Misconduct in Organizations

Kim Sarnoff and Hassan Sayed [*]

November 5, 2025

Click here for latest version.

Abstract

We study how policies that disincentivize misconduct in organizations can generate changes in abusers' behaviors that negatively impact victims. We describe a model where "managers" choose to commit harmful actions of varying intensities against "employees," who can report these actions as "misconduct." We show that when the marginal disutility from managers' actions is particularly small, increasing the ease of reporting misconduct, the severity of punishment for managers, or the efficacy of investigation technology may in fact harm employees. These policies may motivate managers to commit harmful actions that employees do not want to report or induce managers to opt out of interacting with employees altogether. We provide a dynamic extension where reports generate precedents for the organization and employees, showing that the model converges to a steady state where employees are worse off than initially and harmful actions are never punished.

JEL Codes: D23, D73, D83, D91

Keywords: labor abuse, misconduct, harassment, agency, accountability, organizations, grievances, abuse, learning

[*] We thank Leeat Yariv, Alice Hsiaw, and participants in the Princeton Political Economy Research Seminar, Princeton Microeconomic Theory Student Lunch Seminar, and 36^{th} Stony Brook International Conference on Game Theory for helpful comments.

1 Introduction

Workplace misconduct — such as bullying, negligence of safety standards, or overworking — have well-documented and deleterious effects on employees' productivity, job security, and quality of life. Verbal harassment can impact employees' emotional wellbeing or productivity, and may lead to abuse of authority and privilege that accompany broader violations of *norms* of respect and dignity (Aquino & Thau, 2009). A report from McKinsey and Company documents that victims of workplace abuse experience heightened anxiety, lower workplace satisfaction, and increased turnover (Sutton, 2007). One Irish study pegs the productivity losses of these behaviors on the magnitude of hundreds of millions of Euros (Cullinan et al., 2020).

To prevent misconduct, firms utilize accountability mechanisms where victims of abuse can report perpetrators' actions. These reports are investigated, which can result in punishment for an abuser — such as suspension or termination — and compensation for the abused. In light of the #MeToo movement of the late 2010s, there has been considerable interest in how policies intended to discourage sexual harassment in particular may mitigate abuse and help (or harm) prospective victims. Much of the consequent literature has focused on how the reports themselves may impact employees' welfare through channels like retaliation (Dobbin & Kalev, 2019).

Yet, at the same time, policies that discourage harassment and misconduct also affect perpetrators *incentives* to engage in both harmful and unharmful behavior. Men who undergo anti-harassment training may have more anxiety about working with women, limiting the frequency and *quality* of their workplace interactions Tinkler et al. (2022). Perpetrators may limit negative interactions to *microaggressions* — less egregious (but still harmful) actions that victims may feel uncomfortable addressing because they are not worth the psychological or reputational cost of doing so (Bond & Haynes-Baratz, 2022). Gertsberg (2024) and Amano-Patiño et al. (2025) show that, in light of the #MeToo movement, male economists were less likely to interact and coauthor with female colleagues, limiting career trajectories for female junior academics. Thus, well-intended policies, such as increasing the ease of reporting or the stringency of anti-misconduct mechanisms, may affect superiors' willingness to mentor employees or shift their behavior to less egregious actions that employees have no incentive to report.

The goal of this paper is to understand when — and why — policies that *disincentivize* misconduct may change perpetrators' interactions in ways that harm victims of abuse. We study an organizational model where "managers" have the capacity to commit harmful actions against "employees." Employees experience greater harm the more intense managers' actions are, while managers vary in their proclivity for taking more intense actions. At a cost, employees can report managers' actions as *misconduct* to the organization, which can generate a payout for an employee and a punishment for the manager, with the expected payout being greater (and punishment more severe) if a manager's action is more intense.

We show that policies that discourage managers from committing more intense actions may in fact backfire and make employees worse off, depending on the severity of harm that misconduct generates. We focus on two sets of mechanisms that generate these welfare effects. First, discouraging managers from committing misconduct leads them to not interact with employees in the first place. Interacting with a manager and producing some match value at the risk of harm may be preferred by an employee to no interaction at all *if* the marginal disutility from that harm is low. Second, managers may moderate the severity of their actions to avoid being punished, for example, by switching from flagrant and clear harassment to subtle microaggressions that are harder for employees to prove as abusive. On the one hand, these actions harm employees less. On the other hand, it may be harder for employees to prove that these actions are harmful —disincentivizing reports and hence the possibility of compensation.

We show that these dueling welfare effects run through changes to the value of a employee-manager match, the cost of reporting, and the degree to which managers are punished for deviant behavior. In particular, we show that it may not be optimal for employers to punish managers as much as possible, and that optimal punishment may be interior or at the lower bound for possible punishments. We also show how the establishment of precedents for what firms consider misconduct — such as decisions in past cases — can generate dynamic shifts from equilibria where employees report misconduct to "steady-state" equilibria where managers commit only microaggresions that no one is willing to report, leaving employees altogether worse off than initially.

Model Overview Our model considers a set of managers and employees in an infinite time-horizon environment. Each period, a manager and employee pair are born and active for only that period. In that time, managers can choose whether to interact with an employee — generating a match value V — and commit an action a_t. No interaction yields an outside option for both the employee and manager. Managers' actions generate a disutility $h \cdot a_t$ for employees for $h > 0$. Managers are heterogenous, and are characterized by a bliss point or "type" b which represents a "personality type" or affinity for committing more intense actions. Managers would like to take actions a_t closer to their bliss point.

Upon experiencing an action, employees choose whether to report the action at cost c, triggering an investigation. To focus on the role of *incentives*, we abstract from informational frictions and simply assume that employees receive an expected payout that increases in the intensity of a manager's actions a_t, while managers' expected punishment, conditional on being reported, is also increasing in their action. This can represent a deterministic process — sure punishment that is increasing in the intensity of an action — a probabilistic process — where punishment/payout is more likely if a manager commits a more intense action — or a combination of both.

The organization possesses a *precedent* (or norm) for what they consider misconduct, denoted A_t, such that actions above A_t result in a sure and large punishment and, thus, are always discouraged. Employees' expected payout from reporting an action a_t when the norm is A_t is given by a general functional form $p(a_t, A_t)$ that is increasing in a_t and decreasing in A_t. Managers' expected punishment is given by $\gamma \cdot p(a_t, A_t)$, where $\gamma > 0$ represents the relative *severity* of the organization's punishment. We also comment on how our environment can more broadly be construed as an agency model through the lens of labor relations and overtime work, political accountability, or even consumer choice.

We first show that employees report an action as misconduct if and only if it exceeds a reporting threshold $\overline{r}_t$, which depends on the organization's *norm* A_t. Then, we show that manager behavior follows one of two equilibrium cases. In both, managers with blisspoints below the reporting threshold $\overline{r}_t$ simply play their preferred action and remain unpunished. Managers with blisspoints slightly above $\overline{r}_t$ play actions right at the reporting threshold $\overline{r}_t$, keeping employees indifferent to reporting. We refer to actions right at (or below) the reporting threshold as *microaggressions* — harmful behavior that employees do not find worth reporting. From here, behavior bifurcates. In one equilibrium, managers with high blisspoints opt out of interaction entirely. In the second, managers with moderately high blisspoints play "interior actions" that are slightly above the reporting threshold, which employees do report and managers might be punished for, trading off the value of playing their blisspoints with the risk of potential punishment. We denote these "interior actions" $a_t^\dagger$. Managers in the second equilibrium with very high blisspoints opt out of interaction.

We then study how changes in the models' parameters — which represent policy changes that disincentivize misconduct — affect employee welfare. We broadly highlight three mechanisms that influence employees' expected utility, depending on the *marginal* severity of harm h that employees face upon interaction with managers. The first mechanism relates to switches between participation and no participation. If h is relatively small, employees may prefer interacting with a manager — receiving a match value V and risking harm — to the outside option of no interaction. We show that this mechanism is in effect when the match value of interaction V increases; if h is relatively small, increasing V encourages manager participation from a larger range of types b. If h is large, the effects on welfare are potentially ambiguous, since the increase in V for employees is counterbalanced by the harmful participation of managers with high values of b.

The second mechanism considers switches from interior actions $a_t^\dagger$ to the reporting threshold $\overline{r}_t$. If a type b manager switches from playing an interior action to the (lower) reporting threshold $\overline{r}_t$, this generates two effects. On the one hand, $\overline{r}_t$ generates less harm for employees. On the other hand, employees go from having a strict incentive to report actions (and receive a payout) to being indifferent to reporting and not reporting the action as misconduct. If the size of harm h is small, the latter effect dominates and employees are worse off; otherwise, employees are better off.

We show that an increase in the reporting cost c captures the effects of both the first and second mechanisms. Increasing the cost of reporting first encourages the participation of a broader range of managers; the utility from matching with these managers is higher if h is small. However, even if h is large, interacting with certain types of managers could improve employee utility. An increase in c encourages some managers to switch from interior actions $a_t^\dagger$ to the reporting threshold $\overline{r}_t$, which can improve expected utility precisely if h is large. This creates a tension between the negative effect of the first mechanism and the positive effect of the second. These behavioral effects operate in tandem with the negative, mechanical effects of costlier reporting.

Our third mechanism is related to the second, and considers decreases in the intensity of interior actions $a_t^\dagger$. If a manager goes from playing one interior action $a_t^\dagger$ to a less intense interior action $a_t^{\dagger'}$ that is *still* above the reporting threshold, employees again deal with the two effects above. If h is small, this switch leaves employees worse off, and if h is large, they are better off.

We use this mechanism, in tandem with the first two, to analyze how increasing the severity of manager punishment γ affects employee welfare. This change has three effects. First, certain managers switch from playing an interior action $a_t^\dagger$ to the reporting threshold, in line with the second mechanism. Second, managers who play $a_t^\dagger$ moderate their actions in line with the third mechanism. Third, managers with high blisspoints b opt out of interaction altogether. These effects are welfare-decreasing for employees if h is small, welfare-improving for h large, and ambiguous if h is intermediate. We use the combination of our three mechanisms to explore the optimal punishment γ for managers, i.e. that which maximizes employees' expected utility. When the size of harm h is small, we show that the optimal value of γ is at its lower bound. When it is large, the organization should make γ as large as possible. When h is intermediate, an intermediate level of punishment may maximize employees' expected utility.

Finally, we analyze the dynamics of learning on welfare by exploring how decreases in A_t — the organization's understanding of misconduct — affect employee expected utility. We interpret $p(a_t, A_t)$ as the *probability* with which a reported action results in discipline of a manager, and assume that if some past action was punished by the organization, that action must surely be punished in the future as well, thereby generating a decrease in A_t. We show how this interepretation connects to a *microfoundation* where employees have a *private* threshold of what they consider misconduct, and thus how this process may also represent organizational *learning* about what employees find particularly harmful. Thus, when an action a_t is committed and verified as misconduct, the organization and managers know with certainty that any action greater than a_t will surely be punsihed in the future, and the upper bound for what the organization considers misconduct moves to $A_{t+1} = a_t$.

We show that in this setting, the model can converge to a steady state, in the sense that a sequence of A_t values can converge to a value A^* that remains stationary for the rest of time. We then use the tessellation of our three mechanisms to show how this convergence may negatively affect manager participation, the expected impact of participating managers' actions, and thus how the employee welfare may be worse off than prior.

Our primary motivation for studying "misconduct" in organizations is driven by a growing interest in workplace harassment and abuse — particularly sexual harassment — in both the mainstream press and scholarly literature. Sexual harassment can have negative effects on victims' willingness to be hired at or remain in firms (Adams-Prassl et al., 2024) and pursuit of leadership positions (Folke et al., 2020). Harassment increases abseentism and job turnover and decreases productivity and job satisfaction (Hersch, 2015). Studies such as (Hersch, 2018) have calculated statistical values of sexual harassment for women, arguing that its estimates are above the maximal payouts available for victims under American federal law. This research is complemented by an organizational literature that lays out the startling frequency of harassment across a wide array of firms, and how both toxic organizational climates and vertical relationships where harassers have managerial power over victims exacerbate the potential for abuse. Cortina & Areguin (2021) reviews this literature and argues that formal complaint processes often make matters worse for victims, whether by failing to end harassment, triggering retaliation, or putting victims through further psychological stress. To this end, our paper aims to provide a deeper understanding of how policy changes that may be traditionally thought to improve prospective victims' welfare — more effective reporting, lower reporting costs, or increased punishment for managers — may backfire and hurt them through changes in *perpetrators' behavior*.

These empirical insights have paved the way for a theoretical literature on incentives for committing and reporting harassment (as well as other behaviors like corruption) in organizational settings. Our model is most closely related to those of Lee & Suen (2020) and Cheng & Hsiaw (2022), where employees report private experiences with managers who are heterogenous in their propensity for harassment. The latter's model — like ours — argues that disincentivizing harassment may make employees worse off by discouraging manager participation or "mentoring." However, both of these papers focus on the role of coordination problems and verifiability of information in shaping victim reporting in settings where serial harassers are more likely to be punished if more victims come forward. By contrast, our model assumes away these reporting frictions to focus on how, even when reporting may be efficient, policy changes affect managerial *incentives* to commit harm or even interact with employees, potentially making employees worse off. To this end, our paper differs from an emerging literature on how informational frictions — such as coordination problems, false reports, or unverifiable evidence — affect incentives for reporting bad behavior (Chassang & Miquel, 2019; Bac, 2018; Zhu, 2024; Siggelkow et al., 2018).

Our theoretical environment also connects to a long-standing literature on the economics of crime deterrence, first pioneered by Becker (1968) and reviewed by Chalfin & McCrary (2017). This literature often suggests that more stringent punishment reduces crime, which is balanced against the costs of law enforcement or punishing innocents. However, viewing crime prevention through the lens of our model suggests that increased punishment may have more subtle effects — namely, by discouraging manager participation and encouraging the pursuit of actions that are not worth reporting. This means that a moderate level of criminal punishment may be socially optimal, even when law enforcement itself is costless and there is no risk of type-I error.

The structure of the paper is as follows. Section 2 lays out the basic structure of the model — as well as its applications to other organizational, political, or behavioral settings. Section 3 characterizes its static equilibrium. We then analyze how policy changes to the model parameters affect employees' expected utility in Section 4. Section 5 looks at the dynamic effects of reporting on employee welfare, and Section 6 concludes.

2 Setup

Agents and Actions Consider an organization composed of managers (M) and employees (E). Time is discrete beginning at $t = 1$. Each period, a new manager-employee pair (m, e) is born and is active only for that period.

During that period, the manager m (she) chooses whether to interact with an employee e (he), where interaction generates a symmetric benefit $V > 0$ for both parties. V represents the baseline value of mentorship or a project that the manager and employee carry out together. If m decides not to interact, both agents receive 0 and we move to the next period.

Managers are heterogeneous; each m is characterized by a bliss point $b \sim F(b)$ with support on $[0, A_0]$. If m decides to interact, m takes some action $a \in [0, A_0]$ for $A_0 > 0$, where a represents m's behavior towards e. We denote by a_t the action of a manager m at time t. We refer to b as a manager's *type*. Upon taking action $a_t = a$, a type b manager receives a quadratic loss $(a - b)^2$.

Employees, on the other hand, are homogeneous. Given an employer's action a_t, they receive a disutility $h \cdot a_t$. The employee e can choose to report an action as *misconduct* to the organization. Reporting incurs a cost $c > 0$ to e. To abstract from misreporting and focus on the role of *incentives* in shaping manager and employee behavior, we assume that reports are perfectly verifiable, meaning that misreporting an action would simply entail a cost. A (truthful) report results in the punishment of a manager m, who receives an expected loss $\gamma \cdot p(a_t, A_t)$ for $0 < A_t \leq A_0$ and $\gamma \geq 0$. We call A_t the organization's *precedent* for misconduct. We assume $p(a_t, A_t)$ is

- continuous in (a_t, A_t);
- differentiable for $a_t < A_t$;
- strictly increasing in a_t and decreasing in A_t for $a_t < A_t$;
- $p(a_t, A_t) = 1$;
- and $p(0, A_t) = 0$.

For example, $p(a_t, A_t)$ could be $\min\{\frac{a_t^2}{A_t^2}, 1\}$.

Employees' expected payouts $p(a_t, A_t)$ from reporting encompass two sets of interpretations. First, $p(a_t, A_t)$ could represent the outcome of a process where an employee receives a payout of 1 with probability $p(a_t, A_t)$. Actions closer to A_t — the organization's *precedent* — are more likely to be punished, while actions further from the precedent may require the collection and interpretation of novel evidence. Using our example above, an employee could receive a payout of 1 with probability $\frac{a_t^2}{A_t^2}$. Second, since $p(a_t, A_t)$ employees may receive a payout whose value is increasing in the disutility of a manager's action, which is capped at A_t. Continuing the example above, employees could receive a sure payout of $\frac{a_t^2}{A_t^2}$ for reporting an action below A_t.[1] Managers' expected disutility from employee reporting can be interpreted likewise.

The key assumption for our results is that employees' expected utility from reporting and managers' expected disutility is *increasing* in the intensity of managers' actions and decreasing in the organization's precedent (or, equivalently, increasing in its stringency). In Section 5, we utilize the former probabilistic interpretation of reporting to study how past, successful reports of misconduct can generate *changes* in organizational precedents, resulting in decreases in A_t as t progresses.

We additionally assume $c < 1$ so that the payout from reporting managers who commit actions above the precedent A_t are always worth the cost. We also assume $\gamma \geq V$ so that managers have

[1] Both interpretations can also hold at the same time. An employee could receive a payout of $\min\{\frac{a_t}{A_t}, 1\}$ with probability $\min\{\frac{a_t}{A_t}, 1\}$, which gives the functional form above.

no incentives to commit actions that are clearly more intense than the precedent.

Thus, the baseline expected utilities for each agent, conditional on interacting, are:

$$\begin{aligned} &\text{Manager } m: && V - (a_t - b)^2 - \gamma p(a_t, A_t)\mathbf{1}[m \text{ reported}] \\ &\text{Employee } e: && V - ha_t + \big(p(a_t, A_t) - c\big)\mathbf{1}[m \text{ reported}]. \end{aligned}$$

Comments and Interpretations The motivating interpretation for our model is that of *harassment* in organizations. We interpret managers' blisspoints as their proclivity for taking more intense (and potentially offensive) actions due to their "personality type." Organizations often institute reporting procedures for employees; these procedures are costly — for example, due to the psychological burden of coming forward with allegations — and trigger investigations into managers' behavior. As in our model, these investigations are shaped by organizations' precedents for harassment based on internal guidelines, norms, or laws. Punishment (and hence employee compensation) is often more certain severe if those actions are more egregious and, crucially, clearer violations relative to precedent.

Correspondingly, the match value V generated by interaction is orthogonal to the intensity of managers' actions. This is motivated by the fact that managers' actions as a function of their personality type do not have an *inherent* impact on a project's value, and instead impact employees' feelings of discomfort or safety. However, our central mechanisms and welfare results would be identical if managers' and employees' match values were given by a strictly increasing function $V(a_t)$ and would simply generate additional inefficiencies that would operate on top of our existing mechanisms. In this case, we can also view the model through the lens of overtime work and burnout, which we comment on below.

Next, note that employees' marginal disutility from an action a_t is always h. A key feature of our results will be that the welfare effects of policy changes depend on the exact value of h. Allowing for employee heterogeneity in h would simply add a distributional dependence to these results. Relatedly, an organization may only be able to condition rewards and punishments on managers' actions. That is, they may be unable to verify (and hence compensate) the precise disutilities that employees face from identical actions, which can depend on employees' personal discomfort or the idiosyncratic retaliation they face for experiencing harassment. [2] What will matter for our main results is that employees' expected payouts from reporting (and managers' expected losses) are increasing in the intensity of managers' actions, and that employees' willingness to report is thus a function of the intensity of a manager's action.

Additionally, notice that employees themselves face no incentive-compatibility constraints. However, as we will see, many of the inefficiencies that arise in our setting occur when the disutility from misconduct h is relatively small, i.e. precisely when an incentive-compatibility constraint would not bind. Thus, providing employees an outside option would not eliminate the central mechanisms driving our results.

Finally, we provide several interpretations of the model through the lens of organizations, political institutions, and consumer behavior. Although harassment is our motivating example, misconduct in organizations can take many different forms; to achieve a collaborative goal, managers may have idiosyncratic disregard for health and safety regulations, or even engage in fraud and academic dishonesty. For example, the model can be used to analyze overtime work and burnout. Managers can force employees to work overtime on projects, but excessive overtime work may be reportable as workplace abuse by employees. Thus, the model captures managers' willingness to push employees beyond their limits and employees' willingness to entertain the potentially resultant labor abuse. As suggested above, the employee-manager match value in this setting may also be increasing in the intensity of the manager's action a_t. However, this would simply add an additional layer on top of our existing mechanisms.

The model can also describe a setting of political accountability. Consider a political executive (manager) who can use her authority as leader to achieve policy objectives — such as by issuing executive orders that override the oversight of the legislature. Employees in this setting can represent political constituents or the judiciary, who can choose to investigate the executive's actions and put a stop to them. Their ability to hold the executive accountable is a function of the egregiousness of the executive's actions relative to legal precedent, and their disutility may reflect ideological tolerance for an executive's actions. To this end, our paper is related to formal models of the judiciary such as Beim et al. (2014) and Patty & Turner (2021), which study how the

[2] One of our key results will be that policy changes which discourage harassment often harm employees if h is small relative to the marginal expected payout from reporting an action. Thus, pegging the marginal payout above *all* employees' h values would exacerbate the inefficiencies our model speaks to.

review of evidence, the threat of whistleblowing, and the degree of preference alignment affect the judiciary's efficacy in holding others accountable.

Finally, the model can also describe brand affinity and consumers' tolerance for price increases. Consider a firm (manager) selling a product to a set of consumers (employees). The firm has the option to mark up the price of its product. While consumers may stomach small price hikes, they may have a distaste for excessive price hikes and boycott the company's product if markups exceed a threshold.

3 Static Equilibrium

Since managers and employees are short-lived, we consider static equilibria of the game and the implications this generates for the organization. A pure strategy static equilibrium for an interaction between a manager and employee at time t is given by a profile $\{(i_t^*(b), \alpha_t(b)), r_t^*(a_t)\}$, where $i_t^*(b) \in \{I, NI\}$ is the decision of a bliss point b manager to interact/not interact, $\alpha_t(b)$ is the action taken by a type b manager, conditional on interacting, and $r_t^*(a_t) \in \{R, NR\}$ is the decision of the employee to report (R) or not report (NR), conditional on experiencing a_t. Because agents live only a single period and employees make choices after managers, managers' and employees' equilibrium strategies are identical across time.

Employee's Problem First, we write the employee's maximization problem, conditional on experiencing an action a_t:

$$U^E(a_t) = \max \begin{cases} V - ha_t - c + p(a_t, A_t) & R \\ V - ha_t & NR \end{cases}$$

The first line represents the employee's utility if he chooses to report at cost c, which generates an expected reward below A_t of $\frac{a_t}{A_t}$. Not reporting simply involves the match utility with a loss of ha_t.

Lemma 1 shows that reporting follows a threshold rule: there exists a threshold $\overline{r}_t$ such that the employee has a strict incentive to report misconduct if and only if $a_t > \overline{r}_t$.

Lemma 1. *Reporting follows a threshold rule: the employee has a strict incentive to report an action as misconduct if and only if $a_t > \overline{r}_t$. Moreover, $\overline{r}_t \le A_t$ and is increasing in c and A_t.*

We refer to $\overline{r}_t$ as the *reporting threshold* at time t. This threshold is known to managers.

Given that the employee reports if and only if $a_t > \overline{r}_t$, a type b manager's maximization problem, conditional on interacting, is:

$$U^M(b) = \max_{a_t} \begin{cases} V - (a_t - b)^2 & a_t \le \overline{r}_t \\ V - (a_t - b)^2 - \gamma p(a_t, A_t) & \overline{r}_t < a_t \le A_t \\ V - (a_t - b)^2 - \gamma & a_t > A_t \end{cases}$$

Taking action a_t below the threshold $\overline{r}_t$ will never be reported, and so utility is simply a function of V and the disutility of deviating from the bliss point. Taking an action greater than $\overline{r}_t$ additionally entails the expected costs of punishment. For intermediate actions $a_t \in (\overline{r}_t, A_t]$, the manager's expected loss is $\gamma p(a_t, A_t)$; and above that, her loss is $\gamma \ge V$.

The following theorem characterizes employee and manager equilibrium behavior in the static game.

Theorem 1. *Equilibrium behavior in the static game at time t is characterized as follows. The employee reports an action a_t if and only if $a_t > \overline{r}_t$. There exist thresholds $\underline{a}_t$, $\underline{i}_t$, and $\overline{a}_t$ such that a type b manager's behavior, in equilibrium, is characterized as follows:*

- *If $b \in [0, \overline{r}_t]$, then the manager interacts and $\alpha_t(b) = b$.*

- *If $b \in (\overline{r}_t, \min\{\underline{i}_t, \underline{a}_t, A_0\}]$, the manager interacts and plays $\alpha_t(b) = \overline{r}_t$.*

From here, equilibrium behavior bifurcates into one of two cases.

- ***Equililbrium NR**: if $\underline{i}_t < b \le \min\{\underline{a}_t, A_0\}$, the manager does not interact for all $b \ge \bar{i}_t$. We will refer to this as E^{NR}.*

- ***Equililbrium R**: if $\underline{a}_t < \min\{\underline{i}_t, A_0\}$, the manager interacts and plays $\alpha_t(b) = a_t^{\dagger}(b) > \overline{r}_t$ for all $b \in (\underline{a}_t, \min\{\overline{a}_t, A_0\}]$. If $b > \overline{a}_t$, the manager never interacts. We will refer to this as E^R.*

Moreover, $\underline{a}_t, \overline{a}_t$, and $\underline{i}_t > \overline{r}_t$; and, if $\underline{a}_t > \underline{i}_t$, then $\overline{a}_t > \underline{a}_t$.

On the end of the employee, an action a_t is only worth reporting if it is sufficiently egregious. His decision to report is hence given by a simple threshold rule.

A manager's optimal behavior follows a richer cutoff structure. We begin by considering action choices, conditional on interacting. Those with type $b < \overline{r}_t$ simply play their bliss point, since there is no cost to doing so: the action will go unreported, and thus unpunished, with certainty. This delivers the first cutoff for manager actions: the reporting threshold itself.

A manager with $b \geq \overline{r}_t$ has two options. She can play the reporting threshold, which is the non-reportable action that minimizes deviation from her bliss point. Or, she can play some $a_t > \overline{r}_t$, which is an action that would be closer to her bliss point, but would entail punishment in expectation. We refer to the optimal action that takes into account the risk of punishment as an "interior action", and we denote it by $a_t^{\dagger}(b)$.

For managers with b sufficiently close to $\overline{r}_t$, avoiding punishment dominates the cost of an action further from the bliss point. Thus, there exists an interval of b's, with lower bound $\overline{r}_t$, who bunch at the reporting threshold. We refer to managers' actions at or below the reporting threshold as *microaggressions*; these are actions that ostensibly harm employees and disrupt their work environment, but are not egregious enough for them to report.

For managers with b sufficiently far from $\overline{r}_t$, the cost of deviating to $\overline{r}_t$ dominates, so they play an interior action that internalizes the risk of punishment. This gives the second threshold for manager actions: the type that is indifferent between the reporting threshold and an interior action, denoted $\underline{a}_t$.

We then characterize a manager's choice to interact altogether. This decision depends on two additional thresholds: the type that is indifferent between participating and not participating (denoted $\underline{i}_t$), and the type that is indifferent between participating and playing an interior action (denoted $\overline{a}_t$). The ordering of $\underline{i}_t$ with respect to $\underline{a}_t$ and $\overline{a}_t$ plays an important role, as it determines whether reporting occurs in equilibrium.

In particular, two kinds of static equilibria may realize in the model. We will refer to these as a non-reporting equilibrium (E^{NR}) and a reporting equilibrium (E^{R}). The non-reporting equilibrium E^{NR} is illustrated in Figure 1(a). The key feature of this equilibrium is that the participation constraint binds for all managers who would play an interior action, i.e. $\underline{i}_t < \underline{a}_t$. So, managers either participate and play an action that is not incentive compatible to report (the red line), or they do not participate at all. Manager behavior eliminates reporting, and conditional on interacting, they will only ever commit microaggressions against employees.

The reporting equilibrium E^R is illustrated in Figure 1(b). In contrast to E^{NR}, the participation constraint does not bind for some interval of b who would play an interior action (the blue line), i.e. $\underline{i}_t \in (\underline{a}_t, \overline{a}_t)$. As a result, actions that are incentive compatible to report are played with positive probability. This, in turn, generates reporting in equilibrium.

Finally, we can establish conditions on parameters that characterize whether the equilibrium is E^R or E^{NR}.

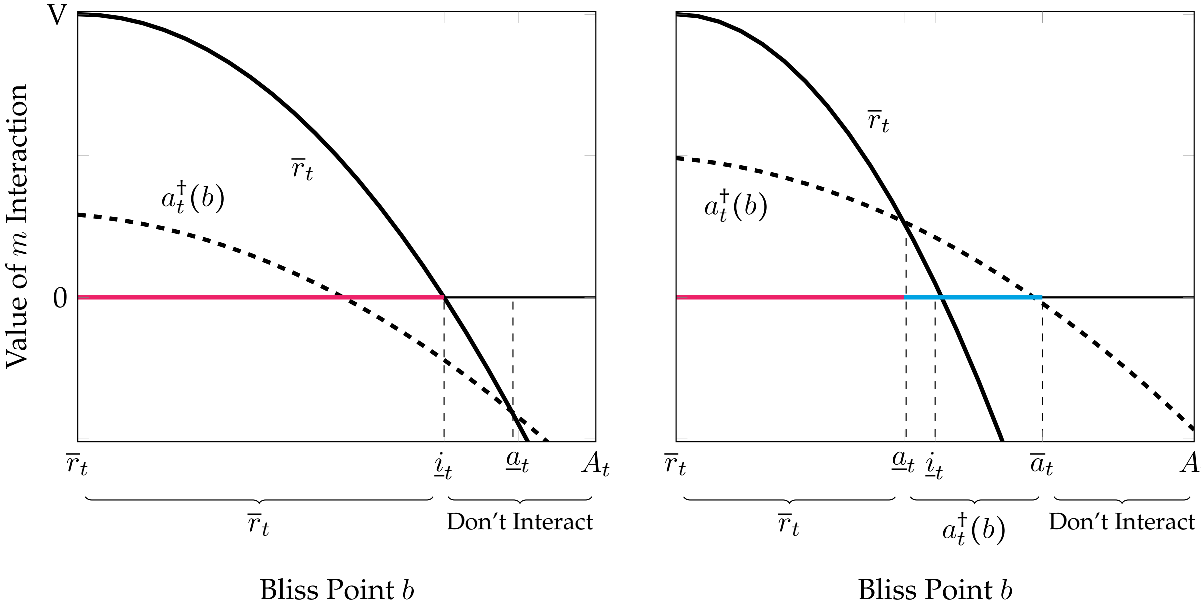
Corollary 1. *The static equilibrium is E^R if and only if V is sufficiently large or γ or c are sufficiently small. In particular, there exists $\overline{\gamma}$ such that if $\gamma \geq \overline{\gamma}$, the equilibrium is E^{NR}.*

Intuitively, when γ is sufficiently large, the manager's expected payoff from an interior action is negative, and so she will not play one. In fact, she will not participate at all: since the type that is indifferent between the reporting threshold and an interior action receives negative payoff from both, so must all b more extreme.

Note that the participation constraint binds for some $b \in (\overline{r}_t, A_0]$, or binds for no one.[3] If it binds for no one, either equilibrium is possible, depending on the values of V and γ. To highlight the main forces of the model, we will focus on the case where it binds for some $b \in (\overline{r}_t + \epsilon, A_0]$.

[3] It binds for $b \in (\overline{r}_t + \epsilon, A_0]$, where $\epsilon > 0$, when either $\underline{a}_t < \underline{i}_t$ and $\overline{a}_t < A_0$ or $\underline{i}_t < A_0, \underline{a}_t$. Note that when $V > 0$, there will always exist some interval of b (with lower bound $\overline{r}_t$) who receive positive utility from playing the reporting threshold. This ensures that the participation constraint can start to bind only for some $b > \overline{r}_t$.

Figure 1: Value of Manager Interactions: Equilibrium NR (E^{NR}) vs. Equilbrium R (E^{R})



(a) E^{NR}: $\underline{i}_t < \underline{a}_t$

(b) E^{R}: $\underline{a}_t < \underline{i}_t < \overline{a}_t$

Threshold Comparative Statics The following proposition summarizes comparative statics of the main thresholds and actions.

Proposition 1. *Comparative statics of the main equilibrium actions for the employee and manager are as follows.*

- *An increase in V generates an increase in $\underline{i}_t$ and $\overline{a}_t$.*
- *An increase in c generates an increase in $\overline{r}_t$, $\underline{a}_t$, and $\underline{i}_t$.*
- *An increase in γ generates an increase in $\underline{a}_t$ and a decrease in $\overline{a}_t$. Moreover, for each b, $a_t^\dagger(b)$ decreases.*
- *A* decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^\dagger(b)$ decreases. The effect on $\underline{a}_t$ is ambiguous.*

Notably, by shifting the ordering of thresholds, changes in manager behavior due to changes in model parameters can change the equilibrium that holds, i.e. induce a switch from E^{NR} to E^{R} or E^{R} to E^{NR}.

4 Welfare

Now that we have characterized static equilibrium behavior, we can study the impact of *policy changes* on employees' wellbeing. In particular, this section shows that policies that disincentivize managers from committing harmful actions may in fact backfire and make employees worse off — either by discouraging managers from interacting with employees altogether, or by causing them to shift their behavior to microaggressions that employees have no incentive to report. That is, policies which may at first glance appear to help employees and other victims of managerial misconduct may inadvertently hurt them.

We first look at the effects of policy changes — represented by changes in match value V, reporting cost c, punishment size γ, or organizational norm A_t — on manager behavior and expected utility. The Corollary below follows from Theorem 1 and Proposition 1.

Corollary 2. *Suppose V increases, c increases, γ decreases, or A_t increases. Then, managers' expected utility increases.*

While these parameter changes have straightforward effects on manager welfare, the effects on employees' welfare are more nuanced. We will show in the value of an interaction V or punishment γ may actually reduce employees' utility in expectation, while increases in the reporting cost c may improve their utility conditional on interacting with certain managers. As a result, even if an organization values both managers' and employees' welfare, changes that negatively impact manager welfare may also negatively impact employee welfare.

We will say that employees' welfare *decreases* if, for every distribution $F(b)$ with support on $[0, A_0]$, employees' expected utility decreases, and if there exists some F such that employees' expected utility *strictly* decreases. This is equivalent to saying that for each $b \in [0, A_0]$, employees' expected utility from interacting with that type either stays the same or *strictly* decreases. We define an increase in welfare analogously. If welfare neither increases nor decreases in the aforementioned sense, we will say that the effect is *ambiguous*. In particular, this means that there exist b and $b' \in [0, A_0]$ such that interacting with a type b manager increases employee utility, while interacting with a type b' decreases utility. Hence, from the perspective of the organization, the overall effect on *expected* utility depends on the distribution $F(b)$.[4]

[4] That is, there may exist a distribution $F(b)$ such that employees' expected utility increases and $G(b)$ where it

First, we write out the indirect utilities of employees in each of the two equilibria.

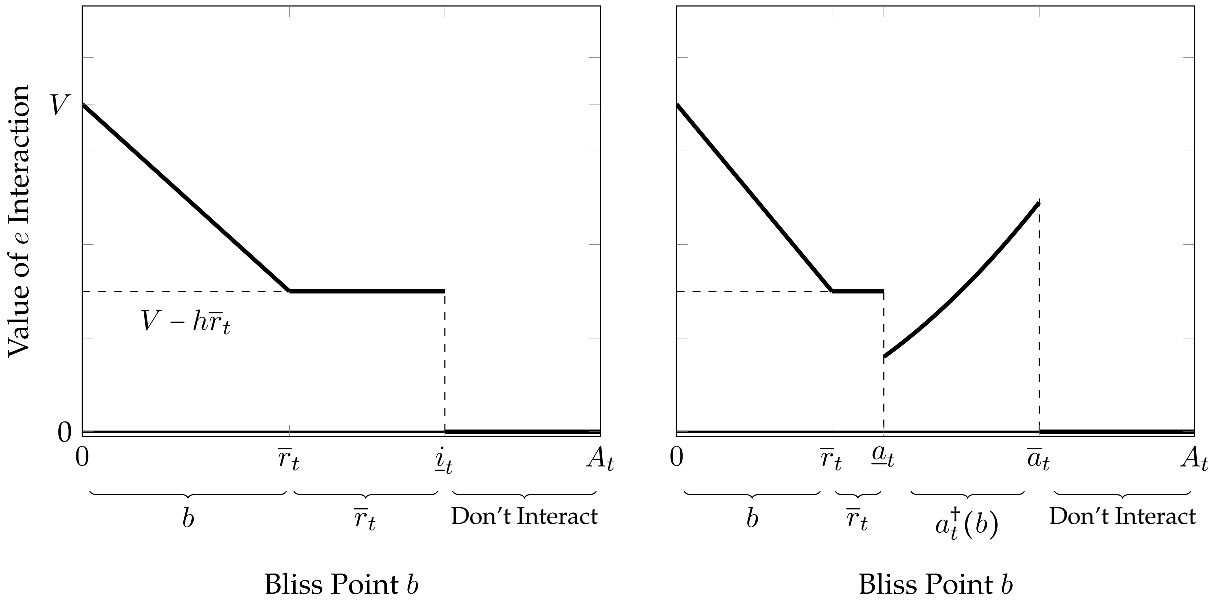
$$U^E_{NL}(b) = \begin{cases} V - hb & b \leq \overline{r}_t \\ V - h\overline{r}_t & b \in (\overline{r}_t, \underline{i}_t] \\ 0 & b > \underline{i}_t \end{cases} \tag{1}$$

$$U^E_{L}(b) = \begin{cases} V - hb & b \leq \overline{r}_t \\ V - h\overline{r}_t & b \in (\overline{r}_t, \underline{a}_t) \\ V - ha^\dagger_t(b) + p(a^\dagger_t(b), A_t) - c & b \in [\underline{a}_t, \overline{a}_t] \\ 0 & b > \overline{a}_t \end{cases} \tag{2}$$

The indirect utilities for the two equilibria are graphed in Figure 2. Both panels are calibrated with $V - h\overline{r}_t > 0$, meaning that matching with a manager and experiencing misconduct is better for employees in expectation than no interaction at all. Calibrating with $V - h\overline{r}_t < 0$ does not change the equilibrium thresholds in the figure, since the value of h does not affect managers' optimal strategy.

decreases.

Figure 2: Value of Employee Interactions: Non-Reporting Equilbrium (E^{NR}) vs. Reporting Equilibrium (E^R)



(a) E^{NR}: $\underline{i}_t < \underline{a}_t$ with $V - h > 0$

(b) E^R: $\underline{a}_t < \underline{i}_t < \overline{a}_t$ with $V - h < 0$

We highlight three mechanisms through which parameter changes affect employee welfare. The first emerges from changes in manager participation. The second and third emerge from changes in whether managers play interior actions and the intensity of these actions.

1. **Mechanism 1: switching between participation and no participation**. An increase in $\overline{a}_t$ or $\underline{i}_t$ induces some previously non-participating managers to participate. This increases welfare if and only if an employee prefers interacting with any newly participating b to not interacting, given the equilibrium (E^{NR} or E^R) after the parameter change. Formally, from (1) and (2), welfare increases if an employee's utility from interacting with type b satisfies:

$$V - h\overline{r}_t \geq 0 \qquad E^{NR} \tag{P1}$$

$$V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \geq 0 \qquad E^R \tag{P2}$$

Decreases in $\overline{a}_t$ or $\underline{i}_t$ result in the opposite effects. This mechanism can be thought of as capturing the decrease in *mentorship* in response to organizations' zero-tolerance policy for misconduct that the literature on harassment has highlighted.

2. **Mechanism 2: switching from interior actions to the reporting threshold**. If $\underline{a}_t$ decreases, or if the initial equilibrium is E^R, managers who previously interacted and played actions $a_t^\dagger(\cdot)$ above the reporting threshold now play the reporting threshold $\overline{r}_t$. From the perspective of the organization, a type b lowering their action to $\overline{r}_t$ is better for employees if and only if

   $$\begin{aligned} V - h\overline{r}_t &\geq V + V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \\ \iff \; \big(a_t^\dagger(b) - \overline{r}_t\big)h &\geq p(a_t^\dagger(b), A_t) - c \end{aligned} \tag{S}$$

   These expressions follow from (2).[5] The reporting threshold constitutes a less egregious action in the first place, as managers switch to committing *microaggressions*. However, switching to the reporting threshold eliminates reporting, and thus the possibility of receiving compensation. For a given b, the organization weighs the expected gain from a lower action against the expected loss from disincentivizing reports.

3. **Mechanism 3: decrease in intensity of interior actions**. Suppose A_t falls to A_t' or γ increases to γ'. Consider a type b manager who played an interior action $a_t^\dagger(b)$ and continues to play an interior action $a_t^\dagger(b)'$. The intensity of the manager's action decreases: $a_t^\dagger(b) > a_t^\dagger(b)'$.

   The tradeoff in terms of welfare is analogous to Mechanism 2. This lower action incurs less expected harm for employees. But, while both actions induce a report, the expected payout is lower. From the organization's perspective, a type b manager switching from $a_t^\dagger(b)$ to $a_t^\dagger(b)'$ is better for employees if and only if

   $$\begin{aligned} -ha_t^\dagger(b)' + p(a_t^\dagger(b)', A_t') &\geq -ha_t^\dagger(b) + p(a_t^\dagger(b), A_t') \\ h &\geq \frac{p(a_t^\dagger(b), A_t) - p(a_t^\dagger(b)', A_t')}{a_t^\dagger(b) - a_t^\dagger(b)'} \end{aligned} \tag{I}$$

   These expressions also follow from (2).

These three mechanisms generate subtleties in how equilibrium welfare responds to changes in

[5] Note also that $p(\overline{r}_t, A_t) = c$. This means that the condition can also be written as $h \geq \frac{p(a_t^\dagger(b), A_t) - p(\overline{r}_t, A_t)}{a_t^\dagger(b) - \overline{r}_t}$, creating a connection with the derivative of $p(\cdot, \cdot)$ with respect to a_t.

the match value of an interaction V, the reporting cost c, and the size of punishment for managers γ.

Changes in Value of Interaction

Proposition 2. *Suppose V increases to V'. If h is sufficiently small, employee welfare increases. In particular:*

- *if we move from E^{NR} to E^{NR} and (P1) holds at V', welfare increases.*
- *If we move from E^{NR} or E^{R}, welfare increases if (P1) holds at V' and (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*
- *If we move from E^{R} to E^{R}, welfare increases if (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}'_t]$.*

Otherwise, welfare changes are ambiguous.

An increase in V does not affect a manager's action, conditional on participating, but does shift the thresholds for participation to the right. This extends the upper bound on which types interact. In regions where managers were already participating, welfare increases: they play the same action, and V is higher. Newly participating managers, however, play actions that are weakly higher than the most extreme action played prior to their entry. So, if h is sufficiently large, these types may generate a negative or ambiguous change in welfare overall — either by committing reportable actions or microaggressions.

The equilibrium may stay in E^{NR} if the increase in V is small, or switch to E^{R} if it is large. However, the equilibrium can never switch from E^{R} to E^{NR}.

Changes in Reporting Costs

Proposition 3. *Suppose c increases marginally to c'.*

- *If we move from E^{NR} to E^{NR}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^{R} to E^{R}, welfare decreases if $h\big(a_t^{\dagger}(b) - \overline{r}'_t\big) \geq p(a_t^{\dagger}(b), A_t) - c$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^{R} to E^{NR}, welfare decreases if $h\big(a_t^{\dagger}(b) - \overline{r}'_t\big) \geq p(a_t^{\dagger}(b), A_t) - c$ for $b \in [\underline{a}_t, \underline{i}'_t]$ and (P2) holds for $b \in [\underline{i}'_t, \overline{a}_t]$.*

Otherwise, changes in welfare are ambiguous.

An increase in c increases the reporting threshold, which has three different effects on welfare. Whether the interaction of these effects increases or decreases welfare depends on the value of h and the value of an interior action $a_t^\dagger(b)$ relative to the new reporting threshold $\overline{r}_t'$.

The first, present in all cases in the Proposition above, is mechanical: increasing reporting costs directly decreases employee welfare in the event of a report.

The second emerges from Mechanism 1: increasing c may crowd in manager participation. If h is sufficiently low, increasing the probability of matching can increase welfare; otherwise, it may not. This is relevant in the first and third cases, as participation cannot change at all when starting and ending in E^R.

The third effect emerges from Mechanism 2, generating shifts from actions reportable as misconduct to microaggressions. If we start in E^R, so that there are managers who play an interior action, then increasing c will induce some of them to play the *new* reporting threshold instead. It is possible that for some of these managers, the new reporting threshold is less intense than their original action, while for others it is more intense. This is because the reporting threshold is higher at $c' > c$. If h is sufficiently large and the old interior $a_t^\dagger(b)$ is still more intense than the new reporting threshold $\overline{r}_t'$, the former dominates the latter in terms of its welfare effect, so welfare increases. The constraint on h in the second bullet point is similar to (S) but additionally internalizes the increase in the reporting threshold.

Mechanism 2 governs whether the equilibrium may change from E^R to E^{NR}. Namely, the new reporting threshold may be sufficiently high that any participating b would rather deviate to it than play a higher, interior action that risks punishment.

Changes with Respect to γ

Proposition 4. *An increase in γ does not change employee welfare if we begin in E^{NR}. If we start in E^R, an increase in γ increases welfare if h is sufficiently large. If h is sufficiently small, welfare decreases. In particular, welfare increases (decreases)*

- *if (S) holds (does not hold) for $b \in [\underline{a}_t, \min\{\underline{a}_t', \underline{i}_t\}]$,*
- *if (I) holds (does not hold) on $[\min\{\underline{a}_t', \underline{i}_t\}, \max\{\underline{i}_t, \overline{a}_t'\}]$,*

- *and if (P2) holds (does not hold) on* $[\max\{\underline{i}_t, \overline{a}'_t\}, \overline{a}_t]$.

Otherwise, welfare changes are ambiguous.

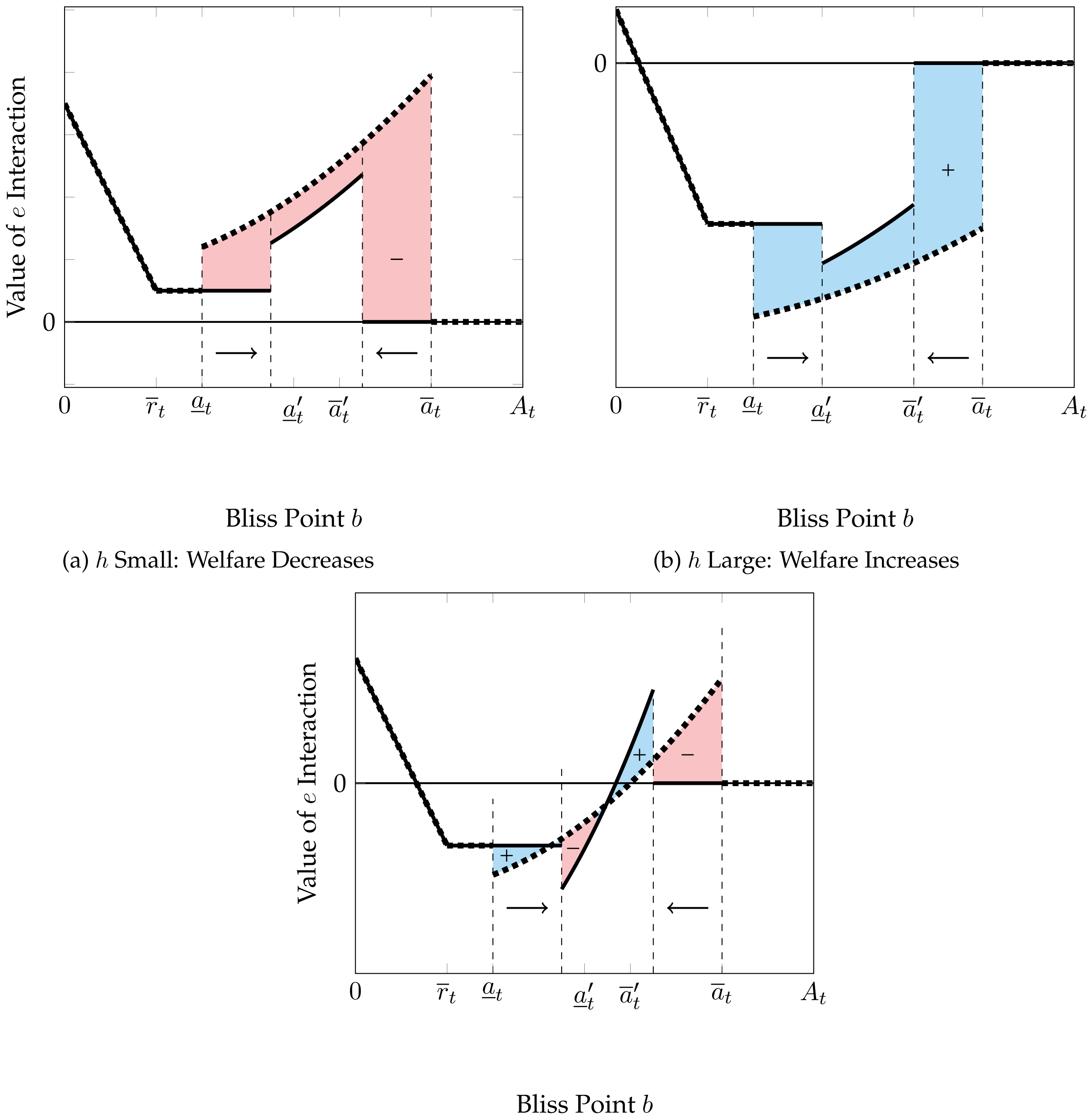
If we begin in E^{NR}, punishment is already large enough that no manager wants to play above the reporting threshold. A fortiori, this will be the case at higher γ as well. So, it is only possible for increasing γ to have an effect if we begin in E^R. When γ increases in E^R, all three mechanisms discussed above are operative, and their overall effect again depends on the value of h.

Figure 3 illustrates welfare changes as a function of h. When γ increases, $[\underline{a}_t, \overline{a}_t]$ shrinks to $[\underline{a}'_t, \overline{a}'_t]$. This triggers all three mechanisms: a range of b to the right select out of participation (Mechanism 1), a range of b to the left now take action $\overline{r}_t$ (Mechanism 2), and the intensity of actions in $[\underline{a}'_t, \overline{a}'_t]$ shifts downwards (Mechanism 3). This generates a benefit in the form of reducing action intensity. But, each mechanism has its concomitant cost: reducing the probability of interaction (Mechanism 1), or reducing the chance of compensation, either through reducing reports by shifting to microaggressions (Mechanism 2) or diminishing incentives to report (Mechanism 3).

Figure 3(a) illustrates the welfare impact of these changes when h is small. The dotted curves indicate the initial equilibrium and the solid the new equilibrium. In this case, there is a strict loss of welfare, shaded in red. Intuitively, since h is small, the match value V is large relative to the disutility of more intense actions, as is the value of preserving the chance of compensation. So, for all three mechanisms, the benefit of reduced action intensity is dominated by the cost.

In contrast, when h is high, the cost of each mechanism is dominated by the benefit of reduced action intensity. This is illustrated in Figure3(b), with the gain shaded in blue.

Figure 3: Effects of Increase in γ in E^R on Equilibrium Utility: h Small vs. h Large



(a) h Small: Welfare Decreases

(b) h Large: Welfare Increases

(c) h Intermediate: Welfare Ambiguous

Finally, 3(c) illustrates the case of intermediate h. In such cases, on some interval(s), h may be sufficiently large to generate a welfare gain, while on others, h is sufficiently small to generate a loss. In the calibration in 3(c), h is large enough that the loss from eliminating reports on $[\underline{a}_t, \overline{a}_t']$ is better than the interior action $a_t^\dagger(\cdot)$ played prior. In the next interval, $[\underline{a}_t', \overline{a}_t']$, h is small enough that decreasing action intensity only improves welfare when b (and so the resulting action) is sufficiently high. In the last interval, $[\overline{a}_t', \overline{a}_t]$, h is sufficiently small that the loss from reduced participation outweighs the benefit of cutting extreme actions.

The value of h characterizes the welfare effects of changes in the size of punishment. The following Theorem synthesizes the patterns in Figure 3, showing that the size of the optimal punishment is increasing in h. When the marginal distuility of managers' actions themselves only have a small effect on employee welfare, managers should be punished as little as possible. When they incur great harm, punishment should be maximal. And in between, the value of optimal punishment can take an interior value.

Theorem 2. *Let $\gamma^*(h)$ be the value of $\gamma \geq V$ that maximizes expected employee utility, as a function of h. $\gamma^*(h)$ is nondecreasing in h. For h sufficiently small, $\gamma^*(h) = V$, and for h sufficiently large, $\gamma^*(h) = \overline{\gamma} > V$.*

5 Dynamics and Non-Learning Steady State

Until now, we have not studied the potentially dynamic impacts of changes in organizational precedents or norms for misconduct. This section studies how organizations' past willingness to punish transgressive managers may generate changes in *precedents* for misconduct, or A_t. In particular, we study how A_t evolves over time and the impact this evolution may have on employee welfare.

To this end, we interpret $p(a_t, A_t)$ as the *probability* that, conditional on being reported, a manager is punished with a loss γ for an action she committed.[6] Suppose an action $a_t < A_t$ is reported at time t, the report was investigated, and the manager was consequently punished. Then, from time $t+1$ onward, any action $a \geq a_t$ must also entail punishment. That is, if a certain behavior was punished today, that behavior must also be punished in the future. Thus, we assume that $A_{t+1} = a_t$. More generally, at time t, let $R(t) \subseteq \{1, 2, \ldots, t\}$ denote the set of time periods t where misconduct was reported and punishment. Then, $A_t = \min\{a_t : t \in R(t) \cup \{0\}\}$, the minimum of all past verified reports. By construction, note that A_t is *decreasing* for all time.

[6] We can also allow this loss to depend on the action itself; for example if $p(a_t, A_t) = \frac{a_t}{A_t}$ below A_t, allowing the punishment to be $\gamma a_t \cdot \frac{a_t}{A_t}$ would not change any of our previous results, since the key assumption is that managers' expected punishment is increasing in γ

This interpretation also relates to a microfoundation, where A_t represents managers' *beliefs* about employees' willingness to tolerate actions. Suppose employees share a common threshold a^* such that they consider an action a_t *misconduct* (and incur disutility) if and only if $a_t \geq a^*$. Thus, a necessary condition for reporting is that $a_t \geq a^*$, and employees report if and only if both $a_t \geq a^*$ and $a_t \geq \overline{r}_t$. However, managers are *uncertain* about where a^* is. If they believe at time 0 that $a^* \sim \mathcal{U}[0, A_0]$, then the probability they are punished for committing an action a_t above the reporting threshold is $p(a_t, A_0) = \min\{\frac{a_t}{A_0}, 1\}$. As reports occur, this prior distribution is repeatedly truncated so that at time t, $p(a_t, A_t) = \min\{\frac{a_t}{A_t}, 1\}$. That is, past misconduct provides *information* about a^*, which is reflected in A_t. [7]

With these interpretations in mind, we show that as actions are reported and verified, the model may eventually converge to a steady state equilibrium. In this steady state, no further revisions of the precedent A_t occur (or, utilizing the microfoundation, organizational *learning* about what employees consider misconduct is halted). This section details conditions for the steady state equilibrium to occur, how employee welfare changes as we converge to this steady state, and assesses the qualitative effects of this process on organizational norms.

Definition A^* corresponds to a *steady state equilibrium* if when $A_t = A^*$, $A_{t+j} = A^*$ with probability 1 for all $j > 0$. E^{NR} is always a steady state, since no changes in A_t can occur (as there is no reporting). Managers either never interact or interact and keep employees precisely indifferent to reporting. E^R is never a steady state, since reporting — and thus changes in A_t — occur with positive probability.

The following proposition establishes that the model converges to a steady state.

Proposition 5. *We converge to a steady state.*

Dynamic Welfare We now analyze the welfare effects of *decreases* in A_t, i.e. how refinement of standards for misconduct can generate reductions in welfare. The following proposition summarizes the welfare effect of a decrease in A_t to A_{t+1}.

[7] Learning in this example involves an assumption that "no news" is uninformative for the organization, while "bad news" in the form of an action a_t being punishment sets a *precedent* that actions at or above a_t will also not be tolerated in the future. The assumption of "no news" being uninformative can be supported by inattention to actions within organizations unless a report "emerges," which may be related to privacy laws protecting employees who have submitted reports of misconduct or harassment.

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}.*

- *If we begin and stay in E^{NR}, welfare increases if h is large, that is, if (P1) does not hold at time t.*
- *If we begin in E^{NR} and move to E^R, welfare increases if h is small. That is, it increases if*
    - *a version of (S) does not hold for $b \in [\underline{a}_{t+1}, \underline{i}_t)$*
    - *and (P2) holds for $b \in [\underline{i}_t, \overline{a}_{t+1}]$ at $t+1$.*
- *If we begin in E^R and stay in E^R, welfare increases if h is large and $\underline{a}_{t+1} > \underline{a}_t$. That is, it increases if*
    - *$\underline{a}_{t+1} < \underline{a}_t$ and $a^{\dagger}_{t+1}(b) \leq \overline{r}_t$ on $[\underline{a}_{t+1}, \underline{a}_t]$ or*
    - *$\underline{a}_{t+1} > \underline{a}_t$ and a version of (S) holds on $[\underline{a}_t, \underline{a}_{t+1}]$;*
    - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$,*
    - *and (P2) does not hold for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*
- *If we begin in E^R and move to E^{NR}, welfare increases if h is sufficiently large. That is, it increases if*
    - *$\underline{i}_{t+1} < \underline{a}_t$ and (P1) does not hold or*
    - *$\underline{a}_t < \underline{i}_{t+1}$ and a version of (S) holds for $b \in [\underline{a}_t, \underline{i}_{t+1}]$;*
    - *and (P2) does not hold for $b \in [\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

A decrease in A_t to A_{t+1} has several effects, depending on the value of b. These are illustrated in Figure 4, for the case where we begin and end in E^R.[8]

We start by considering managers with $b < r_{t+1}$, who play their bliss point regardless, and thus do not impact employee welfare. However, those $b \in [\overline{r}_{t+1}, \overline{r}_t]$ go from playing their bliss point to the new reporting threshold, which is less intense than the old reporting threshold. This results in a strict increase in welfare, since reporting incentives are the same, but managers' actions incur less harm. Similarly, matching with a manager who switches from the old to the new reporting threshold also results in a welfare improvement, since the new reporting threshold is a less egregious action than the old threshold. In the figure, this is illustrated for $b \in [\overline{r}_t, \underline{a}_{t+1}]$.[9]

[8] We choose this case because it illustrates all of the effects generated by changes in A_t, whereas other cases do not.

[9] Note that when A_t changes, it can be that $\underline{a}_{t+1} < \underline{a}_t$, in which case the range would be $b \in [\overline{r}_t, \underline{a}_t]$.

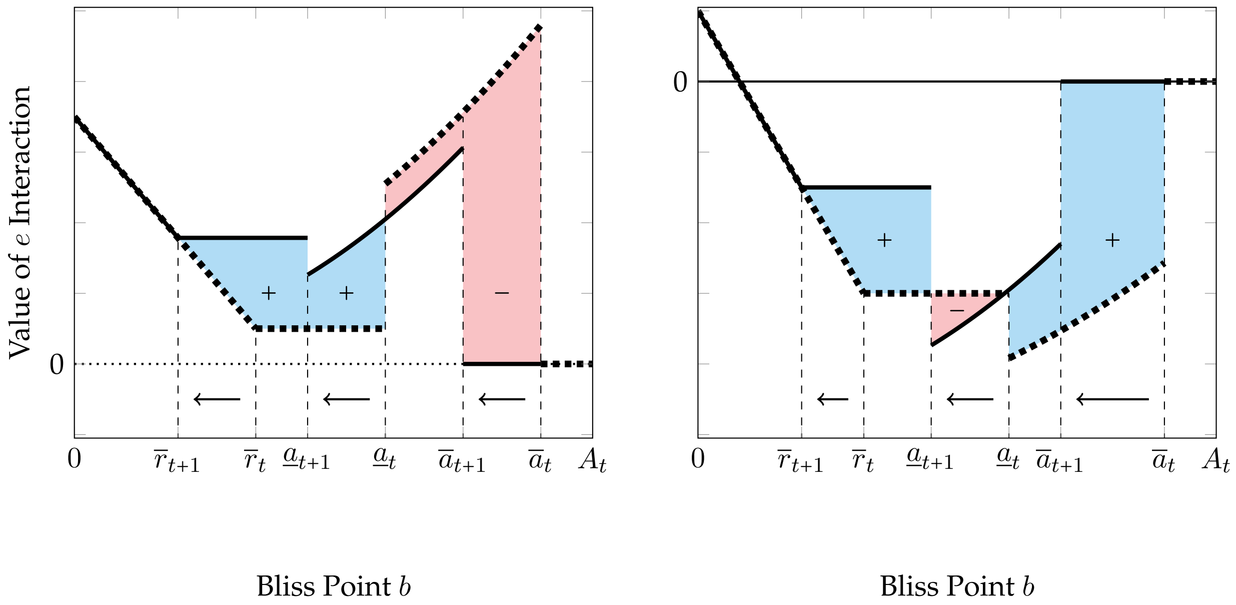
However, despite improvements in welfare conditional on interacting with low b managers, the remaining interval of b's are impacted by the mechanisms discussed earlier. A version of Mechanism 2 operates for $b \in [\underline{a}_{t+1}, \underline{a}_t]$: some managers will shift from an action $a_t^\dagger$ at the old reporting threshold to a (new) interior action $a_{t+1}^\dagger(b) > \overline{r}_{t+1}$. This can increase welfare if h is small (panel a), but can decrease welfare if h is large (panel b).[10] It may also be the case that some new interior actions $a_{t+1}^\dagger(b)$ are less intense than the old reporting threshold $\overline{r}_t$, which will always generate a welfare gain. However, while our calibration shows a case where $\underline{a}_{t+1} < \underline{a}_t$, it may be the case hat $\underline{a}_{t+1} > \underline{a}_t$, which would generate the opposite conclusions.

Mechanism 3 operates for $b \in [\underline{a}_t, \overline{a}_{t+1}]$. Suppose a type b manager takes action $a_t^\dagger(b)$ at A_t and $a_{t+1}^\dagger(b)$ at A_{t+1}. Then, $a_t^\dagger(b) > a_{t+1}^\dagger(b)$. On the one hand, these actions are less intense. On the other hand, reporting generates a smaller benefit in expectation. As in the static case, for h small, welfare decreases (panel a), while it potentially increases for h large (panel b).

Finally, Mechanism 1 operates for $[\overline{a}_{t+1}, A_t]$: managers with high b who previously interacted now no longer interact. As in the static case, this decreases utility if h is small (panel a), and increases if h is large (panel b).

[10] If $\underline{a}_{t+1} < \underline{a}_t$, some managers will shift from the reporting threshold to an interior action. This has the opposite interaction with h.

Figure 4: Effects of Decrease in A_t to A_{t+1} on Equilibrium Utility Beginning and Ending in E^R, $\underline{a}_{t+1} < \underline{a}_t$, h Small vs. h Large



(a) h Small: Welfare Decreases

(b) h Large: Welfare Increases

Convergence to Suboptimal Steady State The proposition above further implies that, when h is sufficiently small, a decrease in A_t that generates a switch from E^R to E^{NR} results can generate a decrease in welfare.

Corollary 3. *Fixing all other parameters, suppose A_t falls to A_{t+1}, such that A_t corresponds to E^R and A_{t+1} to E^{NR}. If h is sufficiently small and $f(b)$ places sufficiently small mass on $b \in [\overline{r}_{t+1}, \min\{\overline{r}_t, \underline{i}_{t+1}]$, welfare decreases.*

Intuitively, the welfare effects of a switch from E^R to E^{NR} are determined by Mechanisms 1 and 2. Some managers switch from interaction to no interaction, while others switch from playing an action $a_t^\dagger$ above the reporting threshold to the new reporting threshold. Both effects decrease welfare when h is small, and will counterbalance any moderation of actions between $\overline{r}_{t+1}$ and either $\overline{r}_t$ or the new interaction threshold $\underline{i}_{t+1}$.

The consequence of this result is that, if E^R is *not* a steady state, learning may result in convergence to a steady state E^{NR} where welfare is *strictly worse* than prior to being in a steady state.[11]

[11] Recall that, from the perspective of the organization, E^R is *never* a steady state. If A_t decreases but we remain in E^R at $t+1$, a necessary condition for a decrease in utility given low h is that $\underline{a}_t < \underline{a}_{t+1}$. A sufficient condition for this, by Proposition 3, is that A_t small.

6 Conclusion

We study an organizational model of misconduct. “Managers” commit harmful actions of varying intensity — such as harassment, overwork, or other forms of workplace abuse — against “employees.” Managers consider first whether to interact with an employee. Managers then take an action of varying intensity; their propensity for different actions is heterogeneous, and is characterized by a (randomly-drawn) blisspoint. Employees face greater disutility from more intense actions, which they can choose to report as misconduct. An employee’s expected payout from reporting is increasing in the intensity of a manager’s action, as does a manager’s expected loss (or punishment) from being reported. This disutility can represent a combination of processes by which rewards/punishment for more intense and egregious actions is greater, or where the probability of issuing a ruling in an employee’s favor increases with the intensity of a manager’s action.

We show that policies that disincentive manager misconduct may have ambiguous or negative effects on employees’ expected utility. We highlight three mechanisms that generate these ambiguous effects. First, employees’ expected utility may decrease when policy changes discourage participation. For example, if managers receive a match value upon interacting with an employee, and that match value decreases, managers may be less likely to interact with an employee ex-ante. In the event the marginal disutility from managers’ actions is large, employees may be better off; no interaction is preferred to being interacted with and experiencing misconduct. However, if the marginal disutility is small, employees may lose out when they are less likely to interact with a manager, since they cannot reap benefits like mentorship from interacting with managers.

Second, managers may switch from committing actions that employees have a strict incentive to report to “microaggressions” they are indifferent to reporting. Policy changes such as increasing the magnitude of managers’ marginal punishment or decreasing reporting costs thus have two effects. On the one hand, because managers commit less intense actions, they are less harmful in expectation. On the other hand, employees have less of an incentive to report them and receive compensation. If the marginal disutility from managers’ actions is small, the latter channel dominates, and employees’ expected utility may decrease. Third, managers may continue committing actions that employees have a strict incentive to report, but again decrease their intensity. These can decrease employees’ welfare if the disutility from experiencing misconduct is small.

We use the combination of these three mechanisms to derive the optimal punishment that maximizes employees' expected utility. We show that optimal punishment is non-decreasing in the marginal disutility of managers' misconduct. When the marginal disutility is low, optimal punishment may involve minimal punishment of abusers. When it is large, maximal punishment is optimal. Otherwise, it may lie in between these two extremes.

Finally, we consider a dynamic extension to the model where reports lead to changes in the organization's standards for misconduct. This utilizes a microfoundation where employees possess a (common) private threshold for what they consider misconduct, and incur disutility only if managers' actions are above that threshold. Managers and the organization are uncertain about where this threshold is. However, as misconduct is reported and successfully adjudicated over time, managers learn about where this threshold is, which gradually constrains the set of actions they can "get away with." While more more precise organizational standards may be thought to improve employees' welfare, we show that learning of this sort encompasses all three of the mechanisms above, and may actually decrease employees' welfare. Moreover, we show that this dynamic model always converges to a steady-state equilibrium. In this steady state equilibrium, no learning or adjustment in organizational standards occurs, and misconduct is never reported or punished in equilibrium.

Future directions for this line of research include richer connections to empirical data on harassment and workplace abuse using micro-data on labor turnover, reporting of abuse, and billed hours in organizations. Our model can be applied to settings beyond organizational misconduct. We have commented on its applications to studying consumer boycotts of excessive price hikes or how judiciaries may hold political executives accountable for abuses of power, but they may also apply to broader policymaking contexts of importance to economists. For example, consider a government that taxes its citizens. Citizens tolerate taxes up to a certain threshold, but beyond that threshold, start evading payment. Or, consider a Central Bank that sets interest rates that affect financial markets; markets again tolerate these hikes up to a certain point, but if they are too excessive, this triggers a sell-off. This framework provides a variety of avenues for further theoretical research on the unintended consequences of disincentivizing harmful actions through their effects on perpetrators' *incentives*.

References

Adams-Prassl, A., Huttunen, K., Nix, E., & Zhang, N. (2024). Violence against women at work. *The Quarterly Journal of Economics*, *139*(2), 937–991.

Amano-Patiño, N., Faraglia, E., & Giannitsarou, C. (2025). Economics coauthorships in the aftermath of metoo. *European Economic Review*, *176*, 105020.

Aquino, K., & Thau, S. (2009). Workplace victimization: Aggression from the target's perspective. *Annual review of psychology*, *60*(1), 717–741.

Bac, M. (2018). Wages, performance and harassment. *Journal of Economic Behavior & Organization*, *145*, 232–248.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, *76*(2), 169–217.

Beim, D., Hirsch, A. V., & Kastellec, J. P. (2014). Whistleblowing and compliance in the judicial hierarchy. *American Journal of Political Science*, *58*(4), 904–918.

Bond, M. A., & Haynes-Baratz, M. C. (2022). Mobilizing bystanders to address microaggressions in the workplace: The case for a systems-change approach to getting a (collective) grip. *American Journal of Community Psychology*, *69*(1-2), 221–238.

Chalfin, A., & McCrary, J. (2017). Criminal deterrence: A review of the literature. *Journal of Economic Literature*, *55*(1), 5–48.

Chassang, S., & Miquel, G. P. I. (2019). Crime, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *The Review of Economic Studies*, *86*(6), 2530–2553.

Cheng, I.-H., & Hsiaw, A. (2022). Reporting sexual misconduct in the# metoo era. *American Economic Journal: Microeconomics*, *14*(4), 761–803.

Cortina, L. M., & Areguin, M. A. (2021). Putting people down and pushing them out: Sexual harassment in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 285–309.

Cullinan, J., Hodgins, M., Hogan, V., & Pursell, L. (2020). The value of lost productivity from workplace bullying in ireland. *Occupational medicine*, *70*(4), 251–258.

Dobbin, F., & Kalev, A. (2019). The promise and peril of sexual harassment programs. *Proceedings of the National Academy of Sciences*, *116*(25), 12255–12260.

Folke, O., Rickne, J., Tanaka, S., & Tateishi, Y. (2020). Sexual harassment of women leaders. *Daedalus*, *149*(1), 180–197.

Gertsberg, M. (2024). The unintended consequences of# metoo: Evidence from research collaborations.

Hersch, J. (2015). Sexual harassment in the workplace. *IZA world of labor*.

Hersch, J. (2018). Valuing the risk of workplace sexual harassment. *Journal of Risk and Uncertainty*, *57*(2), 111–131.

Lee, F. X., & Suen, W. (2020). Credibility of crime allegations. *American Economic Journal: Microeconomics*, *12*(1), 220–259.

Patty, J. W., & Turner, I. R. (2021). Ex post review and expert policy making: When does oversight reduce accountability? *The journal of politics*, *83*(1), 23–39.

Siggelkow, B. F., Trockel, J., & Dieterle, O. (2018). An inspection game of internal audit and the influence of whistle-blowing. *Journal of Business Economics*, *88*, 883–914.

Sutton, R. (2007, May 1). Building the civilized workplace. *McKinsey Quarterly*. Retrieved from `https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/building-the-civilized-workplace`

Tinkler, J. E., Clay-Warner, J., & Alinor, M. (2022). Sexual harassment training and men's motivation to work with women. *Social Science Research*, *107*, 102740.

Zhu, J. Y. (2024). Better monitoring... worse outcome? *The RAND Journal of Economics*, *55*(4), 550–572.

Proofs

Lemma 1. *Reporting follows a threshold rule: the employee has a strict incentive to report an action as misconduct if and only if $a_t > \overline{r}_t$. Moreover, $\overline{r}_t \leq A_t$ and is increasing in c and A_t.*

Proof. For $a_t \geq a^*$, the employee has a strict incentive to report if and only if $-c + p(a_t, A_t) > 0$, which occurs if and only if $a_t > \overline{r}_t$, defined implicitly by $p(\overline{r}_t, A_t) = c$. That $\overline{r}_t < A_t$ follows from the fact that $p(A_t, A_t) = 1 > c$, and that $\overline{r}_t$ is increasing in A_t and c follows likewise. □

Lemma 2. *Conditional on interacting, the optimal action $a^*(t)$ for a type b manager is characterized by thresholds $\underline{a}_t$ and $\overline{b}_t$ such that:*

- *if $b \in [0, \overline{r}_t]$, $\alpha_t(b) = b$ (The manager plays her bliss point.)*
- *if $b \in (\overline{r}_t, \min\{\underline{a}_t, \overline{b}_t, A_0\}]$, then $\alpha_t(b) = \overline{r}_t$ (The manager plays the reporting threshold.)*
- *if $b \in (\underline{a}_t \min\{\overline{b}_t, A_0\})$, then $\alpha_t(b) = a_t^\dagger(b) > \overline{r}_t$ (The manager plays an interior action above the reporting threshold but below her preferred action.)*
- *and if $b > \min\{\overline{b}_t, A_0\}$, then $\alpha_t(b) = b$ (The manager plays her bliss point.)*

Proof. Suppose $a_t \leq \overline{r}_t$. In this case, if $b \leq \overline{r}_t$, the manager simply plays her preferred action, $a_t = b$. If $a_t \geq \overline{r}_t$, the action achieving her maximal payoff is defined implicitly by $a_t^\dagger(b)$, which solves:

$$-2(a_t^\dagger(b) - b) = \gamma p_{a_t}(a_t^\dagger(b), A_t) = 0$$

That $a_t^\dagger(b)$ is decreasing in γ and increasing in A_t follows from monotonicity with the equation above. This interior action is better than $\overline{r}_t$ for $A_t \geq a_t \geq \overline{r}_t$ and $A_t \geq a_t^\dagger(b)$ if and only if

$$V - (a_t^\dagger(b) - b)^2 - \gamma p(a_t^\dagger(b), A_t) \geq V - (\overline{r}_t - b)^2 \tag{3}$$

Note that because $a_t^\dagger(b) < b$, both sides of this equation are decreasing in b. By the Envelope Theorem, the derivative of the left hand side is greater than the right hand side if and only if $2(a_t^\dagger(b) - b) \geq 2(\overline{r}_t - b)$, which always holds. Thus, the left hand side decreases in b at a slower rate than the right hand side. Moreover, because $p(\cdot,\cdot) \leq 1$ and because $a_t^\dagger(b) \to \infty$ as $b \to \infty$, there exists a value of b where

$$V - (a_t^\dagger(b) - b)^2 - \gamma = V - (\overline{r}_t - b)^2$$

Thus, by the Intermediate Value Theorem and monotonicity, there exists a unique point $\underline{a_t} > \overline{r}_t$ which satisfies (3) with equality, and such that the inequality is strict for $b > \underline{a_t}$. By (3), $\underline{a}_t$ is increasing in c (via $\overline{r}_t$) and γ.

If $b \leq \min\{\underline{a}_t, A_t\}$, then a type b manager will commit action $\overline{r}_t$. If $A_t > b > \underline{a}_t$, a type b manager will commit action $a_t^\dagger(b)$. Finally, if $a_t^\dagger(b) > A_t$, the optimal action (conditional on interaction) is simply b, since verification of misconduct will result in a sure loss γ, which is worse than the match value V. This switch is given by the indifference point $\overline{b}_t$ defined by $p(a_t^\dagger(b), A_t) = 1$. Hence, for $b > \overline{b}_t$, the manager simply plays her preferred action b.

□

Lemma 3. *The manager's decision to interact, conditional on playing $\alpha_t(b)$ if she does, is characterized by two additional thresholds, $\underline{i}_t$ and $\overline{a}_t$, such that*

- *if $b \in [0, \min\{\underline{a}_t, \underline{i}_t, A_0\}]$, $i_t^*(b) = I$ (the manager interacts);*
- *if $\underline{a}_t < \underline{i}_t$ and $b \in (\underline{a}_t, \min\{\overline{a}_t, A_0\}]$, $i_t^*(b) = I$ (the manager interacts);*
- *if $\underline{i}_t \leq \underline{a}_t$ and $b > \underline{i}_t$, $i_t^*(b) = NI$ (the manager never interacts);*
- *if $\underline{a}_t < \underline{i}_t$ and $b > \overline{a}_t$, $i_t^*(b) = NI$ (the manager never interacts).*

Proof. For $b \in [0, \overline{r}_t]$, the manager is playing her optimal action and is not punished, so she always interacts. For $b \in (\overline{r}_t, \min\{\underline{a}_t, A_t\}]$, the manager plays at the reporting threshold. Interaction is optimal if and only if

$$V - (\overline{r}_t - b)^2 \geq 0$$

$$\overline{r}_t + \sqrt{V} \equiv \underline{i}_t \geq b.$$

Next, if $\underline{a}_t < A_t$ and $b \in (\underline{a}_t, \min\{\overline{b}_t, A_0\}]$, the manager plays an interior action $a_t^\dagger(b)$. Interaction is optimal if and only if

$$V - (a_t^\dagger(b) - b)^2 - \gamma p(a_t^\dagger(b), A_t) \geq 0$$

At $b = \overline{b}_t$, the expression is equal to $V - (\overline{a}_t^\dagger(\overline{b}_t) - b)^2 - \gamma$. Since $\gamma \geq V$ and since this expression is decreasing in b, there exists $\overline{a}_t < \overline{b}_t$ such that there is an incentive to interact if and only if $b \leq \overline{a}_t$. $\overline{a}_t$ solves the above equation with equality. $\overline{a}_t$, by the implicit equation, is increasing in A_t and V and decreasing in γ.

□

Proposition 1. *Comparative statics of the main equilibrium actions for the employee and manager are as follows.*

- *An increase in V generates an increase in $\underline{i}_t$ and $\overline{a}_t$.*

- *An increase in c generates an increase in $\overline{r}_t$, $\underline{a}_t$, and $\underline{i}_t$.*

- *An increase in γ generates an increase in $\underline{a}_t$ and a decrease in $\overline{a}_t$. Moreover, for each b, $a_t^\dagger(b)$ decreases.*

- *A* decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^\dagger(b)$ decreases. The effect on $\underline{a}_t$ is ambiguous.*

Proof. All comparative statics are immediate, with the exception of $\underline{a}_t$ with respect to A_t. This expression is defined implicitly by

$$V - (a_t^\dagger(\underline{a}_t) - \underline{a}_t)^2 - \gamma p(a_t^\dagger(\underline{a}_t), A_t) = V - (\overline{r}_t - \underline{a}_t)^2$$

Note that $a_t^\dagger(b)$ maximizes the left hand side at an interior point. Denote the derivative of $\underline{a}_t$ with respect to A_t as $\underline{a}_t'$ and of $\overline{r}_t$ as $\overline{r}_t'$. By the Envelope Theorem, since $a_t^\dagger(b)$ maximizes the left hand side at an interior point, $\underline{a}_t'$ is described implicitly by the equation

$$2(a_t^\dagger(\underline{a}_t) - \underline{a}_t)\underline{a}_t' - \gamma p_{A_t} p(a_t^\dagger(\underline{a}_t), A_t) = 2(\overline{r}_t - \underline{a}_t)a_t' - 2(\overline{r}_t - \underline{a}_t)\overline{r}_t'$$

which can be rearranged as

$$\underline{a}_t' = \frac{\gamma p_{A_t}(a_t^\dagger(\underline{a}_t), A_t) + 2(\underline{a}_t - \overline{r}_t)\overline{r}_t'}{2(a_t^\dagger(\underline{a}_t) - \overline{r}_t)}.$$

$\overline{r}_t'$ can be described implicitly by

$$p(\overline{r}_t, A_t) = c \implies p_{a_t}(\overline{r}_t, A_t)\overline{r}_t' + p_{A_t}(\overline{r}_t, A_t) = 0 \implies \overline{r}_t' = -\frac{p_{A_t}(\overline{r}_t, A_t)}{p_{a_t}(\overline{r}_t, A_t)}$$

Since $a_t^\dagger(\underline{a}_t) > \overline{r}_t$ and $\overline{r}_t' > 0$, the sign of the derivative is given by the sign of the numerator. Thus, the derivative is positive if and only if

$$\begin{aligned}
\gamma p_{A_t}(a_t^\dagger(\underline{a}_t), A_t) - 2(\underline{a}_t - \overline{r}_t)\frac{p_{A_t}(\overline{r}_t, A_t)}{p_{a_t}(\overline{r}_t, A_t)} &\geq 0 \\
\gamma p_{A_t}(a_t^\dagger(\underline{a}_t), A_t) &\geq 2(\underline{a}_t - \overline{r}_t)\frac{p_{A_t}(\overline{r}_t, A_t)}{p_{a_t}(\overline{r}_t, A_t)} \\
\frac{2(\underline{a}_t - \overline{r}_t)}{p_{a_t}(\overline{r}_t, A_t)} &\geq \gamma\frac{p_{A_t}(a_t^\dagger(\underline{a}_t), A_t)}{p_{A_t}(\overline{r}_t, A_t)}
\end{aligned}$$

where the last line follows since $p_{A_t} < 0$. Thus, the precise sign depends on the shape of $p(\cdot, \cdot)$. □

Proposition 2. *Suppose V increases to V'. If h is sufficiently small, employee welfare increases. In particular:*

- *if we move from E^{NR} to E^{NR} and (P1) holds at V', welfare increases.*
- *If we move from E^{NR} or E^R, welfare increases if (P1) holds at V' and (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}_t']$.*
- *If we move from E^R to E^R, welfare increases if (P2) holds at V' for $b \in [\max\{\overline{a}_t, \underline{a}_t\}, \overline{a}_t']$.*

Otherwise, welfare changes are ambiguous.

Proof. First, conditional on matching with a manager who has $b \leq \min\{\underline{i}_t, \underline{a}_t\}$, managers' actions stay the same, but employees derive a strictly higher match value V', so their welfare strictly increases.

Next, consider E^{NR}. If we remain in E^{NR} — so $\underline{a}_t' > \underline{i}_t' > \underline{i}_t$, managers who previously did not interact on $[\underline{i}_t, \underline{i}_t']$ now interact and keep employees indifferent to reporting misconduct; welfare on $[\underline{i}_t, \underline{i}'_t]$ increases if and only if $V' - ch \geq 0$.

If we switch from E^{NR} to E^{R} — so $\underline{i}'_t > \underline{a}_t$ — then managers on $[\underline{i}_t, \underline{a}_t]$ switch from not interacting to interacting at $\overline{r}_t$. Welfare hence increases for employees if and only if $V' - ch \geq 0$. Manager behavior changes to the interior action on $[\underline{a}_t, \overline{a}'_t]$, meaning welfare increases if and only if $V' - ha^{\dagger}_t(b) + p(a^{\dagger}_t(b), A_t) - c \geq 0$.

If we begin in E^{R}, note first that we never switch to E^{NR}, since $\underline{a}_t$ does not change and $\underline{i}_t$ only increases. Here, welfare strictly increases below $\overline{a}_t$. Managers on $[\overline{a}_t, \overline{a}'_t]$ who did not interact previously now interact and play an action that an employee reports. Welfare on this range increases if and only if $V' - ha^{\dagger}_t(b) + p(a^{\dagger}_t(b), A_t) - c \geq 0$. □

Proposition 7. *Suppose c increases marginally to c'.*

- *If we move from E^{NR} to E^{NR}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^{R} to E^{R}, welfare decreases if $h\big(a^{\dagger}_t(b) - \overline{r}'_t\big) \geq p(a^{\dagger}_t(b), A_t) - c$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^{R} to E^{NR}, welfare decreases if $h\big(a^{\dagger}_t(b) - \overline{r}'_t\big) \geq p(a^{\dagger}_t(b), A_t) - c$ for $b \in [\underline{a}_t, \underline{i}'_t]$ and (P2) holds for $b \in [\underline{i}'_t, \overline{a}_t]$.*

Otherwise, changes in welfare are ambiguous.

Proof. Both $\overline{r}_t$ and $\underline{i}_t$ increase marginally by a factor of A_t. $\underline{a}_t$ increases by a factor larger than A_t; hence, an equilibrium can move from E^{R} to E^{NR}, but never from E^{NR} to E^{R}.

Next, for managers with $b \leq \overline{r}_t$, their actions do not change. However, managers with $b \in (\overline{r}_t, \overline{r}'_t]$ now play their bliss point, which is strictly higher than $\overline{r}_t$, generating a strictly negative shift in employee utility. Suppose we are E^{NR} and remain in E^{NR}. Welfare on $[\overline{r}'_t, \underline{i}_t]$ strictly decreases, since managers on this range are playing at the new reporting threshold $\overline{r}'_t > \overline{r}_t$. Finally, on $[\underline{i}_t, \underline{i}'_t]$, welfare increases if and only if $V - h\overline{r}_t \geq 0$.

Next, suppose we are in E^{R}. Due to a marginal increase, we have either $\overline{r}'_t < \underline{a}_t < \underline{a}'_t < \underline{i}'_t$, in which case we remain in E^{R}, or $\overline{r}'_t < \underline{a}_t < \underline{i}'_t < \underline{a}'_t$, in which case we move to E^{NR}. As above, welfare decreases on $[\overline{r}_t, \overline{r}'_t]$. Welfare on $[\overline{r}'_t, \underline{a}_t]$ decreases; managers are playing at a strictly higher reporting threshold.

If we remain in E^R, welfare on $[\underline{a}_t, \underline{a}'_t]$ welfare increases if and only if the new reporting threshold is worse than experiencing $a_t^\dagger$ under the old c, i.e. if and only if

$$\begin{aligned} -h\overline{r}'_t &\geq -ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \\ h\big(a_t^\dagger(b) - \overline{r}'_t\big) &\geq p(a_t^\dagger(b), A_t) - c \end{aligned}$$

Notice in particular that the sign of $a_t^\dagger(b) - \overline{r}'_t$ is ambiguous; if it is ≥ 0, this holds if h is large, but otherwise, this never holds. Welfare on $[\underline{a}'_t, \overline{a}_t]$ strictly decreases; employees experience the same interior action $a_t^\dagger$ but pay a higher cost to report it.

Finally, in the case where we move to E^{NR}, welfare on $[\underline{i}'_t, \overline{a}_t]$ increases if and only if no interaction is better than the old $a_t^\dagger$, i.e. if and only if $V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \leq 0$.

□

Proposition 4. *An increase in γ does not change employee welfare if we begin in E^{NR}. If we start in E^R, an increase in γ increases welfare if h is sufficiently large. If h is sufficiently small, welfare decreases. In particular, welfare increases (decreases)*

- *if (S) holds (does not hold) for $b \in [\underline{a}_t, \min\{\underline{a}'_t, \underline{i}_t\}]$,*
- *if (I) holds (does not hold) on $[\min\{\underline{a}'_t, \underline{i}_t\}, \max\{\underline{i}_t, \overline{a}'_t\}]$,*
- *and if (P2) holds (does not hold) on $[\max\{\underline{i}_t, \overline{a}'_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

Proof. Consider E^R, where an increase in γ leads to an increase in $\underline{a}_t$ to $\underline{a}'_t$ and a decrease in $\overline{a}_t$ to $\overline{a}'_t$. If we remain in E^R, on $[\underline{a}_t, \underline{a}'_t]$, managers who previously interacted above the reporting threshold and risked potential punishment now interact right at the reporting threshold. This is better for employees if and only if

$$\begin{aligned} V - h\overline{r}_t &\geq V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \\ h\big(a_t^\dagger(b) - \overline{r}_t\big) &\geq p(a_t^\dagger(b), A_t) - c \\ h &\geq \frac{p(a_t^\dagger(b), A_t) - c}{a_t^\dagger(b) - \overline{r}_t} \end{aligned}$$

On $[\underline{a}_t', \overline{a}_t']$, managers continue to interact at an action above the reporting threshold, but that action becomes less intense. This results in a potential decrease to expected utility (harder to report misconduct) but also a potential increase (the action is less egregious to begin with). An increase is better if and only if

$$\begin{aligned} V - ha_t^\dagger(b)' + p(a_t^\dagger(b)', A_t) - c &\geq V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \\ h\big(a_t^\dagger(b) - a_t^\dagger(b)'\big) &\geq p(a_t^\dagger(b)', A_t) - p(a_t^\dagger(b), A_t) \\ h &\geq \frac{p(a_t^\dagger(b)', A_t) - p(a_t^\dagger(b), A_t)}{a_t^\dagger(b) - a_t^\dagger(b)'} \end{aligned}$$

On $[\overline{a}_t', \overline{a}_t]$, managers who previously interacted now do not interact at all. Because employees on this range had a strict incentive to report, this is an improvement if and only if $V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c \leq 0$.

Finally, if we move from E^R to E^{NR}, we are simply in an equilibrium equivalent to having $\underline{a}_t' = \underline{i}_t = \overline{a}_t'$, and can apply the insights from above.

□

Theorem 2. *Let $\gamma^*(h)$ be the value of $\gamma \geq V$ that maximizes expected employee utility, as a function of h. $\gamma^*(h)$ is nondecreasing in h. For h sufficiently small, $\gamma^*(h) = V$, and for h sufficiently large, $\gamma^*(h) = \overline{\gamma} > V$.*

Proof. Note that conditional on being in E^{NR}, welfare does not change with γ, so suppose we are in E^R. An employee's expected utility conditional on γ is given by:

$$F(\overline{a}_t)V + \int_0^{\overline{r}_t} -hbf(b)db + \int_{\overline{r}_t}^{\underline{a}_t} -h\overline{r}_t f(b)db + \int_{\underline{a}_t}^{\overline{a}_t} \big(-ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) - c\big)f(b)db$$

Noting that $\overline{a}_t$ and $\underline{a}_t$ are both functions of γ and writing the derivative of $a_t^\dagger(b)$ with respect to γ as $a_t^\dagger(b)'$, the derivative of this expression with respect to γ is

$$\begin{aligned} f(\overline{a}_t)\overline{a}_t'V - h\overline{r}_t f(\underline{a}_t)\underline{a}_t' &+ \big(-ha_t^\dagger(\overline{a}_t) + p(a_t^\dagger(\overline{a}_t), A_t) - c\big)f(\overline{a}_t)\overline{a}_t' \\ &- \big(-ha_t^\dagger(\underline{a}_t) + p(a_t^\dagger(\underline{a}_t), A_t) - c\big)f(\underline{a}_t)\underline{a}_t' \\ &+ \int_{\underline{a}_t}^{\overline{a}_t} \big(p_{a_t}(a_t^\dagger(b), A_t) - h\big)a_t^\dagger(b)'f(b)db. \end{aligned}$$

Note that $\overline{a}_t'(\gamma) < 0$, $\underline{a}_t'(\gamma) > 0$, and $a_t^\dagger(b)' < 0$. Hence, after reorganizing terms, the derivative's sign can be represented via the following three components:

$$\underbrace{f(\overline{a}_t)\overline{a}_t'V + \big(p(a_t^\dagger(\overline{a}_t), A_t) - c\big)f(\overline{a}_t)\overline{a}_t'f(\overline{a}_t)\overline{a}_t' - \big(p(a_t^\dagger(\underline{a}_t), A_t) - c\big)f(\underline{a}_t)\underline{a}_t'f(\underline{a}_t)}_{\text{Positive}}$$
$$+h\underbrace{\left[a_t^\dagger(\underline{a}_t)f(\underline{a}_t)\underline{a}_t' - a_t^\dagger(\overline{a}_t)f(\overline{a}_t)\overline{a}_t' - \overline{r}_t f(\underline{a}_t)\underline{a}_t' - \int_{\underline{a}_t}^{\overline{a}_t} a_t^\dagger(b)'f(b)db\right]}_{\text{Negative}}$$
$$+\underbrace{\int_{\underline{a}_t}^{\overline{a}_t} p_{a_t}(a_t^\dagger(b), A_t)a_t^\dagger(b)'f(b)db}_{\text{Negative}}$$

Moreover, differentiating the expression above with respect to h yields a positive expression (simply the middle line enclosed by brackets). Thus, by Topkis' Monotonicity Theorem, $\gamma^*(h)$ is nondecreasing in h. Finally, note that at $h = 0$, the sign of the derivative is negative everywhere, meaning the optimal γ is a corner solution at its lower bound, i.e. $\gamma = V$. As h grows arbitrarily large, for all γ, the middle positive term grows arbitrarily large, meaning that the derivative is everywhere positive and we end up at a corner solution with γ as high as possible. This upper bound is defined by the point where $\underline{a}_t = \underline{i}_t$, which gives the second part of the result; for γ larger than this value, welfare does not change, since we remain in E^{NR}, where utility does not change.

□

Proposition 5. *We converge to a steady state.*

Proof. First, if A_t is such that we begin in E^{NR}, we are already in a steady state. Hence, suppose we begin in E^R.

Next, if $A_{t+1} \neq A_t$, then it must be that $A_{t+1} < A_t$; to see this, note that $A_{t+1} \neq A_t$ only if an employee matches with a manager of type $bin[\underline{a}_t, \overline{a}_t]$, which results in an action $a_t^\dagger(b)$ being played; in this case, $A_{t+1} = a_t^\dagger(b) < A_t$. Hence, by the monotone convergence theorem, each sequence of A_ts converges.

Finally, suppose there exists a sequence A_t and value A^* such that $A_t \to A^*$ but A^* is not a steady state. By definition, we must have that A^* corresponds to E^R. But then, since we are in E^R, with positive probability, an action above the reporting threshold is played, which is reported and, with positive probability, punished. This contradicts that A^* was a steady state. □

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}.*

- *If we begin and stay in E^{NR}, welfare increases if h is large, that is, if (P1) does not hold at time t.*
- *If we begin in E^{NR} and move to E^R, welfare increases if h is small. That is, it increases if*
    - *a version of (S) does not hold for $b \in [\underline{a}_{t+1}, \underline{i}_t)$*
    - *and (P2) holds for $b \in [\underline{i}_t, \overline{a}_{t+1}]$ at $t+1$.*
- *If we begin in E^R and stay in E^R, welfare increases if h is large and $\underline{a}_{t+1} > \underline{a}_t$. That is, it increases if*
    - *$\underline{a}_{t+1} < \underline{a}_t$ and $a^{\dagger}_{t+1}(b) \leq \overline{r}_t$ on $[\underline{a}_{t+1}, \underline{a}_t]$ or*
    - *$\underline{a}_{t+1} > \underline{a}_t$ and a version of (S) holds on $[\underline{a}_t, \underline{a}_{t+1}]$;*
    - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$,*
    - *and (P2) does not hold for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*
- *If we begin in E^R and move to E^{NR}, welfare increases if h is sufficiently large. That is, it increases if*
    - *$\underline{i}_{t+1} < \underline{a}_t$ and (P1) does not hold or*
    - *$\underline{a}_t < \underline{i}_{t+1}$ and a version of (S) holds for $b \in [\underline{a}_t, \underline{i}_{t+1}]$;*
    - *and (P2) does not hold for $b \in [\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$.*

Otherwise, welfare changes are ambiguous.

Proof. A decrease in A_t causes a downward shift in all the major thresholds of the model with the exception of $\underline{a}_t$, whose shift is ambiguous.

Since we are analyzing a marginal decrease, we have $\overline{r}_{t+1} < \overline{r}_t < \min\{\underline{a}_{t+1}, \underline{i}_{t+1}\}$. Utility conditional on interacting with $b \leq \overline{r}_{t+1}$ is identical. Utility on $[\overline{r}_{t+1}, \overline{r}_t]$ increases, since, on this range, $-\overline{r}_{t+1} > -\overline{r}_t$.

$E^{NR} \to E^{NR}$ Next, suppose we begin in E^{NR} and stay in E^{NR}. Welfare conditional on $b \in [\overline{r}_t, \underline{i}_{t+1}]$ increases; it moves from $V - h\overline{r}_t$ to $V - h\overline{r}_{t+1}$. Welfare on $[\underline{i}_{t+1}, \underline{i}_t]$ increases if and only if $V - h\overline{r}_t \leq 0$; employees go from being interacted with and receiving $V - h\overline{r}_t$ to no interaction.

$E^{NR} \to E^R$ Suppose we begin in E^{NR} and move to E^R. By a similar argument as above, welfare conditional on $b \in [\overline{r}_t, \underline{a}_{t+1}]$ increases. Managers on $[\underline{a}_{t+1}, \underline{i}_t]$ previously kept employees indifferent but now interact above the reporting threshold; welfare for employeees increases here if and only if

$$\begin{aligned} V - ha^{\dagger}_{t+1}(b) + p(a^{\dagger}_{t+1}(b), A_{t+1}) - c &\geq V - h\overline{r}_t \\ p(a^{\dagger}_{t+1}(b), A_{t+1}) - c &\geq \big(a^{\dagger}_{t+1}(b) - \overline{r}_t\big)h. \end{aligned}$$

This is the opposite of the condition provided in (S). Note that that $a^{\dagger}_{t+1}(b)$ may be greater than or less than $\overline{r}_t$. If it is greater, employee welfare increases if h is small and decreases if h is large. If $a^{\dagger}_{t+1}(b) < \overline{r}_t$, this always holds.

Finally, employees who match with a manager with $b \in [\underline{i}_t, \overline{a}_{t+1}]$ go from no interaction to interaction at $a^{\dagger}_t$. Their welfare increases if and only if $V - ha^{\dagger}_{t+1}(b) + p(a^{\dagger}_{t+1}(b), A_{t+1}) - c \geq 0$ on this range, i.e. if h is small.

$E^R \to E^R$ Now, suppose we start in E^R and stay in E^R. We either have $\overline{r}_t < \underline{a}_{t+1} < \underline{a}_t < \overline{a}_{t+1} < \overline{a}_t$, or $\overline{r}_t < \underline{a}_t < \underline{a}_{t+1} < \overline{a}_{t+1} < \overline{a}_t$. Utility on $[\overline{r}_t, \min\{\underline{a}_t, \underline{a}_{t+1}\}]$ increases, following the argument above.

If $\underline{a}_{t+1} < \underline{a}_t$, on $[\underline{a}_{t+1}, \underline{a}_t]$, employees go from experiencing the reporting threshold to the interior action $a^{\dagger}_{t+1}$. Welfare increases conditional on matching with $b \in [\underline{a}_{t+1}, \underline{a}_t]$ if and only if

$$\begin{aligned} V - ha^{\dagger}_{t+1}(b) + p(a^{\dagger}_{t+1}(b), A_{t+1}) - c &\geq V - h\overline{r}_t \\ p(a^{\dagger}_{t+1}(b), A_{t+1}) - c &\geq \big(a^{\dagger}_{t+1}(b) - \overline{r}_t\big)h \end{aligned}$$

As before, if $a^{\dagger}_{t+1}(b) \geq \overline{r}_t$, this holds if h is small and does not if h is large. If $a^{\dagger}_{t+1}(b) < \overline{r}_t$, this always holds.

If $\underline{a}_t < \underline{a}_{t+1}$, the exact opposite is true. Managers switch from the old interior action to the new interior action. Welfare improves if and only if

$$\begin{aligned} V - h\overline{r}_{t+1} &\geq V - ha^{\dagger}_t(b) + p(a^{\dagger}_t(b), A_t) - c \\ h &\geq \frac{p(a^{\dagger}_t(b), A_t) - c}{a^{\dagger}_t(b) - \overline{r}_{t+1}} \end{aligned}$$

Notice here that because $a_t^\dagger(b) > \overline{r}_t$, $a_t^\dagger(b) > \overline{r}_{t+1}$ as well. Thus, if $\underline{a}_t < \underline{a}_{t+1}$, welfare increasds if h is large and otherwise decreases.

Next, welfare on $[\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$ increases if and only if

$$\begin{aligned} -ha_{t+1}^\dagger(b) + p(a_{t+1}^\dagger(b), A_{t+1}) \geq -ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) \\ h \geq \frac{p(a_t^\dagger(b), A_t) - p(a_{t+1}^\dagger(b), A_{t+1})}{a_t^\dagger(b) - a_{t+1}^\dagger(b)}. \end{aligned}$$

In both cases, employees experience interaction above the reporting threshold, but with the drop to A_{t+1}, they are less intense. This holds if h is sufficiently large, i.e. if (S) holds.

On $[\overline{a}_{t+1}, \overline{a}_t]$, welfare increases if and only if only if

$$V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) \leq 0$$

Managers here used to play $a_t^\dagger$ but now no longer interact. This holds if h is large, i.e if (P2) does not hold.

$E^R \to E^{NR}$ Finally, suppose we start in E^R and move to E^{NR}. This means $\overline{r}_{t+1} < \underline{i}_{t+1} < \underline{a}_t < \overline{a}_t$ or $\overline{r}_{t+1} < \underline{a}_t < \underline{i}_{t+1} < \underline{a}_{t+1} < \overline{a}_t$.

If $\underline{i}_{t+1} < \underline{a}_t$, welfare increases on $[\underline{i}_{t+1}, \underline{a}_t]$ if and only if no interaction is better than experiencing an action at the old reporting threshold, i.e. if and only if $V - h\overline{r}_t \leq 0$, and $(P1)$ does not hold at time t.

If $\underline{a}_t < \underline{i}_{t+1}$, welfare increases if and only if now experiencing actions at the reporting threshold is better than experiencing an action above the reporting threshold, i.e. if and only if

$$\begin{aligned} -h\overline{r}_{t+1} \geq -ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) \\ h \geq \frac{p(a_t^\dagger(b), A_t)}{a_t^\dagger(b) - \overline{r}_{t+1}} \end{aligned}$$

Because $a_t^\dagger(b) > \overline{r}_{t+1}$, this holds if h is large.

Finally, on $[\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$, welfare increases if and only if no interaction is better than the previous interior actions, i.e. if and only if

$$V - ha_t^\dagger(b) + p(a_t^\dagger(b), A_t) \leq 0$$

his holds if h is large, i.e if (P2) does not hold. □

Corollary 3. *Fixing all other parameters, suppose A_t falls to A_{t+1}, such that A_t corresponds to E^R and A_{t+1} to E^{NR}. If h is sufficiently small and $f(b)$ places sufficiently small mass on $b \in [\overline{r}_{t+1}, \min\{\overline{r}_t, \underline{i}_{t+1}]$, welfare decreases.*

Proof. Note that the previous proposition only considers marginal decreases, i.e. assumes $\underline{i}_{t+1}$ does not fall below $\overline{r}_t$, so we briefly address these cases. If $\overline{r}_{t+1} < \overline{r}_t < \underline{i}_{t+1} < \underline{a}_t$ or $\overline{r}_{t+1} < \overline{r}_t < \underline{a}_t < \underline{i}_{t+1}$, the previous proposition applies. If $\overline{r}_{t+1} < \underline{i}_{t+1} < \overline{r}_t < \underline{a}_t$, welfare on $[\overline{r}_{t+1}, \underline{i}_{t+1}]$ increases. On $[\underline{i}_{t+1}, \overline{r}_t]$, it increases if $V - hb \leq 0$. On $[\overline{r}_t, \underline{a}_t]$, it increases if $V - h\overline{r}_t \leq 0$. Note that if $V - h\overline{r}_t \leq 0$, this is a sufficient condition for $V - hb \leq 0$. Thus, welfare can increase only if h is large and, applying the previous proposition, the welfare effect is ambiguous if h is small.

□