Transgressions in Organizations

Kim Sarnoff and Hassan Sayed [*]

April 14, 2025

Click here for latest version.

Abstract

We study an organizational model where managers commit workplace abuse of unknown harm, or "transgressions." Employees can report transgressive managers for investigation and possible punishment, but managers face uncertainty over what actions employees consider transgressions. When employees' disutility from transgressions is low, we show that policies that disincentivize managers from committing transgressions — increasing the size of manager punishment, the ease of reporting, or the efficacy of investigation technology — may 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 information for the organization and employees, showing that the model converges to a steady state where employees are worse off than initially, harmful actions are never punished, and no organizational learning occurs.

JEL Codes: D23, D73, D83, D91

Keywords: harassment, accountability, organizations, grievances, transgressions, abuse, learning

[*] We thank Leeat Yariv, participants in the Princeton Political Economy Research Seminar, and participants in the Princeton Microeconomic Theory Student Lunch Seminar for helpful comments.

1 Introduction

Organizations often uphold standards that prevent misconduct and abuse of employees at the hands of authority figures like managers. "Transgressions" of workplace norms may include verbal or sexual harassment, forced overtime work, or abuses of power that surpass what employees or the organization find acceptable. Accountability mechanisms for managers who commit transgressions often operate as follows: employees file a report, triggering an investigation into the incident. If a manager is found guilty, employees receive a payout for their grievance, while managers are punished — through suspension, termination of employment, or legal prosecution. However, while many organizations have rules against committing transgressions, managers may not always have a perfect understanding of how egregious their actions must be to constitute a transgression. For example, while firms may have policies banning sexual harassment, managers may not understand exactly what sorts of comments employees consider sexual harassment. Or, while some level of overtime work may be necessary to achieve a firm's objectives, managers may not understand where employees draw the line and consider work excessive.

This paper studies an organizational environment where "managers" have the capacity to commit potentially harmful actions against "employees" but, crucially, are uncertain about which of these actions constitutes a "transgression." In the event that an action crosses a (private) threshold for employees, the action is considered a *transgression* and is thus reportable to and punishable by the organization. We show that policies that discourage managers from committing transgressions may in fact backfire and make employees worse off, depending on the severity of harm that transgressions generate.

We focus on two sets of mechanisms that generate these welfare effects. First, discouraging managers from committing transgressions leads them to not interact with employees in the first place. Interacting with a manager and producing some match value at the risk of experiencing a transgression may be preferred by an employee to no interaction at all *if* the relative harm risked by a transgression is small. Second, managers may moderate the severity of their actions to avoid being punished. On the one hand, these actions are less likely to constitute transgressions, making employees better off. However, in the event they are harmful, it is harder for employees to prove that these actions are transgressions, disincentivizing reporting and making them worse off. When the size of harm is relatively small, the latter channel outweighs the former, and the effect on employee welfare is negative. 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 view the dependence of these welfare effects on the severity of harm as generating insights for different forms of organizational transgressions. “Microaggressions” — subtle but uncomfortable verbal abuse — may be seen as less harmful, and may be particularly privy to the backfiring our model speaks to. More egregious actions that result in physical safety risks for employees, the generation of dangerous precedents for manager behavior, or large and persistent effects on future career prospects may be those that should discourage abuse at all costs.

We finally consider a dynamic extension, where managers and the organization learn about what employees consider transgressions through the verification and adjudication of reports. We show that when only information about verified reports is available publicly, the model converges to a steady state equilibrium where no further organizational learning occurs and transgressions, if they do occur, are never reported or punished. Moreover, we show that employees' expected utility may decrease as learning occurs and we converge towards the steady state.

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. Employees are homogeneous, and are characterized by a common threshold a^*. If a manager's action a_t surpasses this threshold, it is considered a *transgression*, causing employees to incur a disutility h. Managers and the organization do not know the value of a^*, but know that it is below some upper bound A_t.[1] Managers are heterogenous, and are characterized by a bliss point or “type” b, characterizing their affinity for committing transgressions. Managers would like to take actions a_t closer to their bliss point.

Upon experiencing a transgression, employees choose whether to report the action at cost c, triggering an investigation that results in the collection of evidence. Under the interpretation of verifiable evidence, truthful reporting is incentive-compatible for employees. We assume that the probability that the investigation turns up evidence favoring an employee is decreasing in the distance between a_t and A_t, reflecting the idea that A_t reflects an organization's general understanding of what constitutes a transgression, and departures

[1] Formally, we assume that at time $t = 0$, managers' beliefs over $a^* \sim \mathcal{U}[0, A_0]$. If managers observe that some action a' is a transgression, by Bayes' Rule, their posterior over a^* is $\mathcal{U}[0, a']$.

from that understanding may require interpretations or collection of novel evidence. Finally, if an investigation verifies a transgression, managers are punished with a disutility $\gamma \geq V$, while employees receive a payout that compensates for the reporting cost.

We first show that employees report a transgressive action if and only if it exceeds a reporting threshold $\overline{r}_t$, which depends on the organization's understanding of transgressions 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. 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 are reportable as transgressions with positive probability, trading off the value of playing their blisspoints with the risk of potential punishment. We denote these "interior actions" $a_t^\dagger$ Finally, managers in the second equilibrium with very high blisspoints opt out of interaction.

We then study how comparative statics of the model affect employee welfare. We broadly highlight three mechanisms that influence employees' expected utility, depending on the severity of harm h that employees face upon experiencing a transgression. 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 h — 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, $\overline{r}_t$ is ex-ante less likely to constitute a transgression. On the other hand, conditional on being a transgression, employees are strictly worse off; they go from experiencing harm and having a strict incentive to report it (and receive a payout) to being indifferent to reporting and not reporting the transgression. 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 improves expected utility precisely if h is large. This creates a tension between the negative effect of the first mechanism and 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 magnitude 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 transgressive 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 transgressions — affect employee expected utility. We operationalize learning by assuming that the organization keeps a public record of all past *successfully verified* transgressions Thus, when an action a_t is committed and verified as a transgression, the organization and managers know with certainty that the transgression threshold a^* is less than a_t. The upper bound on organizational beliefs about what constitutes a transgression then moves to $A_{t+1} = a_t$. We show that the model always converges to a steady state, in teh sense that the value of A_t remains stationary for the rest of time. This implies that there is no equilibrium learning about the value of a^* upon convergence to the steady state. We then use the tessellation of our three mechanisms to show how learning may negatively affect manager participation and the expected impact of participating managers' actions. We characterize conditions under which the model may move to a steady state where employee welfare is worse off than prior to convergence.

The remainder of this section reviews literature relevant to our setting. We then lay out the basic structure of the model — as well as its applications to other organizational, political, or behavioral settings — and characterize its static equilibrium. We then analyze how changes in model parameters affect employees' expected utility. We finish by looking at the dynamics of learning and reporting on welfare.

Literature Our primary motivation for studying "transgressions" in organizations is driven by a growing interest in workplace harassment and abuse — particularly sexual harassment — in both the mainstream press and scholarly literature. Harassment may include persistent subjugation of employees to unsavory behavior that impacts their emotional wellbeing or productivity, abuse of authority and privilege, or broader violations of workplace *norms* of respect and dignity (Aquino & Thau, 2009). 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, productivity, job satisfaction, and labor turnover (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.

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 managers' *uncertainty* about what actions actually constitute transgressions affect their willingness to commit harm and, subsequently, their effects on victim welfare. 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). In particular, we show that when there is uncertainty over what employees consider harassment, traditional policy levers that discourage managers from committing transgressions can backfire by not only discouraging participation but also encouraging the pursuit of slightly less egregious actions that are harder for employees to report.

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 costs that result from 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 punishment for crimes may be socially optimal due to its, even when law enforcement itself is costless and there is no risk of type-I error.

Finally, our model can also be thought of as a model of political accountability, where an executive who can abuse her authority is uncertain about the degree to which she will be held accountable by the judicial branch. 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 review of evidence, the threat of whistleblowing, and the degree of preference alignment affect the judiciary's efficacy in holding others accountable.

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. This could represent, for example, the value of a project that the manager and employee decide to carry out together. If m decides not to interact, both agents receive 0 and we move to the next period.

Managers are heterogenous; 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. Each employee e shares a common cutoff $a^* \in [0, A_0]$ such that $a > a^*$ imposes a utility cost of $h > 0$ to them, while $a \leq a^*$ induces no cost.[2] a^* represents the threshold for the manager's treatment of the employee to constitute overwork, harassment, or other harmful behavior. We refer to values of a that exceed a^* as *transgressions*. We provide further interpretations of managers, employees, and their actions at the end of the section.

The employee e can choose to report a transgression to the organization, triggering an investigation and possible punishment of the manager. Reporting incurs a cost $c > 0$ to e and either results in the report being *verified* or *unverified*. If the report is *verified*, m is found guilty of a transgression and is punished with a loss γ, while the employee receives a payout normalized to 1. We assume $c < 1$ so that the payout from a successful report is worth the cost. We also assume $V < \gamma$ so that no interaction is always better for a manager than an interaction that is punished with certainty.

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

$$\begin{aligned} &\text{Manager } m: && V - (a-b)^2 - \gamma \mathbf{1}[m \text{ punished}] \\ &\text{Employee } e: && V - h\mathbf{1}[a \geq a^*] - c\mathbf{1}[m \text{ reported}] + \mathbf{1}[m \text{ punished}]. \end{aligned}$$

Updating and Reporting While neither managers nor the organization know a^*, they may learn about its value through reports and their adjudication. We assume that the organization keeps a record of all past, verified reports, which are public knowledge to both employees and managers. We assume that at $t = 1$, the prior over a^* is uniform on $[0, A_0]$.

At time t, let $R(t) \subseteq \{1, 2, \ldots, t\}$ denote the set of time periods t where a transgression was reported and verified. Let $A_t = \min\{a_t : t \in R(t)\}$ be the minimum of all past verified reports. This means, by Bayes' rule, that at time t, the posterior belief over the value of a^* is given by $\mathcal{U}[0, A_t]$.

Given a reported action a_t and A_t, we denote the probability with which a transgression is verified as $p(a_t, A_t)$. We assume the functional form:

$$p(a_t, A_t) = \begin{cases} 1 & a_t > A_t \\ \frac{a_t}{A_t} & a_t \leq A_t \end{cases}$$

This assumption means that an employee's incentive to report depends on the current stan-

[2] The insights of the model are identical if employees are heterogeneous in their thresholds.

dard for what actions constitute a transgression. Specifically, for $a_t < A_t$, there is only a probability of verification, which decreases linearly in the action. We view this as according with the intuition that it is harder to find or interpret evidence when the reported action is further from what is already established as a transgression. Under this interpretation of verification, truth-telling on the part of employees is incentive-compatible. If one reports an action which is not a transgression, the probability of a report being verified is 0, since no evidence would turn up in support of an action that constitutes a transgression, meaning reporting simply entails a cost.

Comments and Interpretations First, note that while managers are heterogeneous in their affinity for committing transgressions, employees are homogeneous and share a common threshold a^*. Since managers are uncertain about the value of a^*, the key welfare insights of the model were employees to have different thresholds would be identical. However, in such a model, there would be no capacity for reports to generate information about employees' thresholds.

Second, we assume the linear functional form of the verification probability $p(\cdot, \cdot)$ for tractability purposes. What matters for the analysis is that the probability of verification decreases as managers commit actions farther from A_t. To this end, we may interpret A_t as reflecting an organization's relative understanding of what constitutes a transgression; actions farther from that standard lack precedent, and may require learning how to interpret evidence in new ways. This implicitly means that the probability of verification depends on a^* insofar as A_t reflects the organization's understanding of the value of a^*. Similarly, the uniform prior over employees' threshold a^* simplifies calculations of managers' expected utility that aid in our later welfare analysis.

Third, note that neither the cost of reporting nor the manager punishment/employee reward for a verified transgression depend on the magnitude of the manager's action a_t or the transgression threshold a^*. However, what matters for the analysis is, that conditional on reporting, a manager's (employees') expected payout is decreasing (increasing) in a_t and increasing (decreasing) in A_t , which is reflected in $p(\cdot, \cdot)$. Hence, allowing for these effects would deliver similar insights as the existing model. Similarly, note that neither the value of a match V nor the disutility of transgressions h not depend on a_t. If the value of a match were to depend on a_t and, in particular, were concave, the model's predictions would be similar, since managers would trade off the benefits of matching and interacting with the risks of punishment and value of non-interaction. Allowing the disutility of transgressions to vary with h would introduce a negative distortionary effect for employees' utility but would leave our basic equilibria unchanged, as well as our basic welfare predictions.

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 a transgression 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. Our motivating interpretation is that transgressions can be seen as "harassment" in an organizational setting. Firms may have policies that protect employees from sexual or racial harassment at the hands of supervisors; however, supervisors may not have a good understanding of what constitutes "harassment" in the first place. Costs of reporting may represent the psychic burden of coming forward with an accusation. The verification process involves the interviewing of witnesses and other parties privy to an incident, and punishment for transgressive managers may range from suspension to termination of employment. Relatedly, the model can be used to analyze overtime work and burnout in firms. Managers can force employees to work overtime on projects, but excessive overtime work may cause employee burnout, which constitutes workplace abuse. Managers are uncertain about employees' tolerance for overtime.

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. The executive faces uncertainty about the extent to which she will be held accountable for potential abuses of authority. 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, acknowledging that their ability to hold the executive accountable is a function of the egregiousness of the executive's actions relative to legal precedent, as well as their ideological tolerance for an executive's actions or an understanding that executive authority may be beneficial in times of crisis. While executives themselves understand precedents for behavior, they may face uncertainty over the degree to which departures from those precedents may be enforced.

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. The firm faces uncertainty over the degree to which they can hike prices, while consumers may face uncertainty about the value of competitors' products if they boycott the original firm.

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:

$$U^E(a_t) = \max \begin{cases} V - c + \mathbf{1}[a_t \geq a^*](-h + \min\{\frac{a_t}{A_t}, 1\}) & R \\ V + \mathbf{1}[a_t \geq a^*](-h) & NR \end{cases}$$

The first line represents the employee's utility if he chooses to report at cost c. If the manager's action is a transgression, he incurs disutility h but a potential reward of 1 with probability $\frac{a_t}{A_t}$. Not reporting simply involves the match utility with a loss of h if $a_t \geq a^*$.

Lemma 1 shows that reporting follows a threshold rule: the employee has a strict incentive to report a transgression if and only if $a_t > \overline{r}_t \equiv cA_t$ and $a_t \geq a^*$.

Lemma 1. *Reporting follows a threshold rule: the employee has a strict incentive to report a transgression if and only if $a_t > \overline{r}_t \equiv cA_t$ and $a_t \geq a^*$.*

We refer to $\overline{r}_t$ as the *reporting threshold* at time t. Note that $\overline{r}_t$ is *not* a function of a^*, meaning it 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:

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

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 A_t additionally entails the cost of sure punishment. For intermediate actions $a_t \in (\overline{r}_t, A_t]$, γ is weighted by the joint probability that the action is a transgression and is reported.

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 $a_t > \overline{r}_t = cA_t$. 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 1**: if $\underline{i}_t < b \leq \min\{\underline{a}_t, A_0\}$, the manager does not interact for all $b \geq \bar{i}_t$. We will refer to this as E^{NL}.*

- ***Equililbrium 2**: if $\underline{a}_t < \min\{\underline{i}_t, A_0\}$, the manager interacts and plays $\alpha_t(b) = a_t^{\dagger}(b) = \frac{bA_t^2}{A_t^2+\gamma} > \overline{r}_t$ 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^L.*

Here, $\underline{a}_t$ is given by $cA_t + c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t}$, $\underline{i}_t$ is given by $cA_t + \sqrt{V}$, and $\overline{a}_t$ is given by $\sqrt{(\frac{A_t^2}{\gamma}+1)V}$.

On the end of the employee, an action a_t is only worth reporting if it is worse than a^* and the probability of verification is sufficiently high. 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 a risk of punishment. We refer to the optimal action that takes into account the risk of punsihment as an "interior action", and we denote by $a_t^{\dagger}(b) = \frac{bA_t^2}{A_t^2+\gamma}$.

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. For those 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 learning can occur in equilibrium.

In particular, two kinds of static equilibria may realize in the model. We will refer to these as a non-learning equilibrium (E^{NL}) and a learning equilibrium (E^L). The non-learning equilibrium E^{NL} 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. Because manager behavior eliminates reporting, no learning can occur.

The learning equilibrium E^L is illustrated in Figure 1(b). In contrast to E^{NL}, 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 (if transgressions) are played with positive probability. This, in turn, makes learning possible.

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

Corollary 1. *The static equilibrium is E^L if and only if*

$$c\frac{\gamma + \sqrt{\gamma(A_t^2 + \gamma)}}{A_t} < \sqrt{V}$$

In particular, there exists $\overline{\gamma}$ such that if $\gamma \geq \overline{\gamma}$, the equilibrium is E^{NL}.

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.

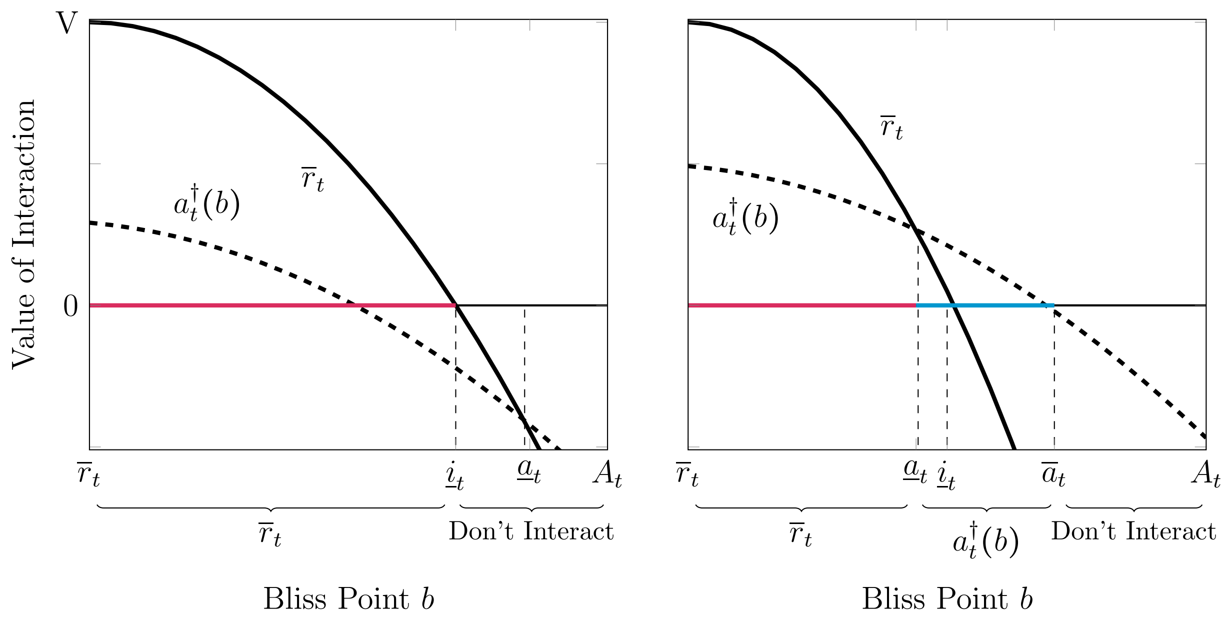
Comments Note that the participation constraint binds for some $b \in (\overline{r}_t, A_0]$, or binds for no one.[3] If it binds for no one, either equilibrium is possible, depending on the values of V

[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

and γ. To highlight the main forces of the model, we will focus on the case where it binds for some $b \in (\overline{r}_t + \epsilon, A_0]$.

Additionally, note that equilibrium learning may *still* halt in E^L if $a^* > a_t^\dagger(\overline{a}_t)$. In this case, managers of type $b \in [\underline{a}_t, \overline{a}_t]$ commit actions above the reporting threshold, but the maximal action played is below the true cutoff for a transgression. So, none of these actions would ever be reported. Conditional on being in E^L, this case occurs with probability $1 - \frac{\overline{a}_t}{A_t}$.

Figure 1: Value of Manager Interactions: Equilibrium 1 (E^{NL}) vs. Equilbrium 2 (E^L)



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

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

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, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases.*

playing the reporting threshold. This ensures that the participation constraint can start to bind only for some $b > \overline{r}_t$.

- A decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases. There exists, $\tilde{A} > 0$ such that the sign of $\underline{a}_t$ is negative for $A_t \le \tilde{A}$ and is positive above.*

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^{NL} to E^L or E^L to E^{NL}.

4 Welfare

With comparative statics of the major thresholds characterizing behavior in hand, we now move to the welfare effects of policy changes. Our analysis will consider welfare as calculated from the perspective of the organization. The following corollary follows from Theorem 1, and characterizes the effects of parameter changes on managers' welfare.

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

Whilethese parameter changes have straightforward effects on manager welfare, the effects on employees' welfare are more nuanced. Increasing 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.

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]

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

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

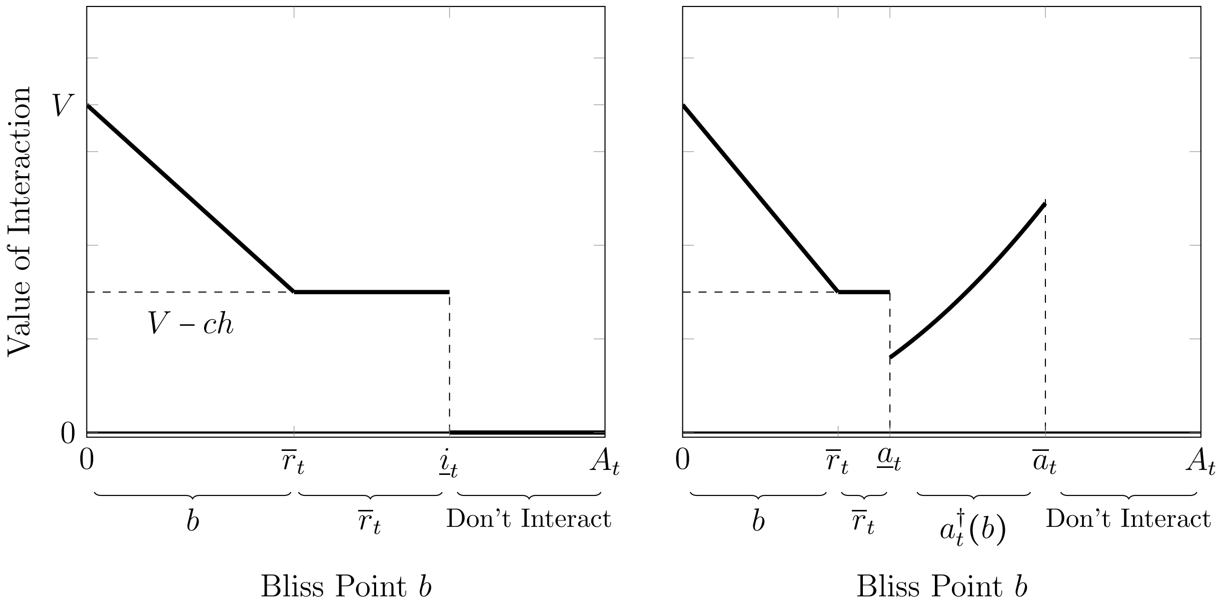
ditional on experiencing a transgression.

$$U_{NL}^{E}(b) = \begin{cases} V - h\frac{b}{A_t} & b \leq \overline{r}_t \\ V - ch & b \in (\overline{r}_t, \underline{i}_t] \\ 0 & b > \underline{i}_t \end{cases} \tag{1}$$

$$U_{L}^{E}(b) = \begin{cases} V - h\frac{b}{A_t} & b \leq \overline{r}_t \\ V - ch & b \in (\overline{r}_t, \underline{a}_t) \\ V + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) & b \in [\underline{a}_t, \overline{a}_t] \\ 0 & b > \overline{a}_t \end{cases} \tag{2}$$

Note that from the perspective of the organization, employees' expected utility from experiencing $\overline{r}_t$ is $V - h\frac{\overline{r}_t}{A_t} = V - h\frac{cA_t}{A_t} = V - ch$, whose value does not depend on b or A_t directly. The indirect utilities for the two equilibria are graphed in Figure 2. Both panels are calibrated with $V - ch > 0$, meaning that matching with a manager and experiencing a transgression is better for employees in expectation than no interaction at all. Calibrating with $V - ch < 0$ does not change the equilibrium thresholds in the figure, since the value of h does not affect managers' optimal strategy.

Figure 2: Value of Employee Interactions: Non-Learning Equilbrium (E^{NL}) vs. Learning Equilibrium (E^{L})



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

(b) E^{L}: $\underline{a}_t < \underline{i}_t < \overline{a}_t$ 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^{NL} or E^L) after the parameter change. Formally, from (1) and (2), welfare increases if an employee's utility from interacting with type b satisfies:

$$V - ch \geq 0 \qquad E^{NL} \tag{P1}$$

$$V + (\frac{bA_t}{A_t^2 + \gamma} - c - h)(\frac{bA_t}{A_t^2 + \gamma}) \geq 0 \qquad E^L \tag{P2}$$

   Decreases in $\overline{a}_t$ or $\underline{i}_t$ result in the opposite effects.

2. **Mechanism 2: switching from interior actions to the reporting threshold**. If $\underline{a}_t$ decreases, or if the initial equilibrium is E^L, 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

$$\begin{aligned} V - ch &\geq V + (\frac{bA_t}{A_t^2 + \gamma} - c - h)(\frac{bA_t}{A_t^2 + \gamma}) \\ \iff h &\geq \frac{bA_t}{A_t^2 + \gamma} \end{aligned} \tag{S}$$

   These expressions follow from (2).

   The reporting threshold is less likely to be a transgression in the first place. But in the event that the original action and the reporting threshold are both transgressions, employees are worse off. The two actions incur the same harm, but 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 of actual transgressions.

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 $a_t^\dagger(b)'$. Since $a_t^\dagger(b) = \frac{bA_t^2}{A_t^2+\gamma} > \frac{bA_t'^2}{A_t'^2+\gamma'} = a_t^\dagger(b)'$, the intensity of the manager's action decreases.

The tradeoff in terms of welfare is analogous to Mechanism 2. This lower action is less likely to be a transgression in the first place. But in the event that the original action and new action are both transgressions, employees are worse off: both actions induce a report, but at the new action, the probability of verification 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

$$\begin{aligned}(\frac{a_t^\dagger(b)}{A_t} - c - h)\frac{a_t^\dagger(b)}{A_t} &\geq (\frac{a_t^\dagger(b)'}{A_t} - c - h)\frac{a_t^\dagger(b)'}{A_t} \\ c + h &\geq a_t^\dagger(b) + a_t^\dagger(b)' \qquad \text{(I)}\end{aligned}$$

These expressions also follow from (2).

These three mechanisms generate subtleties in how equilibrium welfare responds to changes in 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^{NL} to E^{NL} and (P1) holds at V', welfare increases.*
- *If we move from E^{NL} or E^L, 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^L to E^L, 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.

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

Changes in Reporting Costs

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

- *If we move from E^{NL} to E^{NL}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^L to E^L, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma}-c}{\frac{bA_t}{A_t^2+\gamma}-c'}\frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^L to E^{NL}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma}-c}{\frac{bA_t}{A_t^2+\gamma}-c'}\frac{bA_t}{A_t^2+\gamma}$ 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.

The first, present in all cases in Proposition 7, 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^L.

The third effect emerges from Mechanism 2. If we start in E^L, 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, 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^L to E^{NL}. 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^{NL}. If we start in E^{L}, 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.

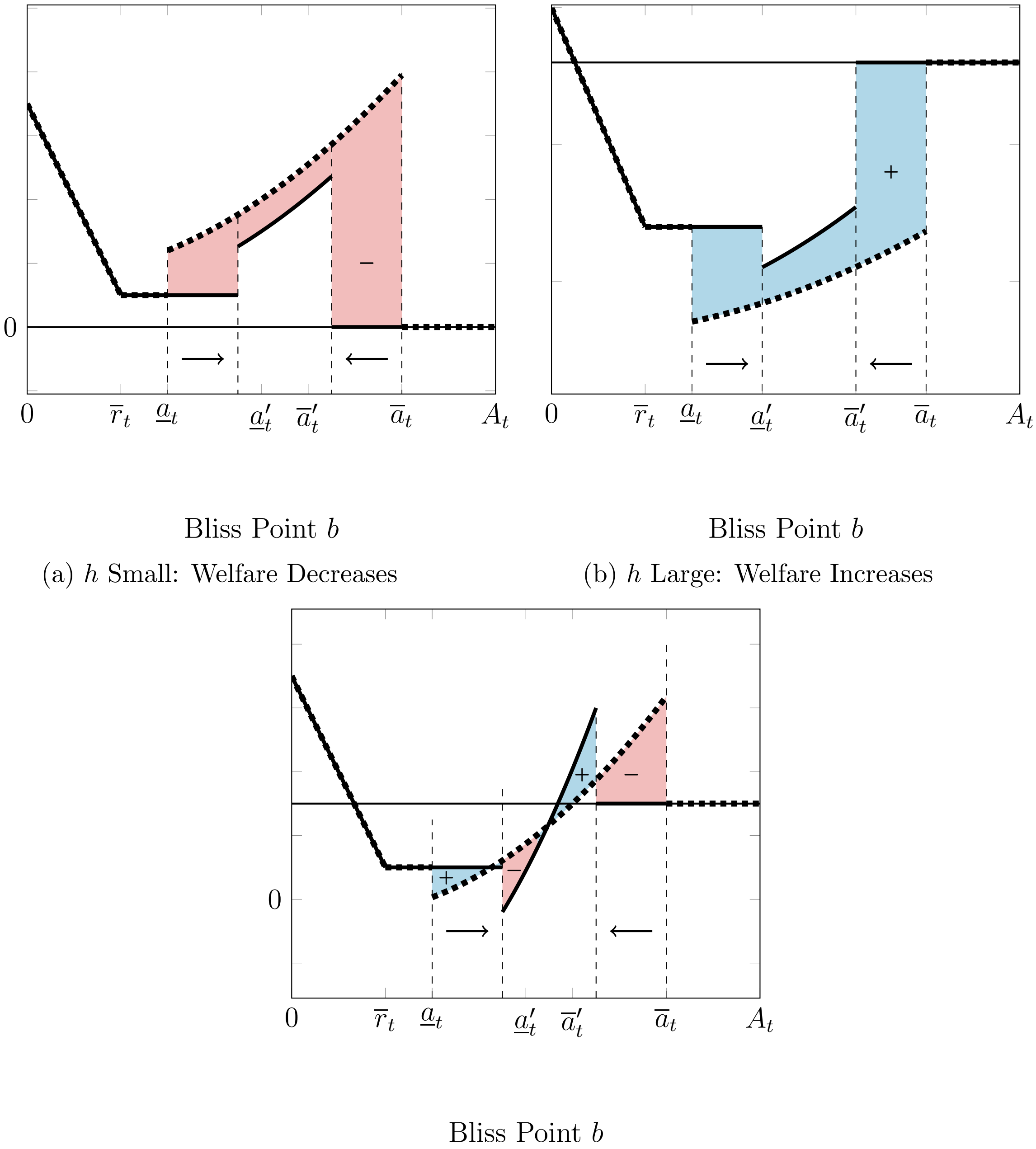
If we begin in E^{NL}, punishment is already large enough that no manager wants to play above the reporting threshold. A fortiori, this will be the case at higher γ as well. So, it is only possible for increasing γ to have an effect if we begin in E^{L}. When γ increases in E^{L}, all three mechanisms 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 (Mechanism 2) or decreasing the probability of report verification (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 cost of a transgression h, 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^L 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 following Theorem synthesizes the patterns in Figure 3, showing that the size of the optimal punishment is increasing in h. When transgressions 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

Within the organizational model of the paper, verified reports update the organizational standard for what actions constitute transgressions. As this process continues, the model may eventually converge to a steady state equilibrium. 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 learning.

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^{NL} is always a steady state, since no learning can occur. Managers either never interact, or interact and keep employees precisely indifferent to reporting. E^L may or may not be a steady state, depending on whether the interior actions being played in fact constitute a transgression. If for some $b \in [\underline{a}_t, \overline{a}_t]$, $a_t^\dagger(b) > a^*$, a report — and therefore learning — can occur. So, this is not a steady state. If $a_t^\dagger(b = \overline{a}_t) \leq a^*$, i.e. the maximal action played in t with positive probability is below the true cutoff for transgressions, then reporting will stop. This is a steady state according to the definition above. However, employees and the organization differ in their perception of whether a steady state has been reached. In particular, the organization still believes that reports will occur with positive probability, since interior actions are being played.

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

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

Dynamic Welfare We analyze the welfare effects of *decreases* in A_t. Since a decrease in A_t corresponds to learning, this allows us to study how refinement of standards for transgression can generate inefficiences in welfare. The following proposition summarizes comparative statics when $A_{t+1} < A_t$.

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}. Conditional on matching with $b \in [0, \overline{r}_{t+1}]$, employee welfare decreases. Moreover,*

- *If we begin and stay in E^{NL}, welfare decreases if h is small, that is, if (P1) holds.*
- *If we begin in E^{NL} and move to E^L, welfare decreases if h is sufficiently large. That is,*
    - *welfare decreases if (S) holds for $b \in [\overline{r}_t, \underline{a}_{t+1}]$*
    - *and (P2) for $b \in [\underline{i}_t, \overline{a}_{t+1}]$.*
- *If we begin in E^L and stay in E^L, welfare decreases if h is sufficiently small and $\underline{a}_t < \underline{a}_{t+1}$. In particular, it decreases if*
    - *$\underline{a}_{t+1} < \underline{a}_t$ and (S) holds;*
    - *$\underline{a}_t < \underline{a}_{t+1}$ and (S) does not hold,*
    - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$, and (P2) holds for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*
- *If we begin in E^L and move to E^{NL}, welfare decreases if h is sufficiently small. That is, welfare decreases if*
    - *$\underline{i}_{t+1} < \underline{a}_t$ and (P1) holds,*
    - *$\underline{a}_t < \underline{i}_{t+1}$ and (S) does not hold for $b \in [\underline{a}_t, \underline{i}_{t+1}]$, and*
    - *and (P2) holds 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^L.[5]

We start by considering managers with $b < r_{t+1}$, who play their bliss point regardless. While their action does not change, the organization's belief that these actions are transgressions increases from $\frac{b}{A_t}$ to $\frac{b}{A_{t+1}}$. This makes employees who match with a manager with low b worse off *in expectation.*

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

Those $b \in [\overline{r}_{t+1}, \overline{r}_t]$ go from playing their bliss point to the new reporting threshold. This, too, results in a strict decrease in welfare, despite being a lower action. Recall that an employee's expected utility from experiencing the reporting threshold is independent of A_t. Since, for b in this range, $\overline{r}_t$ is worse than if they played their bliss point, switching to $\overline{r}_{t+1}$ must necessarily be worse as well. Those who play the reporting threshold in either case, as discussed, incur no change in expected utility. In the figure, this is illustrated for $b \in [\overline{r}_t, \underline{a}_{t+1}]$[6]

The remaining interval of b's are impacted by the mechanisms discussed earlier. Mechanism 2 operates for $b \in [\underline{a}_{t+1}, \underline{a}_t]$: some managers will shift from an action $a_t^\dagger$ above the old reporting threshold to the (new) reporting threshold $\overline{r}_{t+1}$. As in the static case, this decreases welfare if h is small (panel a), but can increase welfare if h is large (panel b).[7]

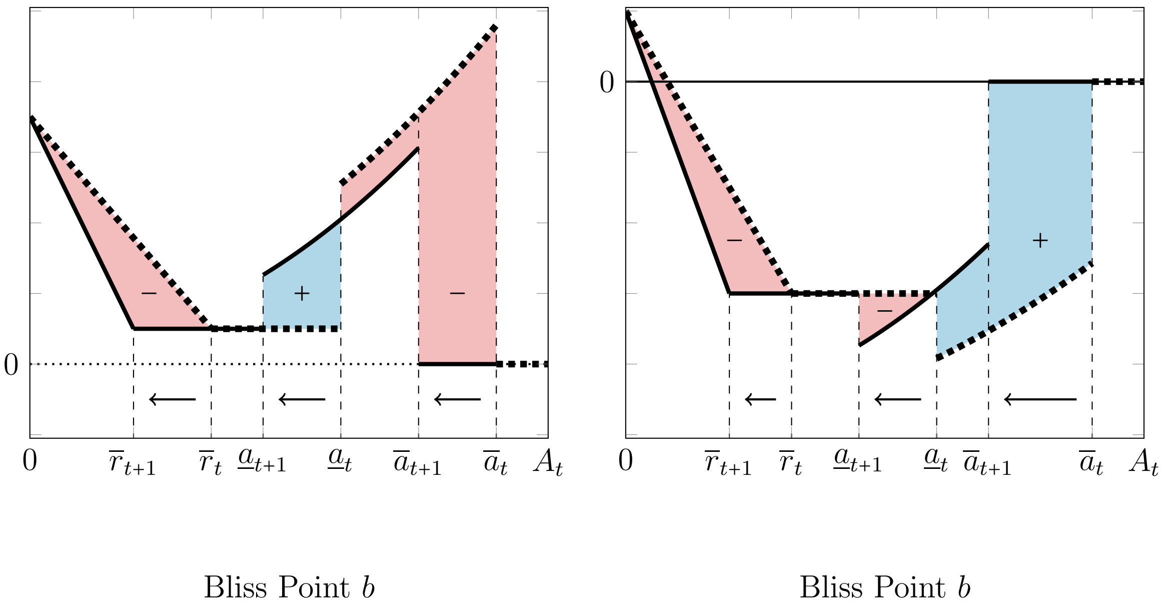
Mechanism 3 operates for $b \in [\underline{a}_t, \overline{a}_{t+1}]$. Suppose a type b manager takes action $a_t^\dagger(b)$ at A_t and $a_{t+1}^\dagger(b)$ at A_{t+1}. Then, $a_t^\dagger(b) > a_{t+1}^\dagger(b)$. On the one hand, these actions are less intense, and hence less likely to constitute a transgression. On the other hand, conditional on a transgression occurring, it is harder to verify. 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).

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

[7] 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; welfare decreases for h large and increases for h small.

Figure 4: Effects of Decrease in A_t to A_{t+1} on Equilibrium Utility Beginning and Ending in E^L, $\underline{a}_{t+1} < \underline{a}_t$, 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^L to E^{NL} results in 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^L and A_{t+1} to E^{NL}. If h is sufficiently small, welfare decreases.*

Intuitively, the welfare effects of a switch from E^L to E^{NL} 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.

The consequence of this result is that, if E^L is *not* a steady state, learning may result in convergence to a steady state E^{NL} where welfare is *strictly worse* than prior to being in a steady state.[8]

[8] Recall that, from the perspective of the organization, E^L is *never* a steady state. If A_t decreases but we remain in E^L 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 where "managers" commit transgressions — 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. While more intense actions are more likely to constitute a transgression, they are uncertain about what employees actually consider transgressions. In the event a manager commits a transgression, employees incur a disutility and face the decision to report this action to the organization at a cost. Verifying a transgression is easier for the organization when it is closer to what has previously been established as a transgression. Verification of a transgression leads to a payout for the victim and a punishment for the manager. Managers are heterogenous and have different propensities for committing more intense actions.

We show that policies that lower managers' expected utility from committing transgressions may have ambiguous or negative effects on employees' expected utility. We highlight three mechanisms that generate these ambiguous effects. First, managers' 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 disutility of experiencing a transgression is large, employees may be better off, no interaction may be preferred to being interacted with and experiencing abuse. However, if the disutility is small, employees' expected utility may decrease.

Second, managers may switch from committing actions that employees have a strict incentive to report to the organization to actions that keep them indifferent. Policy changes that encourage these changes — such as increasing the punishment for committing a transgression — have two effects. On the one hand, because managers commit less intense actions, these are ex-ante less likely to constitute transgressions. On the other hand, in the event an action is actually a transgression, employees now have no incentive to report it and are worse off. If the disutility from a transgression 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 or opt out altogether. These can decrease employees' welfare if the disutility from experiencing a transgression is large.

We use the combination of these three mechanisms to derive the optimal punishment for transgressive managers that maximizes employees' expected utility. We show that optimal punishment depends on the magnitude of harm incurred by a transgression, and is nondecreasing in this magnitude. When the size of harm 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.

Since managers' propensity to commit transgressions is heterogeneous within the organization, the precise welfare effects of these three mechanisms depend on the distribution of managers. However, the general dependence of these different effects allows us to think about policy changes in the setting of different forms of transgressions. One-off microaggressions that make employees feel uncomfortable but may not result in the destruction of employees' work or reputation correspond to transgressions that result in minimal harm. Physical forms of abuse and sabotage that cause long-lasting and persistently negative effects on employees' reputations may correspond to higher levels of disutility.

Finally, we consider a learning extension to the model where reports generate information for the organization and managers. Specifically, we assume that a record of *only* past, successfully-verified reports of transgressions are public information. This means that, as transgressions occur and are reported over time, more precise information about what constitutes a transgression is revealed, which may be thought to improve employees' welfare. We show, however, 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 about transgressions occurs, and if transgressions do occur, they are never punished.

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 harassment and workplace abuse. 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 externalities of harmful actions.

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.

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.

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.

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

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.

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 a transgression if and only if* $a_t > \overline{r}_t \equiv cA_t$ *and* $a_t \geq a^*$*.*

Proof. For $a_t \geq a^*$, the employee has a strict incentive to report if and only if $-c + \frac{a_t}{A_t} > 0$, which occurs if and only if $a_t > cA_t$. □

Lemma 2. *Conditional on interacting, the optimal action* $a^*(t)$ *for a type* b *manager is characterized by the thresholds*

- $\underline{a}_t \equiv cA_t + c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t^2}$*, and*
- $\overline{b}_t \equiv A_t + \frac{\gamma}{A_t}$

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* $\alpha_t(b) = \frac{bA_t^2}{A_t^2+\gamma} \equiv a_t^\dagger(b)$ *(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 by:

$$-2(a_t - b) - 2\frac{a_t}{A_t^2}\gamma = 0 \implies a_t = \frac{bA_t^2}{A_t^2 + \gamma} \equiv a_t^\dagger(b)$$

This interior action is better than $\overline{r}_t$ for $A_t \geq a_t \geq \overline{r}_t$ and $A_t \geq \frac{bA_t^2}{A_t^2+\gamma}$ if and only if

$$\begin{aligned}
V - \left(\frac{bA_t^2}{A_t^2+\gamma} - b\right)^2 - \gamma\left(\frac{bA_t}{A_t^2+\gamma}\right)^2 &\geq V - (cA_t - b)^2 \\
-\left(\frac{b\gamma}{A_t^2+\gamma}\right)^2 - \gamma\left(\frac{bA_t}{A_t^2+\gamma}\right)^2 &\geq -(cA_t - b)^2 \\
-b^2\frac{\gamma^2}{(A_t^2+\gamma)^2} - b^2\frac{\gamma A_t^2}{(A_t^2+\gamma)^2} &\geq -b^2 + 2bcA_t - c^2A_t \\
\left[1 - \frac{\gamma(A_t^2+\gamma)}{(A_t^2+\gamma)^2}\right]b^2 - 2bcA_t + c^2A_t &\geq 0 \\
\left[\frac{A_t^2}{(A_t^2+\gamma)}\right]b^2 - 2bcA_t + c^2A_t &\geq 0 \\
A_t b^2 - 2bc(A_t^2+\gamma) + c^2(A_t^2+\gamma) &\geq 0 \\
b \geq \frac{2c(A_t^2+\gamma) + \sqrt{4c^2(A_t^2+\gamma)^2 - 4A_tc^2k^2(A_t^2+\gamma)}}{2A_t}& \\
b \geq \frac{c(A_t^2+\gamma) + \sqrt{c^2(A_t^2+\gamma)^2 - A_tc^2(A_t^2+\gamma)}}{A_t}& \\
b \geq \frac{c(A_t^2+\gamma) + \sqrt{c^2(A_t^2+\gamma)(A_t+\gamma-A_t)}}{A_t}& \\
b \geq \left(\frac{(A_t^2+\gamma) + \sqrt{\gamma(A_t^2+\gamma)}}{A_t}\right)\left(c\right)& \\
= cA_t + c\frac{\gamma + \sqrt{\gamma(A_t^2+\gamma)}}{A_t} &\equiv \underline{a}_t
\end{aligned}$$

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 $\alpha_t(b) = \frac{bA_t^2}{A_t^2+\gamma}$. Finally, if $\frac{bA_t^2}{A_t^2+\gamma} > A_t$, the optimal action (conditional on interaction) is simply b, since transgressions will always be reported and punished with probability 1, and changing a_t has no marginal impact on punishment probability. This switch is given by the indifference point $\overline{b}_t$ defined by:

$$\frac{\overline{b}_tA_t^2}{A_t^2+\gamma} = A_t \implies \overline{b}_t = A_t + \frac{\gamma}{A_t}$$

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,*

- $\underline{i}_t \equiv cA_t + \sqrt{V}$ *and*
- $\overline{a}_t \equiv \sqrt{\big(\frac{A_t^2}{\gamma} + 1\big)V}$

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 $\frac{bA_t^2}{A_t^2+\gamma}$. Interaction is optimal if and only if

$$V - \Big(\frac{bA_t^2}{A_t^2 + \gamma} - b\Big)^2 - \gamma \frac{b^2 A_t^2}{(A_t^2 + \gamma)^2} \geq 0$$
$$V - \gamma^2 \frac{b^2}{(A_t^2 + \gamma)^2} - \gamma \frac{b^2 A_t^2}{(A_t^2 + \gamma)^2} \geq 0$$
$$V - \frac{\gamma^2 b^2 + \gamma b^2 A_t^2}{(A_t^2 + \gamma)^2} \geq 0$$
$$V - \frac{\gamma}{A_t^2 + \gamma} b^2 \geq 0$$

Note that the expression above is decreasing in b. Moreover, at $b = \overline{b}_t$, the expression is equal to $V - \gamma < 0$. Hence, there exists $\overline{a}_t < \overline{b}_t$ such that there is an incentive to interact if and only if $b \leq \overline{a}_t$. This is i is given by solving the expression above at equality, which yields

$$\overline{a}_t = \sqrt{\Big(\frac{A_t^2}{\gamma} + 1\Big)V}$$

□

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, $a_t^{\dagger}(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases.*
- *A* decrease *in A_t generates a decrease in $\overline{r}_t$, $\underline{i}_t$, and $\overline{a}_t$. For each b, $a_t^{\dagger}(b) = \frac{bA_t^2}{A_t^2+\gamma}$ decreases. There exists, $\tilde{A} > 0$ such that the sign of $\underline{a}_t$ is negative for $A_t \leq \tilde{A}$ and is positive above.*

Proof. All comparative statics are immediate, with the exception of $\underline{a}_t$ with respect to A_t, which is proportional to

$$\frac{(A_t^2+\gamma)+\sqrt{\gamma(A_t^2+\gamma)}}{A_t}$$

The sign of its derivative is positive if and only if:

$$\begin{aligned}
A_t[2A_t + \frac{\gamma A_t}{\sqrt{\gamma(A_t^2+\gamma)}}] - [(A_t^2+\gamma)+\sqrt{\gamma(A_t^2+\gamma)}] &\geq 0 \\
A_t^2 + \frac{\gamma A_t^2}{\sqrt{\gamma(A_t^2+\gamma)}} - \gamma - \sqrt{\gamma(A_t^2+\gamma)}] &\geq 0 \\
(A_t^2-\gamma)\sqrt{\gamma(A_t^2+\gamma)} + \gamma A_t^2 - \gamma(A_t^2+\gamma) &\geq 0 \\
(A_t^2-\gamma)\sqrt{\gamma(A_t^2+\gamma)} - \gamma^2 &\geq 0 \\
(A_t^2-\gamma)^2\gamma(A_t^2+\gamma) &\geq \gamma^4 \\
(A_t^2-\gamma)(A_t^2+\gamma)\gamma(A_t^2-\gamma) &\geq \gamma^4 \\
(A_t^4-\gamma^2)(A_t^2+\gamma) &\geq \gamma^3 \\
A_t^6 + \gamma A_t^4 - \gamma^2 A_t^2 &\geq 0 \\
A_t^2(A_t^2-\gamma) &\geq \gamma^2
\end{aligned}$$

Note that the above is quadratic in A_t^2, is negative below some $\tilde{A}$ which solves the equation above with equality, and is positive above $\tilde{A}$. Moreover, as A_t^2 approaches 0, $\underline{i}_t$ appraoches ∞. □

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

- *if we move from E^{NL} to E^{NL} and (P1) holds at V', welfare increases.*
- *If we move from E^{NL} or E^{L}, 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^{L} to E^{L}, 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^{NL}. If we remain in E^{NL} — 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 a transgression; welfare on $[\underline{i}_t, \underline{i}'_t]$ increases if and only if $V' - ch \geq 0$.

If we switch from E^{NL} to E^{L} — 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 $V - h - c + \frac{bA_t}{A_t^2+\gamma} \geq 0$.

If we begin in E^{L}, note first that we never switch to E^{NL}, 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 $V' + (\frac{bA_t}{A_t^2+\gamma} - c - h)(\frac{bA_t}{A_t^2+\gamma}) \geq 0$. □

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

- *If we move from E^{NL} to E^{NL}, employee welfare decreases if h is large, i.e. if P1 holds at c'.*
- *If we move from E^{L} to E^{L}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma}-c}{\frac{bA_t}{A_t^2+\gamma}-c'}\frac{bA_t}{A_t^2+\gamma}$ for $b \in [\underline{a}'_t, \overline{a}_t]$.*
- *If we move from E^{L} to E^{NL}, welfare decreases if h is small, i.e. if $h \leq \frac{\frac{bA_t}{A_t^2+\gamma}-c}{\frac{bA_t}{A_t^2+\gamma}-c'}\frac{bA_t}{A_t^2+\gamma}$ 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^{L} to E^{NL}, but never from E^{NL} to E^{L}.

Next, for managers with $b \le \overline{r}_t$, 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^{NL} and remain in E^{NL}. 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 $V - c'h \ge 0$.

Next, suppose we are in E^L. 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^L, or $\overline{r}'_t < \underline{a}_t < \underline{i}'_t < \underline{a}'_t$, in which case we move to E^{NL}. 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^L, 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

$$\begin{aligned} -c'h &\ge (-h - c + \frac{bA_t}{A_t^2 + \gamma})\frac{bA_t}{A_t^2 + \gamma} \\ (-c' + \frac{bA_t}{A_t^2 + \gamma})h &\ge (-c + \frac{bA_t}{A_t^2 + \gamma})\frac{bA_t}{A_t^2 + \gamma} \\ h &\ge \frac{\frac{bA_t}{A_t^2+\gamma} - c}{\frac{bA_t}{A_t^2+\gamma} - c'}\frac{bA_t}{A_t^2 + \gamma} \end{aligned}$$

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^{NL}, 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 $V + (-h - c - + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \le 0$.

□

Proposition 4. *An increase in γ does not change employee welfare if we begin in E^{NL}. If we start in E^L, 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 then E^L, 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^L, 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

$$\begin{aligned}
V+(\frac{bA_t}{A_t^2+\gamma}-c-h)(\frac{bA_t}{A_t^2+\gamma}) &\leq V-h\frac{\overline{r}_t}{A_t}\\
(\frac{bA_t}{A_t^2+\gamma}-c-h)(\frac{bA_t}{A_t^2+\gamma}) &\leq -h\cdot c\\
\big(-c+\frac{bA_t}{A_t^2+\gamma}\big)(\frac{bA_t}{A_t^2+\gamma}) &\leq h\big(\frac{bA_t}{A_t^2+\gamma}-c\big)\\
h &\geq \frac{bA_t}{A_t^2+\gamma}
\end{aligned}$$

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 a transgression conditional on it occurring) but also a potential increase (an action for a given type b manager in this range is less likely to be a transgression). An increase is better if and only fi

$$\begin{aligned}
(\frac{bA_t}{A_t^2+\gamma}-c-h)(\frac{bA_t}{A_t^2+\gamma}) &\leq (-h-c+\frac{bA_t}{A_t^2+\gamma'})(\frac{bA_t}{A_t^2+\gamma'})\\
\frac{-h-c}{A_t^2+\gamma}+\frac{bA_t}{(A_t^2+\gamma)^2} &\leq \frac{-h-c}{A_t^2+\gamma'}+\frac{bA_t}{(A_t^2+\gamma')^2}\\
bA_t\left(\frac{1}{(A_t^2+\gamma)^2}-\frac{1}{(A_t^2+\gamma')^2}\right) &\leq (h+c)\big(\frac{1}{A_t^2+\gamma}-\frac{1}{A_t^2+\gamma'}\big)\\
bA_t\left(\frac{1}{A_t^2+\gamma}+\frac{1}{A_t^2+\gamma'}\right) &\leq (h+c)\\
h-\frac{bA_t}{A_t^2+\gamma'} &\geq \frac{bA_t}{A_t^2+\gamma}-c\\
h &\geq \frac{bA_t}{A_t^2+\gamma'}+\frac{bA_t}{A_t^2+\gamma}-c
\end{aligned}$$

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 $0 \geq V+(\frac{bA_t}{A_t^2+\gamma}-c-h)(\frac{bA_t}{A_t^2+\gamma})$.

Finally, if we move from E^L to E^{NL}, 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^{NL}, welfare does not change with γ, so suppose we are in E^L. An employee's expected utility conditional on γ is given by:

$$F(\overline{a}_t)V + \int_0^{\overline{r}_t} -h\frac{b}{A_t}f(b)db + \int_{\overline{r}_t}^{\underline{a}_t} -h\cdot cf(b)db + \int_{\underline{a}_t}^{\overline{a}_t}(-h-c+\frac{bA_t}{A_t^2+\gamma})\frac{bA_t}{A_t^2+\gamma}f(b)db$$

Noting that $\overline{a}_t$ and $\underline{a}_t$ are both functions of γ, the derivative of this expression with respect to γ is

$$f(\overline{a}_t)\overline{a}_t'V - hcf(\underline{a}_t)\underline{a}_t' + (-h-c+\frac{\overline{a}_tA_t}{A_t^2+\gamma})\frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - (-h-c+\frac{\underline{a}_tA_t}{A_t^2+\gamma})\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t'$$
$$+\int_{\underline{a}_t}^{\overline{a}_t} -2\frac{(bA_t)^2}{(A_t^2+\gamma)^3} + (h+c)\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db.$$

Note that $\overline{a}_t'(\gamma) < 0$ and $\underline{a}_t'(\gamma) > 0$. Hence, after reorganizing terms, the derivate's sign can be characterized via the following pieces:

$$\underbrace{f(\overline{a}_t)\overline{a}_t'V + (-c+\frac{\overline{a}_tA_t}{A_t^2+\gamma})\frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - (-c+\frac{\underline{a}_tA_t}{A_t^2+\gamma})\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t'}_{\text{Negative}}$$
$$\underbrace{+h\left[\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' - \frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - cf(\underline{a}_t)\underline{a}_t' + \int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db\right]}_{\text{Positive}}$$
$$\underbrace{-\int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}\left[\frac{2bA_t}{A_t^2+\gamma}-c\right]f(b)db}_{\text{Positive}} \geq 0$$

To see that the middle term is positive, note that

$$\overline{r}_t = cA_t < \frac{\underline{a}_tA_t^2}{A_t^2+\gamma} \implies c < \frac{\underline{a}_tA_t}{A_t^2+\gamma} \implies cf(\underline{a}_t)\underline{a}_t' < \frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' < \frac{\underline{a}_tA_t}{A_t^2+\gamma}[f(\underline{a}_t)\underline{a}_t' - f(\overline{a}_t)\overline{a}_t']$$

Since $\overline{a}_t' < 0 < \underline{a}_t'$, this gives the result. Moreover, differentiating the expression above with respect to h yields

$$\left[\frac{\underline{a}_tA_t}{A_t^2+\gamma}f(\underline{a}_t)\underline{a}_t' - \frac{\overline{a}_tA_t}{A_t^2+\gamma}f(\overline{a}_t)\overline{a}_t' - cf(\underline{a}_t)\underline{a}_t' + \int_{\underline{a}_t}^{\overline{a}_t}\frac{bA_t}{(A_t^2+\gamma)^2}f(b)db\right] > 0$$

suggesting, by Topkis' Monotonicity Theorem, that $\gamma^*(h)$ is nondecreasing in h. Finally, note that at $h = 0$, the sign of the derivative is negative everywhere, suggesting that 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, suggesting 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^{NL}, 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^{NL}, we are already in a steady state. Hence, suppose we begin in E^L.

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 $\frac{bA_t^2}{A_t^2+\gamma}$ being played; in this case, $A_{t+1} = \frac{bA_t^2}{A_t^2+\gamma} < A_t$. 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 A^* corresponds to E^L, $A^* > a^*$, and $a^* \leq a_*^\dagger(\overline{a}_*)$, where $\overline{a}_*$ is the limit of $\overline{a}_t$ values as $t \to \infty$ and $a_*^\dagger$ is defined likewise for $a_t^\dagger$. But note that since $F(b)$ places positive mass on $[\underline{a}_*, \overline{a}_*]$, A^* decreases with positive probability when actions in this range are played, contradicting that A^* is a steady state. □

Proposition 6. *Suppose A_t decreases marginally to A_{t+1}. Conditional on matching with $b \in [0, \overline{r}_{t+1}]$, employee welfare decreases. Moreover,*

- *If we begin and stay in E^{NL}, welfare decreases if h is small, that is, if (P1) holds.*
- *If we begin in E^{NL} and move to E^L, welfare decreases if h is sufficiently large. That is,*
    - *welfare decreases if (S) holds for $b \in [\overline{r}_t, \underline{a}_{t+1}]$*
    - *and (P2) for $b \in [\underline{i}_t, \overline{a}_{t+1}]$.*
- *If we begin in E^L and stay in E^L, welfare decreases if h is sufficiently small and $\underline{a}_t < \underline{a}_{t+1}$. In particular, it decreases if*
    - *$\underline{a}_{t+1} < \underline{a}_t$ and (S) holds;*
    - *$\underline{a}_t < \underline{a}_{t+1}$ and (S) does not hold,*
    - *(I) holds for $b \in [\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$, and (P2) holds for $b \in [\overline{a}_{t+1}, \overline{a}_t]$.*

- *If we begin in E^L and move to E^{NL}, welfare decreases if h is sufficiently small. That is, welfare decreases if*

  - $\underline{i}_{t+1} < \underline{a}_t$ *and (P1) holds,*
  - $\underline{a}_t < \underline{i}_{t+1}$ *and (S) does not hold for* $b \in [\underline{a}_t, \underline{i}_{t+1}]$*, and*
  - *and (P2) holds 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 $\overline{r}_{t+1} < \overline{r}_t < \min\{\underline{a}_{t+1}, \underline{i}_{t+1}\}$. Welfare conditional on interacting with $b \leq \overline{r}_{t+1}$ decreases, since $-\frac{b}{A_t} > -\frac{b}{A_{t+1}}$. Additionally, welfare on $[\overline{r}_{t+1}, \overline{r}_t]$ also decreases, since, on this range, $-\frac{b}{A_t} > -c$.

$E^{NL} \to E^{NL}$ Next, suppose we begin in E^{NL} and stay in E^{NL}. Welfare conditional on $b \in [\overline{r}_t, \underline{i}_{t+1}]$ remains the same; it is $V - ch$ in both cases. Welfare on $[\underline{i}_{t+1}, \underline{i}_t]$ decreases if and only if $V - ch \geq 0$; employees go from being interacted with and receiving $V - ch$ to no interaction.

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

$$\begin{aligned} V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) &\leq V - ch \\ (-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) &\leq (-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})h \\ h &\geq (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \end{aligned}$$

Finally, employees who match with a manager with $b \in [\underline{i}_t, \overline{a}_{t+1}]$ go from no interaction to interaction at $a_t^\dagger$. Their welfare decreases if and only if $V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \leq 0$ on this range.

$E^L \to E^L$ Now, suppose we start in E^L and stay in E^L. 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 $[\overline{r}_t, \min\{\underline{a}_t, \underline{a}_{t+1}\}]$ does not change, 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 $a_t^\dagger$. Welfare decreases conditional on matching with $b \in [\underline{a}_{t+1}, \underline{a}_t]$ if and only if

$$\begin{aligned} V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \leq V - ch \\ h \geq (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \end{aligned}$$

If $\underline{a}_t < \underline{a}_{t+1}$, the opposite holds; welfare decreases conditional on matching with a manager of type $b \in [\underline{a}_t, \underline{a}_{t+1}]$ if and only if $h \leq (\frac{bA_{t+1}}{A_{t+1}^2+\gamma})$.

Next, welfare on $[\max\{\underline{a}_t, \underline{a}_{t+1}\}, \overline{a}_{t+1}]$ decreases if and only if $(-h - c + \frac{bA_{t+1}}{A_{t+1}^2+\gamma})(\frac{bA_{t+1}}{A_{t+1}^2+\gamma}) \leq (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma})$; in both cases, employees experience interaction above the reporting threshold, but with the drop to A_{t+1}, they are less intense. This condition can be simplified as follows:

$$\begin{aligned} (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) &\leq (\frac{bA_t}{A_t^2 + \gamma} - c - h)(\frac{bA_t}{A_t^2 + \gamma}) \\ h(\frac{bA_t}{A_t^2 + \gamma} - \frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) &\leq (\frac{bA_t}{A_t^2 + \gamma} - c)(\frac{bA_t}{A_t^2 + \gamma}) - (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma} - c)(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \\ h &\leq \frac{(\frac{bA_t}{A_t^2+\gamma} - c)\frac{bA_t}{A_t^2+\gamma} - (\frac{bA_{t+1}}{A_{t+1}^2+\gamma} - c)\frac{bA_{t+1}}{A_{t+1}^2+\gamma}}{\frac{bA_t}{A_t^2+\gamma} - \frac{bA_{t+1}}{A_{t+1}^2+\gamma}} \\ h &\leq \frac{(\frac{bA_t}{A_t^2+\gamma})^2 - (\frac{bA_{t+1}}{A_{t+1}^2+\gamma})^2 + c(\frac{bA_t}{A_t^2+\gamma} - \frac{bA_{t+1}}{A_{t+1}^2+\gamma})}{\frac{bA_t}{A_t^2+\gamma} - \frac{bA_{t+1}}{A_{t+1}^2+\gamma}} \\ h &\leq \frac{bA_t}{A_t^2 + \gamma} + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma} - c \end{aligned}$$

On $[\overline{a}_{t+1}, \overline{a}_t]$, welfare decreases if and only if only if $0 \leq V + (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma})$; managers here used to play $a_t^\dagger$ but now no longer interact.

$E^L \rightarrow E^{NL}$ Finally, suppose we start in E^L and move to E^{NL}. 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 decreases on $[\underline{i}_{t+1}, \underline{a}_t]$ if and only if no interaction is worse than experiencing an action at the reporting threshold, i.e. if and only if $V - ch \geq 0$. If $\underline{a}_t < \underline{i}_{t+1}$, welfare decreases if and only if now experiencing actions at the reporting threshold is worse than experiencing an action above the reporting threshold, i.e. if and only if

$$\begin{aligned} V - ch &\le V + (-h - c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \\ (-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})h &\le (-c + \frac{bA_{t+1}}{A_{t+1}^2 + \gamma})(\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \\ h &\le (\frac{bA_{t+1}}{A_{t+1}^2 + \gamma}) \end{aligned}$$

Finally, on $[\max\{\underline{i}_{t+1}, \underline{a}_t\}, \overline{a}_t]$, welfare decreases if and only if no interaction is worse than the previous interior actions, i.e. if and only if $V + (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \ge 0$, as above.

□

Corollary 3. *Fixing all other parameters, suppose A_t falls to A_{t+1}, such that A_t corresponds to E^L and A_{t+1} to E^{NL}. If h is sufficiently small, 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 $[\overline{r}_{t+1}, \underline{i}_{t+1}]$ decreases. On $[\underline{i}_{t+1}, \overline{r}_t]$, it decreases if $V - \frac{b}{A_t}h \ge 0$. On $[\overline{r}_t, \underline{a}_t]$, it decreases if $V - ch \ge 0$. Note that if $V - ch \ge 0$, this is a sufficient condition for $V - \frac{b}{A_t}h \ge 0$

Next, since we are switching from E^L to E^{NL}, we necessarily have $\underline{a}_t < \underline{i}_t$ but $\underline{i}_{t+1} < \underline{a}_{t+1}$, which means that

$$c\frac{\gamma + \sqrt{\gamma(A_t^2 + \gamma)}}{A_t} < \sqrt{V} < c\frac{\gamma + \sqrt{\gamma(A_{t+1}^2 + \gamma)}}{A_{t+1}}$$

Note that since, as $A_{t+1} \to 0$, $\frac{\gamma + \sqrt{\gamma(A_{t+1}^2 + \gamma)}}{A_{t+1}} \to \infty$, that there always exists A_{t+1} small such that this expression holds.

Next, following from the previous proposition, we have that welfare decreases in the following situations:

- $\underline{a}_t < b < \underline{i}_{t+1}$ and $h \le \frac{bA_t}{A_t^2+\gamma}$ OR $\underline{i}_{t+1} < b < \underline{a}_t$ and $V - ch \ge 0$
- $\max\{\underline{i}_{t+1}, \underline{a}_t\} < b < \overline{a}_t$ and $V + (-h - c + \frac{bA_t}{A_t^2+\gamma})(\frac{bA_t}{A_t^2+\gamma}) \ge 0$.

These conditions hold by choosing h sufficiently small without disturbing the initial condition that $c\frac{\gamma+\sqrt{\gamma(A_t^2+\gamma)}}{A_t} < \sqrt{V} < c\frac{\gamma+\sqrt{\gamma(A_{t+1}^2+\gamma)}}{A_{t+1}}$. □