal effects on the number of moral or immoral types willing to work.

[12]Mathematically, this can be seen in the derivations of the welfare function in the Appendix.

cussed and assessed show a varying degree of success and their effectiveness depends also on other factors in the industry's configuration. Unsurprisingly the benefits were highest when the externality was large and the size of the organisation's workforce was higher. The stricter the policies, the larger the benefit to society as well. Interestingly, the level of altruism of the most moral individual plays an important role.

Each of the policies' effectiveness is improved as the altruism level of the most moral individual in the society increases. This can be explained by the morality increasing the intrinsic motivation, making it harder to deter the most moral individuals. When this becomes harder, the other policies are more effective. There are however some limitations to their effectiveness. Despite the ability for a minimum wage to be effective, it may be set as the results of a different policy aim and may be a politically charged device. It could be hard to change with the environmental impact in mind.

Beyond the altruism level of the most moral individual, the distribution of the altruists plays an important role in this model. The altruism level of the most caring individual directly shapes the effectiveness of policy interventions as well as decreasing the likelihood of full deterrence in the organisation. On top of this, the marginal immoral worker also plays a key role. If they shift up in altruism level, the same contract aim of the firm becomes more expensive to run. As described by Schneider et al. (2020) this leads to a higher pay for all immoral types, but it also decreases the profitability of acting poorly. There may also be more types of people than those considered in this model. Instead of differing on their level of altruism alone, some in society may be motivated by the *desire* to do bad. This model has bounded the spectrum of altruism at 0, but such spiteful motivation would lower the threshold into the negative side and lead to a utility gain from harming society. The policies required here would differ because it would be in the organisation's interest not to hinder the employee's actions. Furthermore, as this distribution is a population-wide measure, shifts in attitude

are an important tool to be exploited. In fact, the effects of such aims can have large effects and compound other policies to have an even greater impact. Unlike legislative tools at the disposal of governments, attitude can be affected by individuals and organisations alike.

The aim of activists for many causes around the world is to affect the attitude of everyone by increasing awareness of societal issues. By highlighting malpractices and raising concerns that fall beyond peoples' general remit of concern can have large impacts on the viability and profitability of immoral actions of organisations (in this model B). Examples include boycotts of harmful products or 'bad' producers. The decision of individuals engaging in this can be explained in the framework in this paper. These activists are individuals who feel strongly about a cause and do something about it, even outside the employment at an organisation. Some feel so strongly about their cause that they take action even at great personal cost. For example, the fear of long term impacts of unsustainable environmental policy pushes some to weigh up the costs of large fines and possibly getting arrested to make their voices heard. At unprecedented times, the public can encourage wide reaching action by private enterprises. A present example is the boycott by private and public organisations of the Russian market in response to Russia's actions in Ukraine. This shows that when the will is strong enough swift action can be taken even when it goes against a purely profit maximising motive. The fact that influencing the population as a whole has such large effects shows that the people as well as the government can and should take responsibility, take personal steps and push for wider action to be taken to avoid large welfare costs.

In the context of moral actions in organisations, whistle-blowers are an interesting topic. Many examples exist of whistle-blowers exposing goings-on in organisations and going to authorities with evidence of illegal activities. Such an aim could be another driver for moral people to join and work for immoral organisations not considered in this paper. This possibility has not

been built into the model, but could be added through the altruism mechanism. The long run benefit of working at the immoral organisation could then be a large saving on society ensured through future tighter regulation as well as the threat of fines.

There are a few limitations that must be considered in the applicability of the model. The impacts on welfare were assessed only by their impact on a specific cause. This does not consider the overall utility of all in the economy. To gain a complete picture these preferences would need to be modeled as well. Next, throughout the analysis it was assumed that all employees have the same potential to implement the outcome on society. Decision rights are likely not evenly distributed across every employee in an organisation. Positions of power are skewed towards managers and personnel higher up the structure. The ability to rise up the ranks may be crucial to have long lasting and big impacts. Furthermore, the bonus and pay structures are likely to differ throughout such structures. If the impact of any one decision is determined by the decision rights an individual has, managers making 'good' decisions would have a much higher counterfactual if they were to make other decisions. Making decisions harmful to society comes at a greater personal cost the higher up the structure. To compensate for this, managers with higher decision rights may receive higher bonuses to maintain enough incentives to ensure the destructive decisions are made by making up for the higher costs of acting immorally.

Notably, this paper drew on the distinction made in previous papers between pure and impure altruism and implemented that of pure altruism in the model. The results would look different if we assume individuals not to care about a cause in itself, but also other factors. For example, image concerns could counteract pure altruistic preferences. Such concerns could discourage some from joining a 'bad' organisation for fear of how it will make them look in the eyes of others, despite their intrinsic and pure altruism level.

The tools of the contract being the only option the organisation has to

determine who joins its workforce may be undermined by other forms of screening. Such effort can make use of an individuals' past record or outside interests to gain a better understanding before the hiring process is completed. Under such measured complete (or at least partial) deterrence may be more feasible.

Further extensions to this model are also interesting. This paper has worked on only one organisation with the potential of imposing a negative externality on society, while assuming that the alternative is work with a neutral impact on society. The introduction of a positive organisation or sector adds a level of realism and further counterfactuals to consider for any decision maker. In this environment there are greater opportunity costs to working and mitigating a negative externality at an immoral organisation. An equilibrium would have to include the effect of 'stealing' workers from a positive sector and the benefit and costs of such effects need to be assessed in tandem. In this setting regulation may be better suited at limiting the bad at the bad organisation than to stimulate good behaviour at the good organisation. Thus, ensuring the intrinsically motivated work at firms with positive potential and are not driven to put their effort to mitigating bad outcomes may be important to reap the full benefits moral employees can offer. The same policy may have a greater impact on society.

Three other extensions remain interesting. The first concerns the choice in this model of a binary effort decision. Allowing for a continuous effort choice allows for less black and white decisions to be made. More workers could join the workforce and forego just as much destructive effort (and bonus) as makes them indifferent to their outside option. All else equal, this should increase the share of the population willing to work at the organisation. Particularly, it should increase the number of those in the middle of the distribution working at the firm. Consequently, this allows the pool of workers to come from the extremes (as seen in the bifurcated equilibrium) as well as those saving *some* of the destructive impact from the middle. The

effect on welfare is not immediately clear as these workers replace both the immoral types (saving society on destructive impact) as well as moral types (reducing the saved impact). Further analysis is needed to determine the outcome.

The second could see the number of employees at a firm endogenised. Under situations in which full-deterrence is not possible, it may be desirable for organisations to cut back on their production levels to avoid filling their ranks with moral workers. In doing so, they reduce the number of employees required at the firm. The willingness to do so will depend on the profit differential between moral and immoral workers. If the profit differential is high enough, supply of these goods could be restricted in response.

The final extension concerns the size of the externality. So far the impact on society has been exogenous, assumed to be dependent on the type of organisation itself. However, firms may also be able to change their societal impact. To implement this, the decision space of the firm can be expanded to include investment in technology that decreases the impact on society. In a situation where full-deterrence is not possible, a trade-off exists in which it may be optimal to pay to reduce the size of the externality so less moral individuals apply. This would raise the potential of a further policy instrument: subsidising such investment. Whether such subsidies are welfare increasing or crowd out private investment remains to be explored.

In conclusion, under certain conditions a bifurcated equilibrium can be sustained, resulting in higher welfare of the cause than under a full-deterrence equilibrium. Despite sorting by morality occurring in the labour market, both governments and society as a whole have effective tools at their disposal to reduce the welfare losses. Consequently, the responsibility should not fall on any one agent alone. Instead, everyone who can influence the outcome can and should take responsibility for their part and do what they can.

References

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of political Economy*, 97(6):1447–1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477.

Arunachalam, R. and Shah, M. (2008). Prostitutes and brides? *American Economic Review*, 98(2):516–22.

Ashraf, N. and Bandiera, O. (2018). Social incentives in organizations. *Annual Review of Economics*, 10:439–463.

Auriol, E. and Brilon, S. (2014). Anti-social behavior in profit and nonprofit organizations. *Journal of Public Economics*, 117:149–161.

Bartling, B., Weber, R. A., and Yao, L. (2015). Do markets erode social responsibility? *The Quarterly Journal of Economics*, 130(1):219–266.

Bates, C. and Rowell, A. (2004). Tobacco explained... the truth about the tobacco industry... in its own words.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents. *American economic review*, 95(3):616–636.

Buurman, M., Delfgaauw, J., Dur, R., and Van den Bossche, S. (2012). Public sector employees: Risk averse and altruistic? *Journal of Economic Behavior & Organization*, 83(3):279–291.

Delfgaauw, J. and Dur, R. (2008). Incentives and workers' motivation in the public sector. *The Economic Journal*, 118(525):171–191.

Dur, R. and Zoutenbier, R. (2014). Working for a good cause. *Public Administration Review*, 74(2):144–155.

Dur, R. and Zoutenbier, R. (2015). Intrinsic motivations of public sector employees: Evidence for germany. *German Economic Review*, 16(3):343–366.

Egan, M., Matvos, G., and Seru, A. (2019). The market for financial adviser misconduct. *Journal of Political Economy*, 127(1):233–295.

Falk, A. and Szech, N. (2013). Morals and markets. *science*, 340(6133):707–711.

Frank, R. H. (1996). What price the moral high ground? *Southern Economic Journal*, pages 1–17.

Friebel, G., Kosfeld, M., and Thielmann, G. (2019). Trust the police? self-selection of motivated agents into the german police force. *American Economic Journal: Microeconomics*, 11(4):59–78.

Ghatak, M. and Mueller, H. (2011). Thanks for nothing? not-for-profits and motivated agents. *Journal of Public Economics*, 95(1-2):94–105.

Gill, A., Heinz, M., Schumacher, H., and Sutter, M. (2020). Trustworthiness in the financial industry.

GlobeScan (2021). "new global poll ahead of cop26 in glasgow shows growing support for governments to take strong action on climate change". Retrieved from https://globescan.com/2021/10/27/global-poll-cop26-growing-support- governments-take-strong-action-climate-change/ on 22/01/2021.

Gregg, P., Grout, P. A., Ratcliffe, A., Smith, S., and Windmeijer, F. (2011). How important is pro-social behaviour in the delivery of public services? *Journal of public economics*, 95(7-8):758–766.

Hart, O. and Zingales, L. (2017). Companies should maximize shareholder welfare not market value. *ECGI-Finance Working Paper*, (521).

Heath, D. (2016). Contesting the science of smoking. *The Atlantic. Retrieved from https://www. theatlantic. com/politics/archive/2016/05/low-tar-cigarettes/481116*.

Hjort, J., Streletskiy, D., Doré, G., Wu, Q., Bjella, K., and Luoto, M. (2022). Impacts of permafrost degradation on infrastructure. *Nature Reviews Earth & Environment*, 3(1):24–38.

IPCC (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press. In Press.

Laville, S. (2019). Two-thirds of britons want faster action on climate, poll finds. *The Guardian. Retrieved on 18/07/2021 from https://www.theguardian.com/environment/2019/jun/19/britons-want-faster-action-climate-poll*.

Moffatt, P. G. and Peters, S. A. (2004). Pricing personal services: An empirical study of earnings in the uk prostitution industry. *Scottish Journal of Political Economy*, 51(5):675–690.

Ockenfels, A., Werner, P., and Edenhofer, O. (2020). Pricing externalities and moral behaviour. *Nature Sustainability*, 3(10):872–877.

Prendergast, C. (2007). The motivation and bias of bureaucrats. *American Economic Review*, 97(1):180–196.

Reuters (2019). Americans' attitude on climate action. Retrieved from https://graphics.reuters.com/USA-ELECTION- CLIMAT-ECHANGE/0100B03104Z/index.html on 18/07/2021.

Schneider, F., Brun, F., and Weber, R. A. (2020). Sorting and wage premiums in immoral work. *University of Zurich, Department of Economics, Working Paper*, (353).

US Department of Labor (2017). Child labor in the production of cocoa.

Appendix - Calculations

Maximum Number of Vacancies - n

$$n = N\frac{(w - \bar{u} + b_D - c_D)}{\overline{\gamma} I_D - b_D + c_D}$$
$$n = N\frac{\left(\overline{\gamma} I_D - \frac{\overline{\gamma} n I_D}{N} - \overline{\gamma} I_D\right)}{\overline{\gamma} I_D - \frac{\overline{\gamma} n I_D}{N} - \overline{\gamma} I_D}$$

Welfare Effects - Minimum Wage

$$\begin{aligned}
\Delta W &= W_2 - W_1 \\
&= (\bar{W} - s_2 n I_D) - (\bar{W} - s_1 n I_D) \\
&= n I_D \left(s_1 - s_2\right) \\
&= n I_D \left(1 - \frac{\frac{B - \bar{w} + c_D}{2} - c_D}{I_D \overline{\gamma}}\right) \\
&= n(I_D - \frac{B - \bar{w} - c_D}{2\bar{\gamma}}) \\
&= n(I_D - \frac{(B - c_D) - \bar{w}}{2\bar{\gamma}})
\end{aligned}$$

where we know that before the intervention $s_1 = 1$

Welfare Effects - Maximum Bonus

$$\begin{aligned}
\Delta W &= W_2 - W_1 \\
&= (\bar{W} - s_2 n I_D) - (\bar{W} - s_1 n I_D) \\
&= n I_D \left(s_1 - s_2\right) \\
&= n I_D \left(1 - \frac{\bar{b} - c_D}{I_D}\right) \\
&= n(I_D - \bar{b} + c_D)
\end{aligned}$$

where we know that before the intervention $s_1 = 1$

Welfare Effects - Limit Externality

Full Deterrence Possible:

Full deterrence is possible as there are no binding restrictions on the personnel policies.

$$\begin{aligned}\Delta W &= W_2 - W_1 \\ &= (\bar{W} - s_2 n I_{D2}) - (\bar{W} - s_1 n I_{D1}) \\ &= n\left(s_1 I_{D1} - s_2 I_{D2}\right) \\ &= n\left(I_{D1} - I_{D2}\right)\end{aligned}$$

Binding Minimum Wage:

Personnel contracts offered both before and after the policy implementation are independent of the the size of the externality and instead chosen by 12. Consequently, the bonus set in both periods is the same and we know that $b_{D1} = b_{D2} = b_D$.

$$\begin{aligned}\Delta W &= W_2 - W_1 \\ &= (\bar{W} - s_2 n I_{D2}) - (\bar{W} - s_1 n I_{D1}) \\ &= n\left(\frac{b_{D1} - c_D}{I_{D1}} I_{D1} - \frac{b_{D2} - c_D}{I_{D2}} I_{D2}\right) \\ &= n\left(s_1 I_{D1} - s_2 I_{D2}\right) \\ &= n\left(b_{D1} - b_{D2}\right) \\ &= n(b_D - b_D)\end{aligned}$$