Whistleblower Protection: Theory and Experimental Evidence*

Lydia Mechtenberg†  Gerd Muehlheusser‡  Andreas Roider§

January 28, 2018

Abstract
Whistleblowing by employees plays a major role in uncovering corporate fraud. Recent laws and global policy recommendations aim at facilitating whistleblower protection to enhance the willingness to report and to increase deterrence. We study these issues in a theory-guided lab experiment. Whistleblower protection indeed leads to more reporting of misbehavior. However, our experimental findings suggest that non-meritorious claims are an issue, as they reduce prosecutors’ incentives to investigate, which hampers the intended improvement of deterrence.

JEL-Code: C91, D83, D73, K42, M59.

Keywords: Corporate Fraud, Corruption, Whistleblowing, Business Ethics, Cheap-Talk Games, Lab Experiment

*We gratefully acknowledge financial support by the Fritz Thyssen Foundation (Grant 10.13.2.097).
†Universität Hamburg, Department of Economics, lydia.mechtenberg@uni-hamburg.de
‡Corresponding author: Universität Hamburg, Department of Economics, IZA, and CESifo, gerd.muehlheusser@uni-hamburg.de
§University of Regensburg, Department of Economics, CEPR, IZA and CESifo, andreas.roider@ur.de
1 Introduction

1.1 Motivation

Corporate fraud is a major challenge in both developing and advanced economies, and employee whistleblowers play an important role in uncovering it. Indeed, the issue of protecting employee whistleblowers looms high on the international anti-corruption agenda of the G20 group, the Council of Europe, and the OECD, and some form of whistleblower protection legislation is already in existence in countries such as the U.S. or the UK. To the best of our knowledge, this is the first paper to study whistleblower protection regimes in a theory-guided lab experiment, where the decision to commit fraud, whistleblowing, investigation decisions, and potential retaliation against whistleblowers are endogenous. Our results suggest that the effect of whistleblower protection on deterrence deserves particular attention.

The topicality of corporate fraud is exemplified by high-profile scandals at Volkswagen, Enron, or Worldcom. More systematic evidence is, for example, presented by Dyck, Morse, and Zingales (2014). Using a natural experiment, they estimate the average cost of both detected and undetected fraud in large U.S. corporations in the period 1996-2004 to be $380 billion per year.\(^1\) This evidence suggests that uncovering existing fraud and deterring potential fraud should indeed be a high priority for legislators and policy makers.

In fact, in recent years, the importance of employee whistleblowers (who are not participating in the misbehavior) for uncovering fraud (in particular, fraud involving company insiders) has become evident, primarily because of their access to crucial information.\(^2\) For example, Dyck, Morse, and Zingales (2010) consider all reported cases of fraud in large U.S. corporations between 1996 and 2004. They find that in 17\% of the 216 cases they study, the fraud was uncovered by employee whistleblowers, thereby outnumbering other players such as the SEC, auditors, non-financial market regulators, or the media.\(^3\) The importance of employee whistleblowers has lead to a broad consensus among scholars and practitioners alike that, in

---

\(^1\) According to the Association of Certified Fraud Examiners (2014), the average loss of organizations due to fraud (which includes financial statement fraud, asset misappropriation, and corruption) is estimated to be 5\% of annual revenues. Taken at face value, this number would extrapolate into a worldwide loss from fraud of up to $3.7 trillion. Furthermore, in the latest “Global Fraud Report” (Kroll, 2016), 75\% of surveyed senior executives stated that their company had become a fraud victim in the previous year.

\(^2\) Kroll (2016) finds that in 81\% of all fraud cases where perpetrators were known at least one company insider was involved, and a substantial share of 36\% of these perpetrators came from senior or middle management.

\(^3\) In the notorious corporate fraud scandals of Enron or Worldcom, the misbehavior was uncovered by employee whistleblowers (see, e.g., Healy and Palepu, 2003). Miceli, Near, and Dworkin (2009) survey fraud cases unveiled by whistleblowers in more than 20 countries.
order to uncover and deter corporate fraud, fostering employee whistleblowing would be very desirable. While employees who do report fraud often feel a moral obligation to do so (see, e.g., Jos, Tompkins, and Hays, 1989; Miceli and Near, 1992; Alford, 2001), the fear of retaliation from co-workers or management is often a strong countervailing factor. As a result, the overall willingness of employees to report misbehavior is often perceived as low. For example, Dyck, Morse, and Zingales (2010, p.2245) argue that “the surprising part is not that most employees do not talk, but that some talk at all.”

As a consequence, the best-practice recommendations of international bodies, such as the G20 group, the Council of Europe, and the OECD (Council of Europe, 2014; OECD 2011, 2016), urge for comprehensive legal protection from retaliation for whistleblowers. Such legislation is already in place in the U.S., the UK, and a number of other countries, where examples include the Sarbanes-Oxley Act (SOX), the Dodd-Frank Act, and the Public Interest Disclosure Act (see, e.g., Thüsing and Forst, 2016). For instance, U.S. law aims to shield whistleblowers from tangible employment actions (such as dismissal or demotion) and other forms of retaliation.

There is no doubt that employee whistleblowers who uncover corporate fraud deserve strong protection. A number of scholars, practitioners, and enforcement agencies have at the same time voiced the concern that current whistleblower protection policies might also lead to dysfunctional responses in the form of non-meritorious claims. In particular, it has been argued that low-performing employees might have an incentive to lodge claims to seek shelter from unfavorable actions such as dismissal. For example, in their study of the New York City administration, Anechiarico and Jacobs (1996, p.69) state that “some disgruntled, incompetent, or otherwise poorly performing employees will file whistleblower claims in order to keep their jobs as long as possible or simply to harass their supervisors.” If such claims were indeed successful, the employee would either be retained, or would settle to the effect of receiving a severance payment in exchange for termination.

4See e.g., Near and Miceli (1986) and Alford (2001). Gobert and Punch (2000, pp.33ff) provide a number of striking examples of whistleblowers suffering from various forms retaliation after coming forward.

5See als Gobert and Punch (2000, pp.32ff), Schmidt (2005, p.158), Bowen, Call, and Rajgopal (2010, p.1240), or Blount and Markel (2012, p.1042), who reiterate the arguments and cases of Anechiarico and Jacobs (1996, pp.67ff). Furthermore, a USA Today (2004) article on the practical consequences of SOX quotes practitioner statements such as “some of the more difficult problems I’ve had is whistleblowers who will raise issues in which we find some merit, but where they will raise them to gain personal protection for marginal performance”.

6In a similar vein, it has been argued that monetary rewards for whistleblowers, as for example provided for in the Dodd-Frank Act, might also generate incentives to file non-meritorious claims, see e.g., Hartmann (2011, p.1303), Ebersole (2011, p.135), Blount and Markel (2012, p.1041), Hansberry (2012, p.196), or Rose (2014, p.1283).
Since such non-meritorious claims would be an unintended by-product in the quest for improving the protection of honest whistleblowers, this raises the question of their empirical relevance. In this respect, there might exist several reasons why such claims may not constitute some quantité négligeable with respect to both whistleblower behavior and fraud deterrence.

First, from the viewpoint of the responsible authorities, if the number of non-meritorious claims turns out to be large, reports may in general be perceived to be less informative about the presence of underlying misbehavior. As it is costly to “separate the wheat from the chaff” (Fleischer and Schmolke, 2012), this might reduce the authorities’ responsiveness to reports (see e.g., Ebersole, 2011, p.135 and Rose, 2014, p.1283). In turn, this could reduce deterrence in the first place (see e.g., Casey and Niblett, 2014, pp.1208ff). If this channel is indeed empirically relevant, the deterrence of corporate fraud as one of the major aims of whistleblower legislation (apart from the primary aim of protecting whistleblowers) would be impaired.

Second, many existing whistleblower protection policies (such as SOX and Dodd-Frank) and the above-mentioned best practice recommendations do not seem to provide strong safeguards against non-meritorious claims. This is due to one important feature of these policies, according to which whistleblowers are deliberately not required to provide conclusive proof of their allegations. Instead, they need to demonstrate a reasonable belief with respect to the presence of fraud (for a discussion, see e.g., Kohn, Kohn and Colapinto, 2004, pp.92ff). The rationale behind the use of this judicial standard is to set the bar for protection not too high, thereby encouraging whistleblowers to come forward. However, since claims often involve complex inside information, a claim’s real merit is often hard and costly to assess for a judge ex ante (see e.g., Ebersole, 2011, p.135), and a non-meritorious allegation might appear reasonable (see e.g., Schmidt, 2005, p.158). Hence, while the reasonable belief standard will certainly prevent obviously unsubstantiated claims, this potentially leaves scope for (ex post) non-meritorious claims to result in protection.

Finally, existing policies do usually not sanction whistleblowers whose claims pass the reasonable belief threshold ex ante, but then turn out to be non-meritorious ex post. For example, in their comparative study of whistleblower legislation in 23 countries, Thüising and Forst (2016)...

---

7For example, out of the 27921 cases determined by the U.S. Department of Labor’s Occupational Safety and Health Administration (OSHA) over the time period 2006-2016, 56% were dismissed (see https://www.whistleblowers.gov/factsheets_page/statistics).

8See e.g., Rose (2014, p.1283), who also argues that unwarranted protection might in addition be aided by the vagueness of many legal provisions (see also Ebersole, 2001, p.135, Hansberry, 2012, pp.211ff, or Blount and Markel, 2012, p.1041). Claims that are easily recognizable as fraudulent could for example be punished with sufficiently high fines, and hence be deterred in the first place (Buccirossi, Immordino, and Spagnolo, 2017).
find that, once obtained, protection often remains intact even if, in the end, it turns out that there was no misbehavior. Furthermore, sanctions for claims that turn out to be unsubstantiated are in general either mild or ruled out altogether, as is for example the case for all claims administered by the U.S. Department of Labor (see e.g., Kohn, Kohn and Colapinto, 2004, p.34).

Given the global scale of corporate fraud and the importance of employee whistleblowers in uncovering it, the details of whistleblower protection policies might matter substantially for economic outcomes. However, we believe that the effects of improving the protection of employee whistleblowers who uncover corporate fraud on non-meritorious claims (together with their potentially adverse effects on deterrence) are not yet sufficiently well understood. This paper is an attempt to provide some preliminary insights on these issues.

1.2 Research Question, Framework, and Results

The main goal of this paper is to study, in a unified framework, the effects of legal whistleblower protection on corporate misbehavior, employee whistleblowing, investigations, and retaliation by employers against whistleblowers. To this end, we conduct a theory-guided experiment where predictions are derived from a cheap-talk model in the spirit of Crawford and Sobel (1982). Our framework considers the interaction between an employer (who may misbehave), an employee (who may blow the whistle), and a prosecutor (who may act upon the employee’s report). Moreover, the employer might retaliate against a non-protected whistleblower in the form of dismissal. In this paper, we focus on whistleblower protection in the form of employment protection (i.e., a protected employee cannot be dismissed), which is consistent with common legal practice as discussed above.

We allow employees to be heterogenous with respect to their productivity: The incentive structure is such that, whenever feasible, the employer prefers to dismiss low-productivity employees, while their high-productivity counterparts might face retaliation only if they blow the whistle. Hence, as discussed above, this might give low-productivity employees an incentive to file non-meritorious claims in order to be shielded from dismissal.

We implement a total of six experimental treatments, four main treatments and two ro-

---

9For example, two such cases are discussed in Anechiarico and Jacobs (1996, pp.67ff).
10Employment protection is the most common remedy in whistleblower cases (see, e.g., Kohn, Kohn and Colapinto, 2004, pp.97ff). Thereby, the stated aim is to make whole the whistleblower, i.e., re-establish the attained employment status before becoming a whistleblower (see, e.g., Kohn, Kohn and Colapinto, 2004, pp.102ff).
bustness checks. In treatment $P-R$, protection is obtained by filing a report. This treatment is meant to capture in a stylized way real-world legal regimes (such as U.S. law, UK law, and the G20 group’s policy recommendation), where protection is granted when the employee can demonstrate a reasonable belief with respect to the presence of fraud. In this and all subsequent treatments we focus on the case where whistleblower claims do satisfy this reasonable-belief criterion, i.e., from the prosecutor’s perspective they are not obviously unsubstantiated ex ante.

Treatment $P-R$ is then compared to two benchmarks, one where whistleblower protection (i.e., employment protection) is not available (treatment $NoP$), and one where protection is only granted if, in addition to a report, the employer’s misbehavior is indeed verified in the course of an investigation (treatment $P-RIM$). This latter treatment serves as a useful benchmark relative to $P-R$, as protection is available in both treatments, but in $P-RIM$ there is no incentive to file non-meritorious claims. Finally, we also consider an intermediate setting where a report only leads to protection when it triggers an investigation (treatment $P-RI$).

Our theoretical predictions (developed in Appendix A) can be summarized as follows: First, in treatment $P-R$, all misbehavior is reported, but low-productivity employees also lodge non-meritorious claims. Second, in the benchmark $NoP$ (where protection is not available), non-meritorious claims by employees do not arise, but not all employer misbehavior is reported. Moreover, in treatment $NoP$ deterrence is weaker than in $P-R$. Third, the predictions for $P-R$ and $P-RI$ coincide. Fourth, in treatment $P-RIM$, all misbehavior is reported (as in $P-R$), but non-meritorious claims do not arise and, as a consequence, deterrence is stronger in this benchmark treatment.

The main experimental findings are as follows: First, most of the theoretical predictions (with respect to dismissal, misbehavior, and the reporting behavior of the different productivity types) are broadly supported by the experimental data. Second, there are also interesting deviations. In treatment $P-R$, we find that non-meritorious claims are indeed an issue, but they are also lodged by high-productivity employees (which is not predicted by the theory). Moreover, non-meritorious claims indeed reduce prosecutors’ responsiveness to reports, as these are now less informative about underlying misbehavior. As a consequence, the predicted reduction of misbehavior in $P-R$ relative to $NoP$ does not materialize. Third, as predicted, the behavior in $P-R$ and $P-RI$ is very similar. Fourth, in $P-RIM$, there are substantially fewer non-meritorious claims.  

---

11This is driven by our focus on informative equilibria in which the prosecutor triggers an investigation if and only if there is a report by the employee.
claims than in $P-R$, and employer misbehavior is lower, too. Fifth, in all three treatments where protection is available, the willingness to report misbehavior is higher compared to the benchmark $NoP$ without protection, while the responsiveness of prosecutors to reports is lower.

From a methodological point of view, our lab experiment complements empirical research with field data on whistleblowing. For example, empirically observing an unaltered number of reports after the introduction of whistleblower protection might have at least two possible explanations. First, the scheme might simply be insufficient to trigger more reports. Alternatively, it might indeed increase the willingness to report as intended, which, in turn, deters misbehavior to such a degree that the number of observed reports remains constant. With field data, it is usually difficult to distinguish between these two explanations. Also, with field data one directly observes only those cases of misbehavior that come to light, but has to resort to estimation to gauge the extent of undetected misbehavior. By contrast, a lab experiment allows to directly observe crucial variables such as the underlying (and potentially undetected) level of misbehavior, the willingness to send both truthful and fraudulent reports, and the prosecutors’ response to them. Moreover, in the lab one can run “policy experiments”; thereby (pre-)testing various features of whistleblower protection programs before they are implemented.

The remainder of the paper is structured as follows: Section 2 discusses the related literature. Section 3 introduces the game played and the design of the experiment, while Section 4 presents the theoretical predictions and the underlying intuition. The experimental results are discussed in Section 5. Section 6 concludes. Appendix A contains the description and analysis of the model from which the theoretical predictions of Section 4 are derived. Appendix B contains translations of the experimental instructions. Appendix C provides an overview over the number of observations across decisions and treatments.

2 Related Literature

Our paper is the first one to systematically evaluate (both experimentally and theoretically) how the requirements for obtaining whistleblower protection affect outcomes. In doing so, we complement three strands of the literature on whistleblowing policies. First, there is a theoretical literature on whistleblowing that analyzes the optimal responsiveness of prosecutors to reports. In particular, in Chassang and Padró i Miquel (2016) whistleblowing is fostered

\footnote{These methodological advantages are similar to those advanced in the related experimental literature on leniency programs in antitrust, which is discussed below.}
through investigation policies that generate “garbled” information. They show that, to shield
a whistleblower from retaliation by his employer, the optimal investigation policy (to which
the investigator can commit ex ante) must not be too responsive to reports. The reason is
that a relatively responsive policy would reveal that whistleblowing has in fact occurred, which
would then trigger retaliation. In turn, this would undermine the incentive to report in the
first place.¹³ Like the present paper, Chassang and Padró i Miquel (2016) analyze a cheap-talk
game in which the decisions to misbehave, to report, and to investigate are endogenous. Hence,
from a theoretical perspective, their setup is the one most closely related to ours, but there
are a number of important differences: We compare the impact of different protection schemes
on equilibrium behavior, and allow for heterogeneity of workers with respect to productivity.
Moreover, we focus on pure-strategy equilibria where the investigator has no commitment power
(and hence decides on whether or not to investigate only after a report has arrived). Finally,
we also empirically test our model predictions in a lab experiment. Using a different modeling
approach, Heyes and Kapur (2009) analyze how the optimal responsiveness of investigations
depends on different behavioral motives for whistleblowing such as conscience cleansing, social
welfare considerations, or disgruntlement. Our model captures the first of these motives by
assuming that potential whistleblowers suffer a disutility from undetected misbehavior.

Second, there is a literature that analyzes the role of monetary rewards in fostering whistle-
blowing, as for example implemented in the False Claims Act and the Dodd-Frank Act. Dyck,
Morse, and Zingales (2010) and Zingales (2004) stress the beneficial role of such rewards in un-
covering fraud, while others discuss potentially adverse effects such as fostering non-meritorious
claims.¹⁴ There are also two recent experimental studies on financial rewards. Schmolke and
Utikal (2016) compare financial rewards for whistleblowers and fines for non-reporting, while
Butler, Serra, and Spagnolo (2017) analyze whether monetary rewards might lead to crowding-
out of intrinsic motivation to report.¹⁵ In our paper, we vary the requirements for obtaining
(employment) protection (which shields the employee from a potential downside of whistleblow-

¹³Benoit and Dubra (2004) and Muehlheusser and Roider (2008) show that, even in the absence of a threat
of direct retaliation, reporting might not occur due to the fear of enforcement errors or future non-cooperation.
¹⁴See e.g., Givati (2016), Howse and Daniels (1995), Callahan and Dworkin (1992) and the references given
in Footnote 6. At the moment, financial rewards are not yet very widespread across jurisdictions. For example,
only 30% of the 27 countries surveyed in OECD (2016) have incentives for whistleblowers (such as financial
rewards, expediency of the process, or follow-up mechanisms) in place. Likewise, in a world-wide survey, the
Association of Certified Fraud Examiners (2014) finds that only 11% of organizations had a reward scheme in
place.
¹⁵See Gneezy, Meier, and Rey-Biel (2011) for a survey of the crowding-out literature, and Benabou and Tirole
ing) in order to study how these requirements affect not only reporting, but also misbehavior, investigations, and retaliation. However, we abstract from financial rewards.

Third, there is a literature on leniency programs in anti-trust, which studies the self-reporting of cartel members (see, e.g., the surveys by Spagnolo, 2008, and Marvão and Spagnolo, 2014, and the recent experimental studies by Apesteguia, Dufwenberg, and Selten, 2007, Hinloopen and Soetevent, 2008, Bigoni, Fridolfsson, Le Coq, and Spagnolo, 2012, 2015, and Fel-tovich and Hamaguchi, 2016). This body of (theoretical, empirical, and experimental) research also analyzes how to foster the reporting of illegal activities. However, it considers settings of oligopolistic competition in which every party is involved in the illegal behavior, which is not the case for the whistleblowers in our setup.\footnote{Cotten and Santore (2016) conduct an experiment to analyze the impact of transparency and amnesty rules in the context of corporate fraud by criminal teams.}

Finally, our paper relates to an empirical literature (in fields such as psychology, sociology, organizational behavior, and business ethics) analyzing the impact of situational and personal factors on the reporting decisions of whistleblowers. For example, such factors are the threat of retaliation, whether or not co-workers were harmed, the type of misbehavior and its severity, whether individuals are rather high-performers or low-performers, and the strength of behavioral motivations such as conscience cleansing, see, e.g., the overviews by Miceli and Near (1992), Miceli, Dworkin, and Near (2008), Mesmer-Magnus and Viswesvaran (2005) and Vadera, Aguilera, and Caza (2009), and the recent incentivized lab experiment by Bartuli, Djawadi, and Fahr (2016). In our paper, we investigate how the reporting decisions of employees are affected by their productivity and the underlying legal regime. In addition, we also elicit a number of personal characteristics and situational factors in the post-experimental questionnaire.

3 Experimental Design

This section explains the setup of the experiment. That is, we provide summary information and describe in detail the game played, the incentive structure, the session design and payments, the various treatments and their framing, as well as the post-experimental procedures.

Summary Information The experiment was conducted in the experimental lab of the University of Hamburg between April 2015 and March 2016 and programmed in z-Tree (Fis...
In total, we ran four main treatments (which differ with respect to the requirements for obtaining employment protection) and two further treatments as robustness checks (see Table 1 below). We employed a between-subjects design, so that each subject participated in one treatment (and hence, one session) only. Sessions lasted for approximately 140 minutes, and participants earned 21 Euro on average (including a show-up fee of 12 Euro). For the recruitment of a total of 600 subjects, we used the software tool hroot (Bock, Baetge, and Nicklisch, 2014). Virtually all subjects were undergraduate or master students at the University of Hamburg from a variety of fields (40% majors or minors in economics, business, or a related field), and 51% were female.

**The Game Played in Each Period** In each of 30 periods per session, subjects were randomly (re-)matched into groups of four (stranger-design). They were assigned a role as either employer, employee, prosecutor, or third party (where the role assignments across periods are explained in more detail below). Employees are heterogenous with respect to their (exogenously given) productivity, which is either high (“H-employee”) or low (“L-employee”), drawn randomly anew (with equal probability) at the beginning of each period. The third party is a purely passive player (without any decisions to make) who suffers a loss from employer misbehavior. The third party is included in the experiment to make it more salient that misbehavior causes harm to others. The remaining three players played the following game (summarized in Figure 1), which is an implementation of the game analyzed in the theory part as laid out in Appendix A:

![Figure 1: The Game Played in Each Period](attachment:image)

At date 1, the employer observes the productivity of his employee. She then chooses whether or not to misbehave. Misbehavior entails a gain (which is independent of her employee’s productivity type) but is costly to others. At date 2, we used the strategy method to elicit the
employee’s binary reporting decision for both cases with and without employer misbehavior. Then, the employee observes the actual misbehavior decision of the employer. At date 3, the prosecutor observes whether or not a report is sent by the employee (but the prosecutor observes neither the underlying employer misbehavior decision nor the employee’s productivity type).

The prosecutor then decides on triggering an investigation (which incurs a private cost for the prosecutor). An investigation perfectly reveals whether or not the employer has misbehaved. Moreover, if misbehavior is uncovered this benefits both the prosecutor and the third party, while the employer must pay an (exogenously given) fine. Finally, at date 4, the employer decides whether or not to dismiss the employee. In case of dismissal, the employee receives a payoff of zero and is replaced by a (computerized) outsider, who is more (less) productive than an L-employee (H-employee). However, dismissal is only feasible as long as the employee is not shielded by whistleblower protection. The observability of the employee’s reporting decision to the employer is discussed below when we introduce the various treatments. At the end of each period, subjects learn their individual payoffs from the current period, and the decisions leading to these payoffs.

**Incentive Structure: Monetary Incentives and Parameter Values** In the experiment, the players’ monetary payoff components (which were common knowledge ex ante) had the following properties: Unless detected, an employer’s monetary payoff is higher upon misbehavior. Moreover, the difference between the productivity and the wage of the L-employee (H-employee) is smaller (larger) compared to employing the replacement outsider. Hence, the employer’s monetary payoff is higher when dismissing (retaining) the L-employee (H-employee). In contrast, the monetary payoff of each employee type is always higher when retained. The monetary payoff of the third party is highest under no misbehavior, followed by detected, and then undetected misbehavior. Finally, despite the investigation costs, when there actually is

---

17 Hence, we consider reports that are “external” in the sense of being directed towards the (outside) prosecutor. Some whistleblower laws stipulate that firms must establish internal reporting systems, and that whistleblowers must use these internal channels first, before resorting to outsiders. Incorporating this issue would require a richer framework, which might be an interesting topic for future research.

18 The assumption that the prosecutor has discretion whether to initiate an investigation is in line with both the related literature (see, e.g., Chassang and Padró i Miquel, 2016; Givati, 2016; Heyes and Kapur, 2009) and legal practice (e.g., under SOX). The case where investigations do not perfectly reveal underlying misbehavior is discussed in Section 5.4 below.

19 The motivation for this payoff ranking is that detecting misbehavior might allow to (at least partly) curb the associated harm.

20 The payoff structure of the underlying model is summarized in Table 5 in Appendix A.
misbehavior, the prosecutor’s monetary payoff is higher when he investigates. In contrast, in the absence of misbehavior, the prosecutor’s monetary payoff is higher when he does not investigate.

We used the following parameter values throughout (where the numbers indicate experimental points): The productivities of H-employees, L-employees, and the outside replacement are given by 80, 30, and 70, respectively. Employees receive a fixed wage of 40. The employer’s (gross) payoff from misbehavior is 50, and, in case of detection, she faces a fine of 60. When there is no misbehavior, the prosecutor’s payoff is −20 (0) if he investigates (does not investigate). When there is misbehavior, his payoff is −10 (−20) if he investigates (does not investigate). Note that the fine does not accrue to the prosecutor. Finally, the third party suffers a loss of 50 (70) from detected (undetected) misbehavior. In order to avoid negative payoffs at the end of the experiment, only prosecutors and third parties (who otherwise would face only negative payoff consequences) received per-period endowments of 60 and 40, respectively.

**Incentive Structure: Potential Behavioral Motivations**

As discussed below, our experimental design is intentionally not fully neutral, and we do provide subjects with some information about the context in which they operate. Consequently, in addition to the monetary payoff components, there might also exist moral and psychological motivations that shape subjects’ behavior (which were not incentivized in the experiment): First, employees do not receive a direct monetary reward when reporting misbehavior, and we rely on their potential moral motivation to report misbehavior instead. As discussed above, the literature has identified conscience cleansing as a main motive of whistleblowers to come forward. Second, it is well documented that whistleblowers are often no longer well-liked at their workplace. That is, employers might feel tempted to retaliate in the form of dismissal, even though this might reduce the employers’ profit due to a loss of productivity (i.e., when matched with an H-employee). Third, employers might have moral reservations such that their “net benefit” from misbehavior differs from their pure monetary gain. In the theoretical analysis in Appendix A, on which the predictions of Section 4 are based, we allow for heterogeneity with respect to the intensity of these moral and psychological motives (i.e., employees’ dislike of undetected misbehavior, employers’ dislike of employing whistleblowers, and employers’ aversion against misbehavior).

---

21 Hence, given that there is misbehavior, an investigation is not only beneficial to the third party, but also to the prosecutor (which, in practice, might for example come in the form of a reputation gain).

22 The existence of such heterogeneity on the employers’ side is consistent with empirical findings on the relevance of retaliation. For example, Near and Miceli (1996, pp.517ff) find retaliation rates ranging from 6%
Apart from that, the theoretical predictions rely on the assumption that subjects have standard preferences.

**Session Design and Payments**  In each session, the design of the experiment was common knowledge, and all subjects received the same instructions. Sessions consisted of 30 periods and usually had 24 participants (see Table 1). In addition to the (random) re-matching of groups in each period, also the role assignments varied across periods as follows: Each subject who was assigned the role of employer in the first period retained this role throughout all 30 periods. All other subjects randomly switched roles across periods, either between employees and third partys or between prosecutors and third partys. This was communicated in the instructions, where we also stated that role assignments were independent of subjects’ behavior. The aim of this re-shuffling was to make the negative consequences of misbehavior more salient; in particular to the employee and the prosecutor, whose decisions might (directly or indirectly) curb the harm inflicted by the employer on the third party.

In addition, in order to ensure that subjects indeed understood the game, after going through the instructions, subjects had to answer a series of control questions, and we discussed any wrong answers with them in private before finally launching the experiment. Finally, to determine each subject’s payment, three out of the 30 periods were randomly selected, and the subject’s total points earned in these three periods were converted at the rate of 1 Euro per 15 points. Together with the show-up fee, this was paid out (in private) in cash at the end of the session.

**Treatments**  We consider four main treatments (see Table 1) which differ with respect to the conditions under which protection is obtained (to the effect that a protected employee cannot be dismissed at date 4).

Treatment NoP corresponds to a benchmark setting in which employment protection is not available at all. In addition, there are three treatments where protection is available. In treatment P-R, protection is obtained by sending a report. This treatment is meant to capture real-world legal regimes (such as U.S. law, UK law, and the G20 group’s policy recommendation), where protection is granted when the employee can demonstrate a reasonable belief with respect to the presence of fraud. In this and all subsequent treatments we focus on the to 38%, suggesting that employers do differ with respect to their attitude towards whistleblowing (see also the National Business Ethics Survey of 2013 available at https://www.ibe.org.uk/userassets/surveys/nbes2013.pdf).
case where whistleblower claims do satisfy this reasonable-belief criterion, i.e. they are not obviously unsubstantiated. Moreover, a whistleblower cannot be legally sanctioned in case the investigation reveals that there was no misbehavior. The assumption that a report suffices for protection is relaxed in a third treatment $P\text{-}RI$, where in addition, the report needs to trigger an investigation by the prosecutor. Finally, in treatment $P\text{-}RIM$ protection is only granted if, in addition to the requirements of $P\text{-}RI$, if the investigation indeed reveals the misbehavior by the employer. As discussed in more detail in Sections 5.3 and 5.4 below, as robustness checks we ran two further treatments $RL$ and $IE$, in which employers face a (reputation) loss whenever an investigation occurs and in which there are investigation errors, respectively.

As can be shown, the theoretical predictions for none of the treatments depend on whether the reporting decision of the employee is observed by the prosecutor only or by both the prosecutor and the employer.\textsuperscript{23} Intuitively, this is driven by the fact that the employer can observe the investigation decision and by our focus on informative equilibria where the prosecutor investigates if and only if a report occurs. In the experiment, the prosecutor’s behavior might deviate from this prediction, so that it might be more difficult for the employer to correctly infer the reporting decision from observing the investigation decision only. Hence, as we wanted to rule out the possibility of erroneous updating by the employer as a potential driver for dismissal decisions (rather than any potential dislike of whistleblowing), in treatments $NoP$ and $P\text{-}R$,

\textsuperscript{24}While many whistleblower protection laws require firms to establish anonymous reporting channels, Chassang and Padró i Miquel (2016) argue that the protection offered by a formal requirement of anonymity might be limited in practice as in many cases the set of people informed about misbehavior will be small to begin with (and hence the identity of the whistleblower can be conjectured).

---

**Table 1: Treatments**

<table>
<thead>
<tr>
<th>Conditions for Protection</th>
<th>Never</th>
<th>Report Only</th>
<th>Report + Investigation</th>
<th>Report + Investigation + Detected Misbehavior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main Treatments</strong></td>
<td>$NoP$ (5;120)</td>
<td>$P\text{-}R$ (5;120)</td>
<td>$P\text{-}RI$ (4;88)</td>
<td>$P\text{-}RIM$ (4;96)</td>
</tr>
<tr>
<td><strong>Robustness Checks</strong></td>
<td>$RL$ (Reputation Loss) (4;88)</td>
<td>$IE$ (Investigation Errors) (4;88)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Below each treatment label, the first and the second number indicate the number of sessions played and the total number of subjects participating in the treatment, respectively. In three out of the 26 sessions, the number of participants was 16 (instead of 24) because of no-shows.
both the prosecutor and the employer learn the reporting decision. In treatments \( P-RI \) and \( P-RIM \) (where a report alone is not sufficient for obtaining protection), the employer learns the reporting decision if there is an investigation.\(^{24}\)

**Framing**  In the experiment, we gave subjects some information about the context in which their behavior was placed (e.g., we framed the game as an employer-employee relationship, where the employee could file a report to a prosecutor).\(^{25}\) However, in the experimental instructions (see Appendix B), we avoided the use of strongly judgemental terms such as “misbehavior”, “illegal” or “whistleblowing”. For example, in the experiment, we referred to an employer’s misbehavior decision as a choice between two alternatives CIRCLE (i.e., no misbehavior) and TRIANGLE (i.e., misbehavior). However, all subjects were informed that “a (fictitious) law for the protection of the third party says that TRIANGLE should not be chosen as it harms the third party” (see Appendix B). Moreover, the employee’s reporting decision was not referred to as “whistleblowing”, but as “asking the prosecutor to trigger an investigation”.

**Post-Experimental Procedures**  At the end of the respective session, subjects completed a (non-incentivized) questionnaire in which we elicited socio-demographic information (e.g., age, gender, and field of study), risk preferences (via the “100.000 Euro question” of Dohmen, Falk, Huffman, Sunde, Schupp, and Wagner, 2011), and cognitive abilities (via the “Cognitive Reflection Test” of Frederick, 2005), and their attitudes towards revealing misbehavior (measured on a five-level Likert scale). In addition, we elicited subjects’ “Dutifulness” (i.e., their sense of duty and obligation) as a sub-factor of the Big Five personality trait “Conscientiousness” (where the respective questions were taken from the “NEO Personality Inventory”, see Costa and McCrae, 1992; Berth and Goldschmidt, 2006). As to make these issues not too salient, the above questions were interspersed with some unrelated questions. We also elicited information about subjects’ social preferences by letting them play an incentivized standard one-shot dic-

\(^{24}\)Hence, the comparisons of treatments \( \text{NoP} \) and \( P-R \), and \( P-RI \) and \( P-RIM \), respectively, follow a one-change-at-a-time principle. While this is not the case for the comparison between treatments \( P-R \) and \( P-RI \) (where both the requirements for obtaining protection and the observability of reports change), the theoretical predictions for these two treatments are identical (see Prediction \( P-RI \) below), and this is also borne out in the experiment (see Section 5.3).

\(^{25}\)In experimental economics, there is a discussion about the conditions under which a neutral or a loaded framing is more appropriate, see e.g., Eckel and Grossman (1996) and Alekseev, Charness, and Gneezy (2017). In experimental studies on whistleblowing in organizations, a loaded framing is used in Bartuli, Djawadi, and Fahr (2016) and Cotten and Santore (2016), while Schmolke and Utikal (2016) choose a neutral design. Framing is also discussed in other contexts involving misbehavior, e.g., in experiments on corruption (Abbink and Hennig-Schmidt, 2006; Barr and Serra, 2009) and tort litigation (Loewenstein, Issacharoff, Camerer, and Babcock, 1993; Babcock, Loewenstein, Issacharoff, and Camerer, 1995).
ator game in which they had to decide on how to split 100 points between themselves and a “receiver”. We used the strategy method so that subjects had to make their choice before they knew whether they were actually (randomly) assigned the role of dictator or receiver. We then converted their resulting points at the rate of 1 Euro per 20 points and added this to the monetary payoff they received at the end of the experiment. The post-experimental questionnaire is available upon request.

4 Theoretical Predictions

The theoretical predictions for our main treatments NoP, P-R, P-RI, and P-RIM are derived from the pure-strategy Perfect Bayesian Equilibria of the game described in Section 3, which is formally spelled out and analyzed in Appendix A (see Propositions 1-4). We focus on informative equilibria in the sense that the prosecutor triggers an investigation if and only if the employee sends a report. This directly leads to

Prediction I (Investigation): In all treatments, prosecutors trigger (do not trigger) an investigation upon receiving (not receiving) a report by the employee.

Also, the prediction for the employer’s dismissal decision is straightforward. Intuitively, the employer prefers to dismiss an L-employee whenever this is feasible because the (expected) productivity of the outside replacement is higher. In contrast, an H-employee will only be dismissed upon reporting, and only if the employer’s dislike of employing a whistleblower exceeds the H-employee’s productivity advantage. This leads to

Prediction D (Dismissal): In all treatments: (i) unless protected, L-employees are dismissed. (ii) H-employees are retained when sending no report, while they are dismissed with positive probability when sending a report and not being protected.

The predictions for the reporting and misbehavior decisions are treatment-specific: We start with the comparison of treatments NoP and P-R, and then discuss treatments P-RI and P-RIM.

Prediction R (Reporting): The reporting behavior in treatments NoP and P-R is summarized in Table 2. In particular: (i) In both treatments, misbehavior leads to a (weakly) higher willingness to report for either productivity type. (ii) For either misbehavior decision,
L-employees exhibit a (weakly) higher willingness to report than H-employees. (iii) For either misbehavior decision, both productivity types exhibit a (weakly) higher willingness to report in treatment P-R. (iv) Fraudulent claims are sent by L-employees only, and they occur in treatment P-R only.

Table 2: Theoretical Prediction: Fraction of Employees Sending a Report

<table>
<thead>
<tr>
<th>Treatment</th>
<th>NoP</th>
<th>P-R</th>
<th>P-RIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employee Type</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Misbehavior</td>
<td>1 ∈ [0,1]</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No Misbehavior</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: The prediction for treatment P-RIM is the same as for treatment P-R.

Hence, in treatment NoP, but not in P-R, misbehavior by the employer is a necessary, but not sufficient, condition for reporting to occur (see Table 2). Intuitively, recall that in our model employees are assumed to suffer a disutility from undetected misbehavior, so that either productivity type tends to be more willing to report when misbehavior actually occurs. However, in anticipation of the subsequent investigation and dismissal decisions, the reporting behavior differs across types as L-employees expect to be dismissed whenever feasible, while H-employees are less vulnerable due to their higher productivity. This gives the former a higher incentive to send both truthful and fraudulent reports: When misbehavior actually occurs, H-employees are facing a trade-off between any disutility from undetected misbehavior under no reporting and the higher risk of dismissal when doing so. Moreover, in treatment P-R, L-employees have an incentive to report even when there is no misbehavior, as this protects them from dismissal. Finally, we have:

Prediction M (Misbehavior): Misbehavior in treatments NoP and P-R is summarized in Table 3. In particular: (i) When the employer is matched with an L-employee, the frequency of misbehavior is the same in NoP and P-R. (ii) When the employer is matched with an H-employee, the frequency of misbehavior is strictly lower in P-R than in NoP.

Intuitively, misbehavior of employers with L-employees does not vary across the two treatments as the decision whether or not to misbehave has no effect on the dismissal of L-employees:
<table>
<thead>
<tr>
<th>Treatment</th>
<th>NoP</th>
<th>P-R</th>
<th>P-RIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-employee</td>
<td>$m_{L}^{\text{no}}$ = $m_{L}^{1}$ &gt; $m_{L}^{3}$</td>
<td>$\lor$</td>
<td>$\lor$</td>
</tr>
<tr>
<td>H-employee</td>
<td>$m_{H}^{\text{no}}$ &gt; $m_{H}^{1}$ = $m_{H}^{3}$</td>
<td>$\lor$</td>
<td>$\lor$</td>
</tr>
</tbody>
</table>

Notes: The prediction for treatment P-RI is the same as for treatment P-R. $m_{j}^{\theta}$ denotes the frequency of misbehavior by an employer matched with an employee of productivity $\theta = L, H$ in treatment $j = \{\text{no}, 1, 3\}$.

This productivity type is always dismissed in treatment NoP (irrespective of any earlier decisions), and he is always shielded from dismissal in treatment P-R (as L-employees always report, again irrespective of their employer’s earlier misbehavior decision). Because H-employees report any misbehavior in treatment P-R, the incentive to misbehave is smaller in this treatment.

With respect to treatment P-RI, our focus on informative equilibria implies that the theoretical predictions coincide with those of treatment P-R. The reason is that the only case in which the two treatments would have different implications does not occur on the equilibrium path (i.e., the case that the employee sends a report, but the prosecutor does not investigate, which would lead to protection in P-R, but not in P-RI). This implies

**Prediction P-RI:** The predictions for treatment P-RI coincide with those for P-R.

Predictions change, however, in treatment P-RIM, in which protection is only granted upon a report followed by an investigation and discovery of actual misbehavior by the employer:

**Prediction P-RIM:** In treatment P-RIM: (i) Misbehavior is always reported. (ii) Fraudulent claims do not occur. (iii) The frequency of misbehavior in P-RIM is strictly lower than in NoP (for employers with either employee type) and lower than in P-R (strictly for employers with L-employees and weakly for employers with H-employees).

The results for treatment P-RIM are also displayed in Tables 2 and 3. Intuitively, in P-RIM, by conditioning protection on actual misbehavior, all incentives for fraudulent claims are removed. With respect to truthful claims, the incentives for L-employees (H-employees) are as in treatment NoP (P-R), and hence any misbehavior is reported. The incentive to misbehave is (weakly) decreasing from NoP to P-R to P-RIM. The reason is that in treatment P-RIM,
through her misbehavior decision, the employer can directly affect the employee’s access to protection. As protection is potentially costly for the employer (because of constraining her dismissal decision), this makes the employer more reluctant to misbehave. Comparing treatments $P-R$ and $P-RIM$, whistleblower protection is available in both, but in the benchmark treatment $P-RIM$ the incentive for non-meritorious claims is removed, leading to higher deterrence.

5 Experimental Results

As for our empirical strategy, recall that in each session of the experiment, each subject played 30 periods in a given treatment, but possibly in different roles. Hence, as we observe each subject more than once, in non-parametric tests, our unit of observation are averages on the subject-level (which is explained in more detail in Footnote 26 below), and the resulting numbers of observations are reported in Appendix C. For within-treatment comparisons, we use Wilcoxon Signed-Rank (WSR) tests, while across-treatment comparisons are based on Mann-Whitney-U (MWU) tests. Moreover, in regression analysis, standard errors are clustered at the subject-level. We start by comparing the results of treatment $P-R$ with the benchmark $NoP$, where protection is not available.

5.1 Comparing Treatments $NoP$ and $P-R$

Employers’ Dismissal Decisions: Testing Prediction D  Figure 2 displays the fractions of dismissed workers in treatments $NoP$ and $P-R$, depending on their productivity type and reporting behavior (where the non-feasibility of dismissal upon reporting in treatment $P-R$ is indicated by “n/a”). The results are fully supportive of Prediction D. In particular, in treatment $NoP$, virtually all L-employees are dismissed. In $P-R$ (where dismissal is only feasible when there is no report) this fraction is somewhat lower but still at 0.88, and the difference is not statistically significant according to a MWU test.

Moreover in both treatments, H-employees who do not report are almost always retained. In contrast, and again in line with Prediction D, around 30% of H-employees who do report are dismissed in treatment $NoP$.

\[26\text{As discussed at the end of Section 3, in each of the two treatments shown in Figure 2 there are four conditions under which employers are (repeatedly) observed: with either an L- or an H-employee, who either reports or does not report. For each of these four conditions, we aggregate a given employer’s behavior into an average, and these averages then form the unit of observation in the reported non-parametric tests. We proceed analogously in all other reported non-parametric tests (i.e., when analyzing the behavior of employers, employees, and prosecutors).}\]
This is significantly more compared to non-reporting H-employees (0.30 versus 0.01, WSR, $p < 0.001$) and significantly less compared to L-employees who do report (0.99 versus 0.30, WSR, $p < 0.001$). The considerable fraction of dismissed H-employees is consistent with the prediction of our model, which is based on a utility loss associated with retaining whistleblowers, such that employers might be willing to forgo a higher productivity in order to avoid that loss.

**Employees’ Reporting Decisions: Testing Prediction R** Figure 3 illustrates our results concerning the reporting behavior of employees in treatments *NoP* and *P-R*. It turns out that *Prediction R* is broadly supported as summarized in Table 2. In particular, as for *Prediction R(i)*, the reporting rates of both types are higher when there is misbehavior. These differences are all statistically significant (all with $p < 0.001$, WSR) and hence, with only one exception, also in line with the prediction. The exception is the significant difference in reporting rates of L-employees in treatment *P-R*, for which our theory predicts no difference. In the experiment, at least some L-employees do condition their reporting behavior on whether or not there is misbehavior.

As for *Prediction R(ii)*, in both treatments also the reporting rates of L-employees are generally higher than those of H-employees, irrespective of whether or not misbehavior actually occurred (see, again, Figure 3). Again, all of these four differences are statistically significant. For three of these four differences, this is in line with *Prediction R(ii)*, the exception being the
reporting of misbehavior in treatment $P-R$, where misbehavior should always be reported by either productivity type (see Table 2). For this latter case (i.e., comparing fractions $0.97$ and $0.89$ in Figure 3), a WSR test yields $p < 0.028$. For the other three cases, we have $0.81$ versus $0.38$ ($p < 0.001$), $0.76$ versus $0.64$ ($p < 0.019$), and $0.23$ versus $0.14$ ($p < 0.013$).

Also Prediction $R(iii)$ is broadly supported: Comparing the reporting behavior across treatments for each type of employee reveals that reporting is generally higher in treatment $P-R$. Again, all of these treatment differences are statistically significant (all with $p < 0.0001$, MWU test), although for truthful reports of L-employees (fractions $0.76$ in $P-R$ and $0.97$ in $NoP$) and false reports of H-employees (fractions $0.14$ in $P-R$ and $0.38$ in $NoP$), the absence of a treatment difference was be predicted (see Table 2).

Overall, it can be seen that a whistleblower protection scheme such as $P-R$ leads to a pronounced increase of reporting rates of misbehavior for both productivity types. The downside is that also the fraction of non-meritorious claims rises sharply in treatment $P-R$, in particular by L-employees, which is in line with Prediction $R(iv)$. However, we also observe an (unpredicted) increase in the fraction of non-meritorious claims by H-employees. This issue and its potential implications are discussed in Section 5.2 below.
Prosecutors’ Investigation Decisions: Testing Prediction I

Figure 4(a) illustrates the experimental results for the investigation decisions in treatments NoP and P-R. First, in both treatments, prosecutors indeed seem to perceive an employee’s report as an informative signal about the presence of misbehavior, and the number of investigations is significantly higher when a report occurs (0.93 versus 0.19, and 0.65 versus 0.18, both with $p < 0.001$, WSR).

Moreover, the point predictions of Prediction I are broadly confirmed in treatment NoP, where the fraction of investigations following a report is 0.93 (and hence, indeed close to one as predicted). When there is no report, the fraction of investigations is 0.19 (and hence, somewhat further away from the predicted level of zero).

In treatment P-R, we find very similar results for the case of no report (0.18). However, compared to treatment NoP, the willingness to investigate conditional on a report is significantly lower in P-R (0.65 versus 0.93, $p < 0.001$, MWU test). This difference might be driven by the observed reporting behavior, in particular the high number of non-meritorious claims by H-employees in treatment P-R. A high number of non-meritorious claims reduces the informativeness of reports, and hence might dilute the incentive of prosecutors to trigger costly investigations. This issue will be discussed in more detail in Section 5.2 below.\textsuperscript{27}

\textsuperscript{27}We have also checked how the behavior of prosecutors varies across periods. In particular, there is no period effect in treatment NoP, while the frequency of investigations conditional on a report decreases slightly over time.
Employers’ Decisions to Misbehave: Testing Prediction M  Figure 4(b) displays the fractions of employers who chose to misbehave in treatments NoP and P-R. As can be seen, Prediction M(i) is strongly supported: The fractions of misbehaving employers with L-employees are basically identical in the two treatments (0.38 versus 0.39, where the difference is not statistically significant). Prediction M(ii) is not borne out by the data: For employers matched with H-employees, there is no statistically significant difference in misbehavior between treatments NoP and P-R (0.42 and 0.56, respectively). If anything, there is more (rather than less) misbehavior in P-R compared to the case without whistleblower protection.

To summarize, many of the experimental results for treatments NoP and P-R strongly support the theoretical predictions. However, there are also some deviations, in particular in treatment P-R, which are discussed next.

5.2 Treatment P-R: A Closer Look at Deviations From the Predictions

Relative to the theoretical predictions for treatment P-R, we observe (i) a considerable number of non-meritorious claims by H-employees, (ii) a lower responsiveness of prosecutors to reports, and (iii) no reduction in the level of misbehavior relative to treatment NoP. In this subsection, we first show that the latter two findings can be rationalized given the assumption that prosecutors and employers correctly anticipate the actual behavior in the experiment (rather than the theoretically predicted one). Furthermore, as for the first finding, we investigate potential drivers for sending non-meritorious claims, using the post-experimental questionnaire.

Investigation Decisions and the Informativeness of Reports  Recall from Prediction R that both employee types should exhibit a higher overall willingness to report misbehavior in treatment P-R compared to NoP. However, there is also an incentive for L-employees to send non-meritorious claims. Hence, even from a theoretical perspective, the treatment comparison for the fraction of truthful reporting decisions (i.e., sending a report when there is misbehavior, and not sending a report when there is no misbehavior) is ambiguous. Actual play in the experiment reveals that the fraction of truthful reporting decisions is actually lower in treatment P-R than in NoP (0.66 versus 0.76). As a result, in P-R, reports are less informative about underlying misbehavior: in particular, if there is a report, the empirical frequency of underlying misbehavior is 0.71 in treatment NoP, but only 0.58 in treatment P-R (see Figure 5). In contrast, the empirical frequencies are basically identical when there is no report.
In a next step, using the information in Figure 5, we derive the optimal investigation decisions of prosecutors under the assumption that they correctly take into account the actual empirical relationship between reporting and underlying misbehavior. In addition, we also allow for the possibility that prosecutors internalize the harm from misbehavior inflicted on the third party (with weight $\alpha \in [0,1]$). Hence, for $\alpha = 0$, the prosecutor only cares about his own payoff, while for $\alpha = 1$, he fully internalizes the third party’s harm.\(^{28}\)

Under these assumptions, it is optimal for the prosecutor (i) to refrain from investigating in treatments NoP and P-R if no report is sent (irrespective of $\alpha$), (ii) to investigate in treatment NoP whenever a report is sent (irrespective of $\alpha$), and (iii) to investigate in treatment P-R when a report is sent and at the same time $\alpha > 0.22$ holds.\(^{29}\) Hence, under these assumptions, in treatment P-R it would not necessarily be optimal for the prosecutor to trigger an investigation upon receiving a report. All in all, this modified prediction for P-R is well in line with the

\(^{28}\)Note that the theoretical predictions of Section 4 are based on $\alpha = 0$. However, assuming that prosecutors exhibit (the same) $\alpha > 0$ would not necessitate an extension of the model. The reason is that the case of $\alpha > 0$ would simply correspond to an increase in the cost of not detecting actual misbehavior as incurred by the prosecutor.

\(^{29}\)To see this, recall that prosecutors receive an endowment of 60 points, their investigation cost is 20 points, and their payoff is reduced by 20 (10) points in the case of undetected (detected) misbehavior. Moreover, third parties receive an endowment of 40, which is reduced by 50 (70) points in case of detected (undetected) misbehavior. For example, based on the information in Figure 5 in treatment NoP and conditional on receiving a report, the prosecutor’s expected payoff when choosing $I = 1$ is $0.71 \cdot (50 - 10\alpha) + 0.29 \cdot (40 + 40\alpha)$. Analogously, when choosing $I = 0$ instead, the prosecutor expects a payoff of $0.71 \cdot (40 - 30\alpha) + 0.29 \cdot (60 + 40\alpha)$, which leads to a payoff difference of $1.3 + 14.2\alpha > 0$. Hence, the prosecutor would optimally trigger an investigation independent of $\alpha$. For all the other cases, the calculations are analogous.
lower number of investigations in this treatment as reported in Figure 4(a).

In the following, we proxy $\alpha$ by the offer in the (incentivized) dictator game, which was played at the end of the experiment (where the offer to the other party was an integer between 0 and 100). Table 4 reports the results of a linear probability model for treatment $P-R$. In line with Figure 4(a), a crucial driver for the investigation decision is indeed whether or not a report arrives. Moreover, a higher offer (i.e., a higher $\alpha$) increases the probability of investigation only in the case in which a report is sent. Again, this is in line with our modified prediction. Finally, the effect of social preferences is considerable: For example, for a prosecutor who dictates an equitable outcome (which corresponds to an offer of 50, and which was chosen by around 20% of subjects) the likelihood of an investigation is 27 percentage points larger compared to a prosecutor who keeps everything to himself (which corresponds to $\alpha = 0$).

**Employer Misbehavior** According to Prediction M, there should be no treatment effect on misbehavior for employers with L-employees (which is supported by the data), while, for employers with H-employees, misbehavior should be lower in treatment $P-R$ than in NoP (which is not borne out in the data). We find that, similar to above, the observed relative frequencies of misbehavior across treatments and employee types (see Figure 4(b)) might be rationalized given the assumption that employers correctly anticipate the actual payoff consequences of their misbehavior decision. To show this, in a first step we determine the difference of the employer’s average payoff when choosing $M = 1$ and $M = 0$, respectively, from the experimental data. We do this separately for each treatment and for each productivity type of the employee with whom the employer might be matched. For treatment NoP, these payoff differences are 4.32 when the employer is matched with an L-employee, and 7.37 when the employer is matched with an H-employee. Proceeding analogously for treatment $P-R$, we get 7.66 and 13.18. Note that the

---

30 We do not report the regression results for treatment NoP as the likelihood of investigations is strongly determined by whether or not a report is sent (which is fully in line with Figure 4(a)), and the behavior in the dictator game has no effect.

31 The coefficients for Offer and the interaction term are also jointly significant (F-test, $p < 0.001$).

32 The condition for existence of the informative equilibrium (on which the prediction that every report triggers an investigation is based) is more likely to be satisfied when the cost of undetected misbehavior as incurred by the prosecutor is sufficiently large (for details see Lemma 8 in Appendix A). Moreover, as discussed in footnote 28 above, a higher value of $\alpha > 0$ can simply be interpreted as an increase in this cost, so that the informative equilibrium is more (less) likely to exist when $\alpha$ is high (low). It can be shown that there always exists a babbling equilibrium in which the L-employee (H-employee) always (never) sends a report irrespective of actual misbehavior, and the prosecutor never investigates. However, in treatment $P-R$ reports do trigger a substantial number of investigations (the fraction is 0.65, see Figure 4(a)). This suggests that, in the experiment, prosecutors are heterogenous with respect to their willingness to investigate.
Table 4: Regression Analysis: Investigation Decisions in Treatment P-R

<table>
<thead>
<tr>
<th></th>
<th>Investigate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report</td>
<td>0.344***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Offer</td>
<td>-0.00236</td>
</tr>
<tr>
<td></td>
<td>(0.177)</td>
</tr>
<tr>
<td>Report x Offer</td>
<td>0.00780**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>(0.481)</td>
</tr>
<tr>
<td>Observations</td>
<td>860</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.242</td>
</tr>
</tbody>
</table>

Notes: The table reports the results from a linear probability model with the investigation decision as the dependent variable. P-values are reported in parentheses. Standard errors are clustered at the subject level, where *, **, and *** indicate statistical significance at the 5%, 1%, and .1% level, respectively. Further controls included are: age, gender, proxies for (i) risk aversion, (ii) cognitive reflection, (iii) attitude towards revealing misbehavior, (iv) dutifulness, and (v) a dummy for a major or minor in a field related to economics or business. The coefficients of these controls are all insignificant, and hence are not reported.

ranking of these four payoff differences (4.32, 7.37, 7.66, and 13.18) is the same as the ranking of the corresponding misbehavior frequencies observed in the experiment (0.38, 0.39, 0.42, and 0.56) and as reported in Figure 4(b). Hence, large (small) payoff differences correspond to high (low) levels of misbehavior. While the monetary payoff differences are all positive, an employer will prefer not to misbehave when his (moral) aversion towards misbehavior is sufficiently large. This is more likely to occur the smaller the monetary payoff is in the first place.\textsuperscript{33}

Non-Meritorious Claims by H-Employees  As argued above, in treatment P-R the higher than predicted number of non-meritorious claims by H-employees can rationalize the decisions not to investigate and, in turn, also the decisions to misbehave. That is, prosecutors and employers seem to understand that in treatment P-R (where protection is obtained upon reporting) non-meritorious claims are a crucial issue which they (directly or indirectly) seem to take into account.

\textsuperscript{33}An (unreported) regression akin to the one of Table 4 reveals that the propensity to misbehave is negatively related to employers’ risk aversion (which is line with a similar finding by Minor, 2015) and to the intensity of social preferences (again proxied by the amount offered in the dictator game), where $p = 0.013$ and $p = 0.067$, respectively.
characteristics of the H-employees who lodge them. In total, there are 44 distinct subjects in treatment $P-R$ who played the role of an H-employee and whose employer did not misbehave. Out of these 44 subjects, 16 behaved exactly in line with Prediction R(iv), i.e., they never sent a non-meritorious claim. This is also the modal behavior. However, there is also a substantial fraction of 10 subjects who always sent non-meritorious claims. It turns out that, on average, these subjects exhibit a degree of risk aversion (which, recall, we elicited in the post-experimental questionnaire) that is 0.5 standard deviations higher compared to those 16 subjects who never report. While, in the experiment, the frequency of dismissal of non-reporting H-employees is negligible (see Figure 2), these risk-averse subjects might nevertheless prefer to insure themselves against this risk. This is straightforward to achieve in treatment $P-R$. Below, we will contrast this with treatments $P-RI$ and $P-RIM$ where a report no longer suffices to obtain protection.

Note also that there is no time trend in the frequency of non-meritorious claims across periods. In particular, it is not the case that H-employees (erroneously) file such claims in early periods, while, after experiencing that they are not dismissed when remaining silent, they refrain from reporting in later periods.\footnote{Decomposing the fraction of non-meritorious claims by H-employee (0.38) across periods yields 0.37, 0.40, and 0.38 for periods 1-10, 11-20, and 21-30, respectively.}

### 5.3 Testing the Robustness of Treatment $P-R$

The results from treatment $P-R$ suggest that, when the requirements for protection are relatively low, non-meritorious claims might indeed constitute a serious issue: They seem to dilute the responsiveness of prosecutors to reports, as reports are no longer good indicators for underlying misbehavior. In turn, this seems to reduce the deterrence effect of the whistleblower protection scheme. Of course, at the present stage, it would be premature to derive sound policy conclusions from the admittedly stylized treatment $P-R$. For example, it is assumed that (i) a mere (yet ex ante reasonable) report suffices to obtain protection (independent of whether or not an investigation is actually triggered) and that (ii) lodging a non-meritorious claim is morally cheap for the employee, as it does not impose any additional cost on the employer. In this section, we consider two further treatments, which address these issues in turn (see Table 1 in Section 3): First, in treatment $P-RI$ a whistleblower obtains protection if and only if his report also triggers an investigation by the prosecutor. Second, in treatment $RL$
any report that triggers an investigation (whether truthful or not) leads to a reputation loss for the employer. In what follows, we focus on the reporting, investigation, and misbehavior decisions. As for dismissals, the results remain strongly in line with Prediction D, and hence are not reported here.\footnote{For example, conditional on dismissal being feasible, the fraction of dismissed L-employees (H-employees) is 0.93 (0.00) in treatment $P-RI$.}

**Treatment $P-RI$** The only difference between treatments $P-R$ and $P-RI$ is that in the latter the employee only obtains protection when his report also triggers an investigation. As discussed in Section 4, the theoretical predictions for both of these treatments coincide. In fact, this is also borne out in the experiment. The results for treatment $P-RI$ are illustrated in Figures 6 and 7 where, for the sake of comparability, we repeat the results for treatments NoP and $P-R$. It turns out that there are no statistically significant differences compared to treatment $P-R$ with respect to either reporting or misbehavior. In particular, the number of non-meritorious claims does not drop significantly compared to $P-R$. That is, none of the pair-wise tests (MWU) for differences in reporting behavior between treatment $P-R$ (0.97, 0.81, 0.89, 0.38) and $P-RI$ (0.9, 0.72, 0.81, 0.27) is statistically significant (see Figure 6(b) and (c)). The tests for treatment differences with respect to misbehavior and investigations (see Figure 7) are performed analogously. Here, the only significant difference occurs with respect to the frequency of investigations conditional on a report, which is higher in treatment $P-RI$ (0.79 versus 0.65, $p = 0.02$, MWU).

**Treatment RL: Reputation Cost of Investigations** Recent findings from the experimental literature on lying (see, e.g., Gneezy, 2005; Gneezy, Rockenbach, and Serra-Garcia, 2013; Fischbacher and Föllmi-Heusi, 2013) suggest that many individuals are subject to lying aversion. Moreover, as shown in Gneezy (2005), this aversion seems to be the stronger the bigger the harm imposed on others through a lie. It is therefore interesting to see whether the non-meritorious claims in treatments $P-R$ and $P-RI$ are (partially) driven by the fact that, apart from the fact that a protected whistleblower cannot be dismissed, filing a non-meritorious claim does not impose any further cost on the employer. In practice, such an additional cost might, for example, come in the form of a reputation loss triggered by an investigation. Hence, as a robustness check, we ran treatment $RL$, which differs from $P-RI$...
Figure 6: Varying the Requirements for Protection: Fraction of Employees Sending a Report

(a) NoP

(b) P-R

(c) P-RI

(d) P-RIM

(e) IE

(f) RL

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior

L-Emp. | H-Emp.
--- | ---
No Misbehavior | Misbehavior
only by the fact that the employer’s payoff is reduced by 10 points whenever an investigation occurs. We find, however, that this does not affect behavior. In particular, all pair-wise tests for treatment differences between RL and P-RI (with respect to reporting, investigations, and misbehavior) are not significant (for an illustration, see Figures 6 and 7).

In summary, in all treatments considered so far where whistleblower protection is available, there is a sizeable number of non-meritorious claims. Moreover, while protecting honest whistleblowers, the predicted improvement in deterrence does not materialize in these treatments. To investigate in more detail the interplay of non-meritorious claims and deterrence in the presence of whistleblower protection, we now consider a further benchmark treatment P-RIM, where protection is available, but where all incentives for non-meritorious claims are removed.

5.4 Disincentivizing Non-Meritorious Claims

Treatment P-RIM To obtain protection in the benchmark treatment P-RIM, in addition to a report and an investigation (as in P-RI), actual misbehavior is also required. As stated in Prediction P-RIM (i) and (ii) above, relative to P-R, this change should not affect the truthful reporting of misbehavior, but the incentives for non-meritorious claims are eliminated. Indeed, this prediction is strongly supported by the experimental data (see Figure 6(d)): Note first that the truthful reporting of misbehavior (black bars) remains at high levels and there is no statistically significant difference to either treatment P-R or P-RI (again using pair-wise comparisons). In fact, in all treatments with whistleblower protection, the willingness to report misbehavior is significantly higher (at the 1% or 5% level, MWU tests) compared to NoP.

In P-RIM, non-meritorious claims (white bars) indeed go down strongly for both productivity types, and they are significantly lower than in any other treatment with whistleblower protection. For example, compared to P-RI, they are virtually fully eliminated for H-employees (a drop from 27% to 5%), and, for L-employees, we also observe a substantial decrease of more than 50 percentage points (both effects with p < 0.001, MWU). The fractions of non-meritorious claims are also lower in P-RIM compared to treatment NoP, but these differences (0.19 versus 0.23, and 0.05 versus 0.14) are not statistically significant (MWU).

Furthermore, as can be seen in Figure 7(a), in treatment P-RIM the frequency of investigations upon receiving a report is significantly higher compared to P-R (p < 0.01, MWU). This finding is in line with the above reasoning: the low number of non-meritorious claims
in \textit{P-RIM} seems to lead to a higher responsiveness of prosecutors to reports. However, in all treatments with whistleblower protection, the frequency of investigations conditional on receiving a report is significantly lower compared to \textit{NoP} (in all cases, \( p < 0.05 \), MWU). In contrast, when reporting does not occur, we find no significant difference in any pair-wise comparison. Together, this suggests that prosecutors seem to have more doubts about the truthfulness of reports when a whistleblower protection scheme is in place.

Consider next the frequency of employer misbehavior in \textit{P-RIM} (see \textit{Prediction WPB3}(iii) and Table 3). As displayed in Figure 7(b), for employers with L-employees, the frequency of misbehavior in \textit{P-RIM} (0.2) is significantly lower than in \textit{P-R} (0.39; \( p < 0.01 \), MWU),
which is in line with the prediction. However, the difference to NoP (0.38) is not significant ($p = 0.16$, MWU). For employers with H-employees, as predicted there is no statistically significant treatment difference between P-RIM and P-R. As discussed above, between P-R and NoP there is no treatment difference for employers matched with H-employees. It is therefore not surprising that the same holds true for the comparison of P-RIM and NoP.

To summarize, treatment P-RIM is effective in bringing down non-meritorious claims without decreasing truthful reporting and also fares well with respect to deterrence. It serves as a theoretical benchmark to analyze how non-meritorious claims and whistleblower protection affect prosecutor behavior (and hence, deterrence). Treatment P-RIM is not meant as a policy recommendation, because it abstracts from a number of important real-world aspects.

For example, in treatment P-RIM the merit of a claim is revealed immediately in the course of an investigation. In practice, however, in constituencies where protection is only granted after the presence of misbehavior has been established in court, it often takes a painfully long time for a whistleblower until it is clear whether or not protection is awarded.36

Moreover, even if there were no such time lag, whenever the court finding misbehavior is a necessary condition for protection, whistleblowers will worry about investigation errors, as protection is not granted if their case is erroneously rejected. In all treatments considered so far, it was assumed that an investigation perfectly reveals whether or not there was misbehavior. We have relaxed this assumption in a further robustness treatment IE. In particular, this treatment modifies P-RIM by assuming that investigations are prone to type-1 error (i.e., an investigation detects misbehavior only with probability 0.7). Hence, even a truthful report no longer necessarily ensures protection, while exposing the employee as a whistleblower. As can be seen in Figure 6, in contrast to the comparison of treatments P-RIM and P-R, there is significantly less truthful reporting of each employee type in treatment IE compared to P-R.37 Moreover, some of the (reduced number of) whistleblowers do not obtain protection. Finally, with respect to deterrence, there is no significant difference in the frequency of misbehavior between treatments IE and P-R (seen Figure 7(b)). Hence, the decline in misbehavior in

---

36For example, this is highlighted by the prominent Heinisch v. Germany case, where several German courts had refused to reverse the dismissal of a whistleblower (a geriatric nurse who had (correctly) reported misbehavior by her employer) before protection was eventually affirmed by the European Court of Human Rights (see, e.g., Thusing and Forst, 2016, pp. 12).

37In particular, 0.86 versus 0.97, $p = 0.002$ for L-employees and 0.76 versus 0.89, $p = 0.027$ for H-employees (both MWU). With respect to non-meritorious claims, as in the comparison of treatments P-RIM and P-R, there is also a strong and highly significant decline (0.81 versus 0.22, $p = 0.001$ for L-employees and 0.38 to 0.17, $p = 0.001$ for H-employees, both MWU) for the comparison of treatments IE and P-R.
treatment $P-RIM$ compared to $P-R$ is offset when enforcement errors are present.

## 6 Conclusion

In this paper, we have studied employee whistleblowing as an instrument to fight corporate fraud. To this end, we have considered, experimentally and theoretically, a setting where employees (as potential whistleblowers) interact with employers (as potential wrong-doers) and prosecutors (who may investigate the allegations of whistleblowers against their employers).

Our main goals was to study the effects whistleblower protection on reporting behavior (both truthful and non-meritorious) and deterrence. To this end, we have analyzed a treatment ($P-R$) which is meant to capture a stylized version of current whistleblower laws (where these laws also serve as a role model for recommendations by the OECD and the G20 group). We have compared this scheme with two benchmark treatments, one where protection is not available ($NoP$), and one where protection is conditional on whether the whistleblower’s claim turns out to be true in the course of an investigation ($P-RIM$). We find that whistleblower protection indeed has the intended effect of enhancing the truthful reporting of existing misbehavior.

We also find that non-meritorious claims indeed play a significant role when a whistleblower protection scheme provides incentives to lodge them. Moreover, non-meritorious claims makes reports less informative about the presence of misbehavior, and prosecutors react by investigating less often when receiving a report. In turn, this effect makes misbehavior more attractive. We indeed find that, despite the higher willingness for truthful reporting of misbehavior, the desired improvement in deterrence is not realized. This latter effect does not arise in our benchmark where the incentive to file non-meritorious claims is eliminated. Overall, our results suggest that the issue of non-meritorious claims and the potential reactions of other parties as a result of such claims, should be seriously taken into account in the discussion on how to design whistleblower protection schemes.
References


Appendix (For Online Publication)

A Theory

This Appendix is structured as follows: In Section A.1, the model is presented, and in Section A.2, we derive the equilibrium outcome for each treatment. We focus on pure-strategy Perfect Bayesian Equilibria that are informative equilibria in the sense that the prosecutor triggers an investigation if and only if the employee sends a report. Hence, we do not consider babbling equilibria throughout. The theoretical predictions of Section 4 follow immediately from Propositions 1 - 4. The comparisons of the fractions of employers who misbehave (as stated in Table 3) are derived at the end of Section A.2.

A.1 Model

The Game Played  We consider a game played by three players, an employer, an employee, and a prosecutor (see also Figure 1 in the main text). The employer (she) is matched with an employee of type $\theta$ whose productivity $x_\theta$ the employer appropriates. In addition, the employer decides whether or not to misbehave denoted by $M \in \{0, 1\}$ (where $M = 0$ indicates no misbehavior), which is observed by the employee, but not by the prosecutor.

The employee has productivity $x_\theta$, $\theta = L, H$, which is either high ($\theta = H$: H-employee) or low ($\theta = L$: L-employee, where $x_H > x_L$). This productivity is known to the employer but not to the prosecutor who only knows that there is a share $h \in (0, 1)$ of H-employees in the population. The employee’s only choice is whether or not to send a report $R \in \{0, 1\}$ to the prosecutor indicating that the employer engaged in misbehavior, where $R = 1$ indicates that the employee sends a report. As a tie-breaking rule, we assume that employees refrain from reporting when being indifferent between reporting and not reporting. The prosecutor always observes whether or not a report is sent. In treatments NoP and P-R, this is also observed by the employer. In treatments P-RI and P-RIM, the employer learns the reporting decision in the course of an investigation (see also the discussion in Section 3 above).

After the employee’s reporting decision, the prosecutor decides on initiating an investigation, $I \in \{0, 1\}$, where $I = 1$ indicates an investigation. Upon investigating the prosecutor learns

---

38 As discussed in Section 5.2, in the experiment we have added a “third party”, which is a purely passive player without any decisions to take. In the experiment, it is only included to make it more salient that misbehavior causes harm to others.

39 Our results would also hold in a model where employees face an arbitrarily small reporting cost.
whether or not the employer indeed has misbehaved. Whether or not an investigation is initiated and whether or not the employer is found to be guilty is publicly observable.

Finally, before production eventually takes place, the employer decides whether or not to dismiss the employee, \( D \in \{0, 1\} \), where \( D = 1 \) indicates a dismissal. A dismissed employee is replaced by an outsider of some intermediate productivity \( \overline{x} \), with \( x_L < \overline{x} < x_H \). In this case, the employee appropriates the outsider’s productivity.

**Treatments** We capture the following four legal regimes, which correspond to the four main treatments in the experiment (see Table 1 in the main text): In treatment NoP, the employer is free to dismiss the employee. In treatment P-R, a dismissal is prohibited if and only if \( R = 1 \). In treatment P-RI, a dismissal is prohibited if and only if \( R = I = 1 \). Finally, in treatment P-RIM, a dismissal is prohibited if and only if \( R = I = M = 1 \).

**Payoffs** All payoffs (monetary and non-monetary) are summarized in Table 5. First, the payoff of the employer depends on whether or not she misbehaves, whether or not there is an investigation, and whether or not she employs a whistle-blower. The employer’s potential net gain \( y \) from misbehavior consists of a monetary payoff \( z \) minus some disutility from misbehavior \( \zeta \) (which might reflect moral reservations of the employer). We assume that \( \zeta \) is randomly distributed (and the realization is private information of the employer), and hence this is also the case for \( y \). In particular, we assume that \( y \) is distributed according to \( H(\cdot) \), with full support, and mean \( \overline{y} \). If the prosecutor investigates and there is misbehavior, the employer faces an (exogenously given) fine \( f > 0 \). The employer receives the employee’s or the outside replacement’s productivity (i.e., \( x_L, x_H, \) or \( \overline{x} \)) and pays a fixed wage \( \omega \). Last, but not least, the employer dislikes employing a whistle-blower, and the respective disutility is denoted by \( \tau > 0 \). It is drawn from a distribution \( G(.) \), and it is the employer’s private information. The employer forms a belief \( \beta \in [0, 1] \) that her employee has sent a report.

Second, the employee gets a fixed wage \( \omega \) if he is not dismissed by the employer, and zero otherwise. In addition, misbehavior that remains undetected by the prosecutor imposes a disutility \( \delta > 0 \) on the employee, which could reflect a preference for conscience cleaning as discussed in the main text, and which is the employee’s private information. From the viewpoint of the other players, \( \delta \) is drawn from a distribution \( F(\cdot) \). We assume \( F(\omega) < 1 \) which ensures that there exist values of \( \delta \) for which the respective disutility outweighs the
Table 5: Payoffs

(a) **NoP, P-R, P-RI, P-RIM**: Payoffs When There is No Protection

<table>
<thead>
<tr>
<th>Misbehavior</th>
<th>Investigation</th>
<th>Dismissal</th>
<th>Employee</th>
<th>Prosecutor</th>
<th>Employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>ω</td>
<td>0</td>
<td>$(x_i - ω) - β \cdot τ$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$(x - ω)$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>ω</td>
<td>$-K_1$</td>
<td>$(x_i - ω) - β \cdot τ$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$-K_1$</td>
<td>$(x - ω)$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>$-K_1 - K_2$</td>
<td>$(x_i - ω) + y - f - β \cdot τ$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>ω</td>
<td>$-K_1 - K_2$</td>
<td>$(x_i - ω) + y - f - β \cdot τ$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>$-δ$</td>
<td>$-K_2 - K_3$</td>
<td>$(x - ω) + y$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$ω - δ$</td>
<td>$-K_2 - K_3$</td>
<td>$(x_i - ω) + y - β \cdot τ$</td>
</tr>
</tbody>
</table>

(b) **P-R**: Payoffs When There is Protection

<table>
<thead>
<tr>
<th>Misbehavior</th>
<th>Investigation</th>
<th>Dismissal</th>
<th>Employee</th>
<th>Prosecutor</th>
<th>Employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>n/a</td>
<td>ω</td>
<td>0</td>
<td>$(x_i - ω) - β \cdot τ$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>n/a</td>
<td>ω</td>
<td>$-K_1$</td>
<td>$(x_i - ω) - β \cdot τ$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>n/a</td>
<td>ω</td>
<td>$-K_1 - K_2$</td>
<td>$(x_i - ω) + y - f - β \cdot τ$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>n/a</td>
<td>$ω - δ$</td>
<td>$-K_2 - K_3$</td>
<td>$(x_i - ω) + y - β \cdot τ$</td>
</tr>
</tbody>
</table>

(c) **P-RI**: Payoffs When There is Protection

<table>
<thead>
<tr>
<th>Misbehavior</th>
<th>Investigation</th>
<th>Dismissal</th>
<th>Employee</th>
<th>Prosecutor</th>
<th>Employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>n/a</td>
<td>ω</td>
<td>$-K_1$</td>
<td>$(x_i - ω) - β \cdot τ$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>n/a</td>
<td>ω</td>
<td>$-K_1 - K_2$</td>
<td>$(x_i - ω) + y - f - β \cdot τ$</td>
</tr>
</tbody>
</table>

(d) **P-RIM**: Payoffs When There is Protection

<table>
<thead>
<tr>
<th>Misbehavior</th>
<th>Investigation</th>
<th>Dismissal</th>
<th>Employee</th>
<th>Prosecutor</th>
<th>Employer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>n/a</td>
<td>ω</td>
<td>$-K_1 - K_2$</td>
<td>$(x_i - ω) + y - f - β \cdot τ$</td>
</tr>
</tbody>
</table>

Notes: The table depicts the players’ payoffs as a function of the employer’s misbehavior and dismissal decisions and the prosecutor’s investigation decision. As the employee’s reporting decision (R) has no direct effect on the payoff of neither player, we omit a separate column for the sake of readability. The payoffs in Panel (a) apply whenever the employee is not shielded from dismissal, and hence (i) always in treatment NoP, (ii) in treatment P-R if $R = 0$ holds, (iii) in treatment P-RI either if $R = 0$ holds or if both $R = 1$ and $I = 0$ hold, and (iv) in treatment P-RIM if $R = I = M = 1$ does not hold. The payoffs in Panel (b) apply in treatment P-R if the employee is protected from dismissal (i.e., if $R = 1$). The payoffs in Panel (c) apply in treatment P-RI if the employee is protected from dismissal (i.e., if $R = I = 1$). The payoffs in Panel (c) apply in treatment P-RIM if the employee is protected from dismissal (i.e., if $R = I = M = 1$).
(H-employee’s) fear of dismissal. Moreover, in case of undetected misbehavior, δ accrues to the employee independently of whether or not he is dismissed.

Finally, the payoff of the prosecutor depends on whether there is misbehavior and whether an investigation takes place. When there is no misbehavior, the prosecutor’s payoff is \(-K_1(0)\) if he investigates (does not investigate). Hence, \(K_1 > 0\) can be considered as investigation costs. When there is misbehavior, his payoff is \(-K_1 - K_2\) if he investigates and \(-K_2 - K_3\) if he does not investigate, where we assume \(K_3 > K_1\).\(^{40}\) Hence, when there is (no) misbehavior, the prosecutor’s payoff is higher if he conducts (does not conduct) an investigation.

A.2 Equilibrium Analysis

A.2.1 Preliminaries

When deriving our predictions, we focus on Perfect Bayesian Equilibria (PBE) in pure strategies (i.e., all players choose best responses given their beliefs and given the strategies of the other players, where beliefs are formed in accordance with Bayes’ Rule whenever possible), that are informative in the sense that the prosecutor’s investigation decision varies with the employee’s report:

Definition 1. A Perfect Bayesian Equilibrium is called informative equilibrium if the prosecutor’s equilibrium strategy is given by \(I(R) = R\) for all \(R \in \{0, 1\}\).

In the following, we provide conditions for the existence of an informative equilibrium under each treatment, and we assume that it is always played given that it exists. To derive our predictions, we proceed as follows: First, under the assumption that the prosecutor plays his equilibrium strategy \(I^*(R) = R\), we characterize optimal behavior with respect to misbehavior, reporting, and dismissal, denoted by \(M^*(\cdot), R^*(\cdot),\) and \(D^*(\cdot)\), respectively. Note that in informative equilibrium, the employer’s belief that the employee has sent a report satisfies \(\beta^* \in \{0, 1\}\). Second, we derive conditions under which \(I^*(R) = R\) is in fact optimal for the prosecutor (i.e., for each treatment, we provide conditions that ensure existence of informative equilibrium). Third, this leads to the equilibrium outcome, which depends on the realizations of the random variables \(\delta, \tau,\) and \(y\) (where these realizations are unknown to the experimenter). Taking into account the prior distributions of these random variables, the predictions of Section 4 are then based on the expected equilibrium outcomes (see Propositions 1 - 4).

\(^{40}\)The experimental payoffs of the prosecutor as reported in the main text are obtained when setting \(K_2 = -10\) and \(K_3 = 30\).
A.2.2 Treatment NoP: Equilibrium Outcome

In the following, we assume that the report is observed by both the prosecutor and the employer (as in the experiment), and we solve the game backwards, starting with the employer’s dismissal decision at date 4 (see Figure 1 in the main text). In doing so, we write $D^*(\cdot)$ as a function of I rather than $R$, because $I = R$ for all $R \in \{0, 1\}$ in the informative equilibrium:

**Lemma 1 (NoP: Dismissal).** In the informative equilibrium, the following holds: The L-employee is always dismissed. The H-employee is dismissed only if both a report occurs and the employer’s disutility from retaining a known whistle-blower is sufficiently large. That is,

$$D^*(x_\theta, I, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L, \\
1 & \text{if } x_\theta = x_H, \text{ and } R = 1 \text{ and } \tau > \tau, \text{ and} \\
0 & \text{else},
\end{cases}$$

where $\tau := x_H - \overline{x}$.

**Proof.** First, when $R = I = 0$, the employer gets $x_\theta$ if retaining the employee, and $\overline{x}$ if dismissing him. Since $x_L < \overline{x} < x_H$, in this case, the L-employee (H-employee) is dismissed (retained).

Second, when $R = I = 1$, the employer gets $x_\theta - \tau - M \cdot f$ if retaining the employee and $\overline{x} - M \cdot f$ if dismissing him. Hence, the L-employee is again dismissed, while the H-employee is dismissed only if $\tau$ is sufficiently large, i.e., for $\tau > \tau := x_H - \overline{x}$.

In the informative equilibrium, the employee’s optimal reporting behavior at date 2 can be characterized as follows:

**Lemma 2 (NoP: Reporting).** In the informative equilibrium, the following holds: The L-employee reports if and only if the employer misbehaves. The H-employee reports if and only if there is both misbehavior and his disutility $\delta$ from undetected misbehavior is sufficiently large. That is,

$$R^*(x_\theta, M, \delta) = \begin{cases} 
1 & \text{if } M = 1 \text{ and } x_\theta = x_L, \\
1 & \text{if } M = 1, \text{ } x_\theta = x_H \text{ and } \delta > \overline{\delta}, \text{ and} \\
0 & \text{else},
\end{cases}$$

where $\overline{\delta} := (1 - G(\overline{\tau})) \cdot \omega$.

**Proof.** The L-employee is always dismissed independent of his reporting decision (see Lemma 1). Hence, the L-employee’s payoff is $-\delta \cdot M$ if he does not report and 0 if he reports. Again from Lemma 1, when not reporting, the H-employee is not dismissed, and hence gets $\omega - \delta \cdot M$. Upon reporting, he is retained with probability $G(\overline{\tau})$, and hence his payoff is $G(\overline{\tau}) \cdot \omega$. 

\[\square\]
Next, consider the employer’s misbehavior decision at date 1:

**Lemma 3 (NoP: Misbehavior).** *In the informative equilibrium, the employer’s misbehavior decision is given by:*

\[
M^*(x_\theta, y, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L \text{ and } y > f, \\
1 & \text{if } x_\theta = x_H \text{ and } \tau < \overline{\tau} \text{ and } y > y_1, \\
1 & \text{if } x_\theta = x_H \text{ and } \tau > \overline{\tau} \text{ and } y > y_2, \text{ and} \\
0 & \text{else},
\end{cases}
\]

*where* \( y_1 := (1 - F(\delta))(f + \tau) \) *and* \( y_2 := (1 - F(\delta))(x_H - \overline{\tau} + f) \).

**Proof.** First, suppose the employer faces an L-employee. In this case, Lemmas 1 and 2 imply that the employer’s payoff is \( x + y - \omega - f \) if she misbehaves, and \( x - \omega \) if she does not misbehave. Hence, misbehavior is optimal if \( y > f \). Second, consider the situation where the employer is facing an H-employee. When the employer chooses \( M = 0 \), then Lemmas 1 and 2 imply that her payoff is \( x_H - \omega \). When choosing \( M = 1 \) instead, then the employer’s payoff also depends on the subsequent dismissal decision, and hence it also depends on \( \tau \). Case (i): \( \tau < \overline{\tau} \) (no subsequent dismissal). From Lemma 2, it follows that the employer’s expected payoff when choosing \( M = 1 \) is \( x_H + y - \omega - (1 - F(\delta))(f + \tau) \). In this case, the employer optimally misbehaves if \( y > y_1 := (1 - F(\delta))(f + \tau) \). Case (ii): \( \tau > \overline{\tau} \) (subsequent dismissal). Here, the expected payoff from choosing \( M = 1 \) is \( y - \omega + F(\delta)x_H + (1 - F(\delta))(\overline{\tau} - f) \). In this case, the employer optimally misbehaves if \( y > y_2 := (1 - F(\delta))(x_H - \overline{\tau} + f) \).

Finally, consider the prosecutor’s investigation decision, and recall that the prosecutor does not observe the employee’s productivity. Define the prosecutor’s equilibrium belief with respect to misbehavior conditional on \( R \) as \( B_0 := \Pr \{ M = 1 \mid R = 0 \} \) and \( B_1 := \Pr \{ M = 1 \mid R = 1 \} \). Given Lemmas 1 - 3, in equilibrium this leads to \( B_1 = 1 \) (as there are no non-meritorious claims) and \( B_0 < 1 \) (as misbehavior is not always reported). In particular,

\[
B_0 = \frac{h \cdot p^0_H \cdot F(\delta)}{h \cdot (p^0_H \cdot F(\delta) + 1 - p^0_H) + (1 - h) \cdot H(f)},
\]

where

\[
p^0_H := G(\overline{\tau}) E_{\tau} [1 - H(y_1) \mid \tau < \overline{\tau}] + (1 - G(\overline{\tau}))(1 - H(y_2))
\]

and where in (2) expectations are formed over \( \tau \) (as \( y_1 \) is a function of \( \tau \)). Intuitively, in (1) the numerator states the probability of unreported misbehavior (recall that this occurs with
H-employees only), and the denominator states the overall probability that no report is sent.

**Lemma 4 (NoP: Investigation).** Given the behavior of the other players as described in Lemmas 1 - 3, if \( B_0 \leq \frac{K_1}{K_3} \) holds, then choosing \( I^*(R) = R \) is optimal for the prosecutor.

**Proof.** First, if \( R = 0 \), upon choosing \( I = 0 \), the prosecutor’s expected payoff is \(-B_0 \cdot (K_3 + K_2)\). When choosing \( I = 1 \) instead, he gets \(-K_1 - B_0 \cdot K_2\). Hence, given \( R = 0 \), \( I = 0 \) is optimal iff \( B_0 \leq \frac{K_1}{K_3} \). Second, if \( R = 1 \), when choosing \( I = 0 \), the prosecutor’s expected payoff is \(-B_1 \cdot (K_3 + K_2)\). When choosing \( I = 1 \) instead, he gets \(-K_1 - B_1 \cdot K_2\). Hence, given \( R = 1 \), \( I = 1 \) is optimal iff \( B_1 > \frac{K_1}{K_3} \). Since in equilibrium \( B_1 = 1 \), this is always satisfied (recall that \( K_1 < K_3 \) by assumption).  

Lemmas 1 to 4 characterize behavior in informative equilibrium. As this also depends on the random variables \( \tau, \delta \) and \( y \) (which are unobservable to the experimenter), we now state the expected equilibrium outcome given the prior distributions of these random variables. This expected equilibrium outcome is the basis for the predictions in Section 4:

**Proposition 1 (NoP: Expected Equilibrium Outcome).** The informative equilibrium in treatment NoP has the following expected equilibrium outcome: (i) L-employees always (never) report if there is (no) misbehavior. (ii) L-employees are always dismissed. (iii) Given misbehavior, the probability of observing a report by an H-employee is \( E_\delta [R^*(x_H, 1, \delta)] = 1 - F(\delta) \), and, in the absence of misbehavior, H-employees never send a report. (iv) Given that an H-employee sends a report, the probability of observing his dismissal is \( E_\tau [D^*(x_H, 1, \tau)] = 1 - G(\tau) \), while when sending no report, he is never dismissed. (v) The probability of observing misbehavior by the employer when matched with an L-employee is \( m_{L}^{\text{no}} := E_{y, \tau} [M^*(x_L, y, \tau)] = 1 - H(f) \). (vi) The probability of observing misbehavior by the employer when matched with an H-employee is \( m_{H}^{\text{no}} := E_{y, \tau} [M^*(x_H, y, \tau)] = p_{H}^{0} \) as defined in (2). (vii) When (not) receiving a report, prosecutors always (never) trigger an investigation.

### A.2.3 Equilibrium Outcome in Treatment P-R

Again, we assume that the report is observed by both the prosecutor and the employer (as in the experiment), and we solve the game backwards:

**Lemma 5 (P-R: Dismissal).** In the informative equilibrium, the following holds: The L-employee is dismissed whenever this is feasible (i.e., if \( R = 0 \)). The H-employee is never
dismissed. That is,

\[ D^*(x_\theta, I, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L \text{ and } R = 0, \text{ and} \\
0 & \text{else.}
\end{cases} \]

Proof. In treatment P-R, a dismissal is only feasible when \( R = 0 \). Analogously to Lemma 1, the L-employee is always dismissed (when feasible). Moreover, the employer might only want to dismiss the H-employee, if the latter sends a report (which, however, is not feasible). \( \square \)

In informative equilibrium, the employee’s optimal reporting behavior at date 2 can be characterized as follows:

**Lemma 6 (P-R: Reporting).** *In the informative equilibrium, the following holds: The L-employee always sends a report, irrespective of whether or not there is misbehavior. In contrast, the H-employee sends a report if and only if there is misbehavior. That is,*

\[ R^*(x_\theta, M, \delta) = \begin{cases} 
1 & \text{if } x_\theta = x_L, \\
1 & \text{if } x_\theta = x_H \text{ and } M = 1, \text{ and} \\
0 & \text{else.}
\end{cases} \]

**Proof.** From Lemma 5, the L-employee anticipates that he will be dismissed unless sending a report (thereby obtaining protection). For \( M = 1 \), his payoff upon choosing \( R = 1 \) is \( \omega \) (since the report triggers an investigation), while he would get only \( -\delta \) when choosing \( R = 0 \) instead. For \( M = 0 \), the L-employee still gets \( \omega \) upon choosing \( R = 1 \), but would get zero upon choosing \( R = 0 \). Hence, always sending a report is optimal for the L-employee. An H-employee who observes \( M = 1 \) gets \( \omega \) when choosing \( R = 1 \), and \( \omega - \delta \) when choosing \( R = 0 \). If \( M = 0 \), he gets \( \omega \) regardless of his reporting decision. Since we assume no reporting in case of indifference, the optimal response to \( M = 0 \) is \( R = 0 \). \( \square \)

Next, consider the employer’s misbehavior decision at date 1.

**Lemma 7 (P-R: Misbehavior).** *In the informative equilibrium, the employer’s misbehavior decision is given by:*

\[ M^*(x_\theta, y, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L \text{ and } y > f, \\
1 & \text{if } x_\theta = x_H \text{ and } y > f + \tau, \text{ and} \\
0 & \text{else.}
\end{cases} \]

**Proof.** Given Lemmas 5 and 6, when matched with an L-employee, the employer anticipates that the employee always reports, and hence always triggers an investigation. Therefore, when
choosing $M = 1$, the employer gets $x_L - \omega + y - f - \tau$. Upon choosing $M = 0$, she gets $x_L - \omega - \tau$. By contrast, when matched with an H-employee, the employer anticipates that a report is sent if and only if $M = 1$ is chosen. Hence, upon choosing $M = 1$ she gets $x_H - \omega + y - f - \tau$, and $x_H - \omega$ upon choosing $M = 0$.

Finally, consider the prosecutor’s investigation decision. Given Lemmas 5 - 7, his equilibrium beliefs with respect to misbehavior conditional on $R$ are given by $B_0 = 0$ (in equilibrium, any misbehavior is reported) and

$$B_1 = \frac{h \cdot p_H^1 + (1 - h) \cdot (1 - H(f))}{h \cdot p_H^1 + (1 - h)} \in (0, 1),$$

where

$$p_H^1 := E_\tau [1 - H(f + \tau)].$$

**Lemma 8 (P-R: Investigation).** Given the behavior of the other players as described in Lemmas 5 - 7, if $\frac{K_1}{K_3} \leq B_1$ holds, then choosing $I^*(R) = R$ is optimal for the prosecutor.

*Proof.* First, if $R = 0$, then, when choosing $I = 0$, the prosecutor’s expected payoff is $-B_0 \cdot (K_3 + K_2) = 0$ due to $B_0 = 0$. When choosing $I = 1$ instead, the prosecutor gets $-K_1 - B_0 \cdot K_2 < 0$, which is strictly worse. Second, if $R = 1$, when choosing $I = 0$, the prosecutor’s expected payoff is $-B_1 \cdot (K_3 + K_2)$. When choosing $I = 1$ instead, he gets $-K_1 - B_1 \cdot K_2$. Hence, given $R = 1$, $I = 1$ is optimal iff $\frac{K_1}{K_3} \leq B_1$.

Note that the condition $\frac{K_1}{K_3} \leq B_1$ in Lemma 8 can be reformulated as follows (inserting for $B_1$): $\frac{K_1}{K_3} \leq \frac{1}{1 + \frac{1}{1 - H(f)}}$. Furthermore, note that $\frac{1}{1 - H(f)}$ represents the fraction of the reporting frequency of L-employees (which is 1) and the frequency of misbehavior of employers matched with L-employees (which is $1 - H(f)$). This fraction is nothing else than the inverse of the measure of informativeness of the reports sent by L-employees: If it is 1, these reports are perfectly informative; but for higher values, they are less informative. If the percentage of non-meritorious claims among reports sent by L-employees converges to 100%, i.e., if the frequency of misbehavior of employers matched with L-employees converges to zero, the measure of informativeness converges to zero, too; its inverse converges to infinity, and thus, $B_1$ converges to zero. Hence, for $\frac{K_1}{K_3}$ bounded away from zero, the condition for existence of the informative equilibrium is violated for a sufficiently high percentage of non-meritorious claims among reports sent by L-employees. For now, we assume that the condition is not violated. We come
back to the possibility of violation below.

Lemmas 5 - 8 characterize behavior in informative equilibrium. As this also depends on the random variables $\tau$, $\delta$ and $y$ (which are unobservable to the experimenter), we now state the *expected* equilibrium outcome given the prior distributions of these random variables. This expected equilibrium outcome is the basis for the predictions in Section 4:

**Proposition 2** (*P-R: Expected Equilibrium Outcome*). *The informative equilibrium in treatment P-R has the following expected equilibrium outcome: (i) L-employees send a report regardless of whether or not there is misbehavior. (ii) L-employees are never dismissed. (iii) H-employees always (never) report if there is (no) misbehavior. (iv) H-employees are never dismissed. (v) The probability of observing misbehavior by the employer when matched with an L-employee is $m_{L} = E_{y,\tau}[M^{*}(x_{L},y,\tau)] = 1 - H(f)$. (vi) The probability of observing misbehavior by the employer when matched with an H-employee is $m_{H} = E_{y,\tau}[M^{*}(x_{H},y,\tau)] = p_{H}$ as defined in (4). (vii) When (not) receiving a report, prosecutors always (never) trigger an investigation.*

A.2.4 Equilibrium Outcome in Treatment *P-RI*

Note that the only difference between treatments *P-R* and *P-RI* is that in the latter, an investigation must be triggered if the employee is to obtain protection after sending a report. Since in an informative equilibrium we have $I = R$, reporting (no reporting) always results in (no) protection, as in treatment *P-R*. It follows that the respective equilibrium outcomes are the same in both treatments:

**Proposition 3** (*P-RI: Expected Equilibrium Outcome*). *In treatments P-RI and P-R, the expected equilibrium outcomes coincide.*

A.2.5 Equilibrium Outcome in Treatment *P-RIM*

In treatment *P-RIM*, the reporting decision is not directly observed by the employer, but since $I = R$ holds in an informative equilibrium, the employer can perfectly infer the reporting decision from observing whether or not an investigation occurs. We solve the game backwards starting with the employer’s equilibrium dismissal decision at date 4:

**Lemma 9** (*P-RIM: Dismissal*). *In the informative equilibrium, the following holds: The L-employee is always dismissed whenever this is feasible. The H-employee is dismissed if $I = 1$,
\( M = 0 \), and \( \tau \) sufficiently large. That is,

\[
D^*(x_\theta, I, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L \text{ holds and } R = I = M = 1 \text{ does not hold}, \\
1 & \text{if } x_\theta = x_H, M = 0, I = 1, \text{ and } \tau > \bar{\tau}, \\
0 & \text{else}.
\end{cases}
\]

where \( \bar{\tau} = x_H - \bar{x} \) as defined in Lemma 1.

Proof. In treatment \( P-RIM \), the employee is protected from dismissal when \( R = I = M = 1 \). Since \( I = R \) in the informative equilibrium, dismissal is, hence, feasible if either \( I = 0 \) holds (irrespective of \( M \)) or if both \( I = 1 \) and \( M = 0 \). If \( I = 0 \), the employer always dismisses the L-employee and always retains the H-employee (because \( x_L < \bar{x} < x_H \)). If \( I = 1 \) and \( M = 0 \), the L-employee is dismissed (because \( x_L - \tau < \bar{x} \)), while the H-employee is dismissed if \( x_H - \tau < \bar{x} \), i.e., for \( \tau > \bar{\tau} \).

In informative equilibrium, the employee’s optimal reporting behavior at date 2 can be characterized as follows:

**Lemma 10 (\( P-RIM \): Reporting).** In the informative equilibrium, both the L- and the H-employee send a report if and only if there is misbehavior. That is,

\[
R^*(x_\theta, M, \delta) = \begin{cases} 
1 & \text{if } M = 1, \text{ and} \\
0 & \text{else}.
\end{cases}
\]

Proof. The L-employee anticipates that he will be dismissed unless obtaining protection. For \( M = 1 \), his payoff upon choosing \( R = 1 \) is \( \omega \), while he would get only \(-\delta\) when choosing \( R = 0 \). Hence, he reports. For \( M = 0 \), the L-employee gets zero in any case, and thus no reporting is a best response. To summarize, for the L-employee we have \( R = M \) for all \( M \). Next, consider the H-employee: For \( M = 1 \), when choosing \( R = 1 \), the H-employee gets \( \omega \) and \( \omega - \delta \) otherwise. Hence, he reports. If \( M = 0 \), when choosing \( R = 1 \), he is retained with probability \( G(\tau) \) and hence gets \( G(\tau) \omega \). When choosing \( R = 0 \), he gets \( \omega \), which is strictly larger. To summarize, also for the H-employee, we have \( R = M \) for all \( M \).

Next, consider the employer’s misbehavior decision at date 1:

**Lemma 11 (\( P-RIM \): Misbehavior).** In the informative equilibrium, the employer’s misbe-
Behavior decision is given by:

\[ M^*(x_\theta, y, \tau) = \begin{cases} 
1 & \text{if } x_\theta = x_L \text{ and } y > f + \tau - x_L + \overline{x}, \\
1 & \text{if } x_\theta = x_H \text{ and } y > f + \tau, \text{ and} \\
0 & \text{else.} 
\end{cases} \]

**Proof.** Given Lemmas 9 and 10, the employer anticipates that both employee types will report if and only if \( M = 1 \), which then leads to protection. When matched with an L-employee, when choosing \( M = 1 \) the employer gets \( x_L - \omega + y - f - \tau \). When choosing \( M = 0 \), he gets \( \overline{x} - \omega \). Hence, \( M = 1 \) is preferred if \( y > f + \tau - x_L + \overline{x} \). When matched with an H-employee, when choosing \( M = 1 \) the employer get \( x_H - \omega + y - f - \tau \). When choosing \( M = 0 \), he gets \( x_H - \omega \). Hence, \( M = 1 \) is preferred if \( y > f + \tau \).

Finally, consider the investigation decision of the prosecutor. It follows from Lemma 10 that \( B_0 = 0 \) and \( B_1 = 1 \). Since \( K_1 < K_3 \), the condition \( B_0 \leq \frac{K_1}{K_3} < B_1 \) (see the proofs of Lemmas 4 and 8) is always satisfied.

**Lemma 12 (P-RIM: Investigation).** Given the behavior of the other players as described in Lemmas 9 - 11, choosing \( I^*(R) = R \) is optimal for the prosecutor.

Lemmas 9 - 12 characterize behavior in informative equilibrium. As this also depends on the random variables \( \tau, \delta \) and \( y \) (which are unobservable to the experimenter), we now state the expected equilibrium outcome given the prior distributions of these random variables. This expected equilibrium outcome is the basis for the predictions in Section 4:

**Proposition 4 (P-RIM).** The informative equilibrium in treatment P-RIM has the following expected equilibrium outcome: (i) Employees of either productivity type send a report if and only if there is misbehavior. (ii) L-employees are always dismissed whenever this is feasible. (iii) H-employees are never dismissed. (iv) The probability of observing misbehavior by the employer when matched with an L-employee is \( m_3^L := E_{y,\tau}[M^*(x_L, y, \tau)] = p_3^L \), where \( p_3^L = E_\tau[1 - H(f + \tau - x_L + \overline{x})] \). (v) The probability of observing misbehavior by the employer when matched with an H-employee is \( m_3^H := E_{y,\tau}[M^*(x_H, y, \tau)] = p_3^H \) as defined in (4). (vi) When (not) receiving a report, prosecutors always (never) trigger an investigation.

**A.2.6 Comparing Employer Misbehavior**

Propositions 1 - 4 directly lead to the predictions concerning investigations, dismissals, and reporting as presented in Section 4. The comparison of employer misbehavior across treatments
and employee productivity types (see Table 3) requires some further elaboration: From Lemmas 3, 7, and 11, for a given productivity type of the employee, the employer misbehaves if \( y \) exceeds a certain threshold. First, when the employer is matched with an L-employee, Lemmas 3 and 7 imply that both in treatment \( \text{NoP} \) and \( P-R \), the employer misbehaves if \( y > f \), while Lemma 11 implies that in treatment \( P-RIM \) the employer misbehaves if \( y > f + \tau - x_L + \overline{\alpha} \), where \( \tau - x_L + \overline{\alpha} > 0 \). Hence, \( m_{1L}^{no} = m_{1L}^1 > m_{1L}^3 \). Second, when the employer is matched with an H-employee, Lemmas 7 and 11 imply that both in treatment \( P-R \) and \( P-RIM \), the employer misbehaves if \( y > f + \tau \), and hence \( m_{1H}^{no} = m_{1H}^1 = m_{1H}^3 \). Moreover, the discussion of the threshold levels above immediately implies \( m_{1L}^1 > m_{1H}^1 = m_{1H}^3 > m_{1L}^3 \). It remains to show that \( m_{1L}^no > m_{1H}^1 \) holds.

From Lemma 7, the threshold for \( y \) that determines \( m_{1H}^1 \) is \( f + \tau \). From Lemma 3, the threshold for \( y \) that determines \( m_{1H}^{no} \) depends on \( \tau \): First, if \( \tau < \overline{\alpha} \), the threshold is \( (1 - F(\overline{\delta}))(f + \tau) < (f + \tau) \). Second, if \( \tau > \overline{\alpha} \), the threshold is \( (1 - F(\overline{\delta}))(x_H - \overline{\alpha} + f) = (1 - F(\overline{\delta}))(f + \overline{\alpha}) < (f + \tau) \) because \( \overline{\alpha} = x_H - \overline{\alpha} \) and we are in the case \( \tau > \overline{\alpha} \).
B Instructions

Note: We report here a translation of the instructions (originally in German) for treatments NoP and P-R, where all changes in P-R are indicated in square brackets as follow: [In P-R only: ...]. The respective modifications for the other treatments were made accordingly and are available upon request.

Welcome to today’s experiment!

You are taking part in a decision situation, where you can earn some money. How much you will earn depends on your decisions and on the decisions of the other participants that are allocated to you. Moreover, your earnings depend on the role that is randomly assigned to you. The experiment consists of two parts. You now receive the instructions for the first part. After having finished the first part, you will get the instructions for the second part. What happens in the first part of the experiment will not have any influence on the amount of money that you might earn in the second part of the experiment. And vice versa. After having completed both parts, you will also have to answer a short questionnaire.

Please note that from now on until the end of the experiment it is not allowed to communicate! If you have any questions, please raise your hand out of your cubicle. One of the experimenters will come to you. Throughout the experiment, it is forbidden to use mobile phones, smartphones, tablets, or alike. Participants intentionally violating the rules may be asked to leave the experiment and may not be paid. All decisions are made anonymously, i.e., none of the participants will learn about the identity of the others. The payment for both parts of the experiment will also be made anonymously at the end of the experiment.

Instructions for the first part of the experiment

Please notice that if subsequently we refer to the “experiment”, this relates to the first part of the experiment.

1. What it is about - A short overview
This experiment is about making decisions in a group of four people that consists of an employer, an employee, a third party, and a prosecutor, where these decisions may affect the payoffs of all members of the group. All decisions are made by the employer, the employee, and the prosecutor; the affected person cannot make any decisions. The employer chooses between two alternatives, CIRCLE and TRIANGLE. A (fictitious) law for the protection of the third party says that TRIANGLE should not be chosen as it harms the third party. Nevertheless, if an employer chooses TRIANGLE, he goes completely unpunished and even earns a higher profit - provided that the prosecutor does not initiate an investigation. The employer’s decision between the two alternatives can only be observed by the employee. The employee - and only him - can (but does not have to) ask the prosecutor to initiate an investigation. The prosecutor may initiate an investigation even if the employee has not asked him to do so. The employer learns whether an investigation is initiated or not. He also learns whether the employee asked the prosecutor to initiate an investigation or not. At the end of a given round (of which there will be several) the employer decides on whether the employee is dismissed or not. [In P-R only: If, however, the employee has asked the prosecutor to conduct an investigation, a dismissal of the employee is not possible. This applies regardless of whether the employer chose CIRCLE or TRIANGLE and regardless of whether the prosecutor initiated an investigation or not.] In the following, the experiment will be explained more in detail.

2. The assignment of roles

At the beginning of the experiment, the computer randomly assigns every participant a role either as employer, employee, third party or prosecutor. Employers will stay employers throughout the whole experiment. However, over the course of the experiment, prosecutors and employees will sometimes also take the role of third party; and third parties will sometimes take the role of either employee or prosecutor. Prosecutors will never take the role of employer, and employees will never take the role of prosecutor. The change of roles occurs randomly, and is consequently not affected by current or prior decisions. The change of roles only takes place between rounds. During a given round of the experiment, each member of the group remains in his or her role. In each round, the computer randomly matches the participants into groups of four consisting of an employer, an employee, a third party, and a prosecutor. The employee is also randomly assigned a productivity level (high or low).
Both productivity levels are equally likely, and the productivity level is independent across rounds, i.e., the productivity level of an employee might change from round to round. In the following, the course of events in a given round will be described. The experiment consists of 30 rounds.

3. The sequence of events in a given round

3.1. The sequence of events in a given round from the perspective of the employer

The employer does not receive an initial endowment; i.e., his earnings depend exclusively on his decisions and the decisions of the other group members. First, the employer learns whether the productivity level of his employee is high or low. A high-productivity employee, who does not get dismissed, will earn the employer 80 points for the current round; a low-productivity employee, who does not get dismissed, is worth 30 points. If the employer dismisses his employee at the end of the round [In P-R only: (which is only possible if the employee did not ask the prosecutor to conduct an investigation)], he will get a new employee whose productivity will earn him 70 points. Each employee who is not dismissed (and also any new employee replacing a dismissed employee) earns a wage of 40 points. An employee who got dismissed does not earn a wage in the current round.

Before the employer decides on whether to dismiss the employee or not, he has to take another decision: He has to choose between two alternatives, CIRCLE and TRIANGLE. This decision is observed by the employee only.

If CIRCLE is chosen

If the employer chooses CIRCLE, he will not receive any extra earnings, and he will not cause any financial loss for the third party. In this case, his earnings in the current round only result from the productivity of the employee (80, 30, or 70 points, depending on the productivity of the initial employee and depending on whether the initial employee is replaced by a new one) minus the employee’s salary (40 points).

- An employer with a high-productivity employee, who chooses CIRCLE, gets 80 - 40 = 40 points if he keeps the employee. If the employee gets replaced by a new one, the employer receives 70 - 40 = 30 points.
• An employer with a low-productivity employee who chooses option CIRCLE gets 
30 - 40 = -10 points if he keeps the employee. If the initial employee is replaced by a 
new one, the employer receives 70 - 40 = 30 points.

• These payments are irrespective of the prosecutor’s decision for conducting an 
investigation or not.

If TRIANGLE is chosen

If the employer chooses TRIANGLE, there are two [In P-R: four] distinct cases, depending 
on [In P-R only: whether the employee asked the prosecutor to investigate or not, and on] 
whether the prosecutor conducts an investigation or not.

In any of these cases if the employer chooses TRIANGLE, then he receives an extra payment 
of 50 points in addition to the productivity of his employee. In the case of no 
investigation, the employer goes unpunished and does not have to pay a fine, while in the 
case of an investigation, he has to pay a fine of 60 points, which, hence, exceeds the extra 
payment resulting from the choice of TRIANGLE. [In P-R only: Furthermore, the employee 
can only be dismissed if he did not ask the prosecutor to conduct an investigation, i.e., if he 
kept silent.]

• If the prosecutor does not conduct an investigation, and the employer consequently 
remains unpunished, the following holds:

  – An employer with a high-productivity employee who chooses TRIANGLE 
    gets 80 + 50 - 40 = 90 points if he keeps the employee. If the employee is replaced 
    by a new one [In P-R only: (which is only possible if the employee kept silent)], the 
    employer receives 70 + 50 - 40 = 80 points.

  – An employer with a low-productivity employee who chooses TRIANGLE 
    gets 30 + 50 - 40 = 40 points if he keeps the employee. If the old employee is 
    replaced by a new one [In P-R only: (which is only possible if the employee kept 
    silent)], the employer receives 70 + 50 - 40 = 80 points.

• If the prosecutor conducts an investigation, the following holds:
An employer with a high-productivity employee who chooses TRIANGLE gets $80 + 50 - 60 - 40 = 30$ points if he keeps the employee. If the employee gets replaced by a new one [In P-R only: (which is only possible if the employee kept silent)], the employer receives $70 + 50 - 60 - 40 = 20$ points.

An employer with a low-productivity employee who chooses TRIANGLE gets $30 + 50 - 60 - 40 = -20$ points if he keeps the employee. If the old employee is replaced by a new one [In P-R only: (which is only possible if the employee kept silent)], the employer receives $70 + 50 - 60 - 40 = 20$ points.

The potential fine is higher than the extra payment the employer receives when choosing TRIANGLE. Thus, it depends on the prosecutor’s decision to conduct an investigation or not whether the employer earns more when choosing TRIANGLE or when choosing CIRCLE.

However, the employer choosing TRIANGLE implies a loss of 70 points for the third party. As the third party has an initial endowment of 40 points, if the employer chooses TRIANGLE, the third party loses 30 points in the current round. However, this only applies if the prosecutor does not conduct an investigation, because choosing TRIANGLE violates the (fictitious) law for the protection of the third party. If the prosecutor conducts an investigation (potentially because he was asked to do so by the employee), the third party receives a partial refund of his damage in the form of a compensation of 20 points. In the role of third party, it is thus possible to complete the first part of the experiment with a loss. However, no participant will finish the entire experiment with a loss.

The total payoff (for the current round) of the employer (depending on the productivity of his employee as well as on his own decisions and the decision of the prosecutor) is summarized in the below table. In the experiment, this table is shown on the employer’s decision screen. [In treatment P-R, the part of the table marked by the red bold frame is displayed in addition to the remainder of the table.]

The employer should keep in mind that the employee observes his choice between the two alternatives and may ask the prosecutor to initiate an investigation. [In P-R only: In this case, a dismissal of the employee is not possible.]

3.2 The sequence of events in a given round from the perspective of the employee
The employee does **not receive an initial endowment**, i.e., his earnings depend exclusively on his decisions and the decisions of the others. First, the employee is informed about whether his **productivity level is high or low**. Both productivity levels are equally likely. At the end of the round, the employer can dismiss the employee. [In P-R only: However, a dismissal is only possible, if the employee did not ask the prosecutor to conduct an investigation, i.e., if he kept silent.] If the employee gets **dismissed**, he earns **0 points** in the current round. If the employee **does not get dismissed**, he receives a **wage of 40 points** from the employer.

The employee observes whether the employer chose **CIRCLE** or **TRIANGLE**. He then decides on whether to ask the prosecutor to conduct an investigation. This decision is taken as follows: The employee indicates both whether he wants to ask the prosecutor to conduct an investigation in case that the employer chose CIRCLE and also whether he wants to ask the prosecutor to conduct an investigation in case that the employer chose TRIANGLE. The computer then effectuates the decision (depending on the actual decision of the employer). Also the **employer** observes whether or not the employee decides to ask the prosecutor to conduct an investigation. If the **prosecutor conducts an investigation**, the following applies: If the employer chose CIRCLE, nothing happens. If, however, the employer chose TRIANGLE, the employer has to **pay a fine of 60 points**, while the **third party receives a compensation payment of 20 points**.

The total payoff (for the current round) of the **employee and the third party**, respectively, (depending on his own decision as well as on the decisions of the employer and the prosecutor)
are summarized in the below table. In the experiment, this table is shown on the employee’s decision screen. [In treatment P-R, the part of the table marked by the red bold frame is displayed in addition to the remainder of the table.]

<table>
<thead>
<tr>
<th>Employer chooses</th>
<th>Ask prosecutor to investigate</th>
<th>Keep silent</th>
<th>Third Party</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Investigation initiated?</td>
<td>Are you being dismissed?</td>
<td>Your payment</td>
</tr>
<tr>
<td>CIRCLE</td>
<td>No</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>CIRCLE</td>
<td>No</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>CIRCLE</td>
<td>Yes</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>CIRCLE</td>
<td>Yes</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>TRIANGLE</td>
<td>No</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>TRIANGLE</td>
<td>No</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>TRIANGLE</td>
<td>Yes</td>
<td>No</td>
<td>40</td>
</tr>
<tr>
<td>TRIANGLE</td>
<td>Yes</td>
<td>No</td>
<td>40</td>
</tr>
</tbody>
</table>

The employee should keep in mind two things. Firstly, if the employer chooses TRIANGLE, the employee may ask the prosecutor to conduct an investigation, and, if the prosecutor acts on his request, thereby reduce the loss of the affected person. Secondly, the employer can observe whether the employee asks the prosecutor to conduct an investigation or not.

3.3 The sequence of events in a given round from the perspective of the prosecutor

The prosecutor receives an initial endowment of 60 points at the beginning of each round. His task is to decide on whether to investigate the employer or not. If he conducts an investigation, he has costs of 20 points. If he does not conduct an investigation, the employer chose CIRCLE, the prosecutor keeps his initial endowments.

If the employer chose TRIANGLE, the prosecutor loses 20 points if he does not conduct an investigation. If he investigates (and in spite of the investigation cost of 20 points), he only has to bear a (smaller) loss of 10 points. When deciding on whether to investigate or not, the prosecutor can observe whether the employee asked him to investigate or not.

The total payoff (for the current round) of the prosecutor and the third party, respectively, (depending on his own decision and the decisions of the employer and employee) are summarized in the below table. In the experiment, this table is shown on the prosecutor’s decision screen.
The prosecutor should keep in mind two things: If the employer chose TRIANGLE, the prosecutor is the only one who can reduce both his own loss and the loss faced by the third party. If the employer chose CIRCLE, an investigation only leads to expenses. Thus, it is important for the prosecutor to think about how informative the employee’s request (or lack of a request) to conduct an investigation is.

3.4 The sequence of events in a given round from the perspective of the third party

The third party gets an initial endowment of 40 points and does not have any own decisions to make. If the employer chooses CIRCLE, the third party can keep its initial endowment, irrespective of what the employee and the prosecutor do. If the employer chooses TRIANGLE and the prosecutor does not conduct an investigation, the third party loses 70 points, so that its payoff in the current round is -30 points. If the employer chooses TRIANGLE and the prosecutor does conduct an investigation, the third party again loses 70 points. However, in this case the third party also receives a compensation payment of 20 points so that its earnings in the current round are -10 points. In the experiment, this table is shown on the third party’s decision screen.

4. Summary of the sequence of events in a given round

- Each participant learns his or her role.
• The employer and the employee learn the productivity level of the employee (high or low).

• The employer chooses between two alternatives: CIRCLE and TRIANGLE

• The employee decides whether he wants to ask the prosecutor to conduct an investigation in case that the employer chooses CIRCLE, and also whether he wants to ask the prosecutor to conduct an investigation in case that the employer chooses TRIANGLE.

• The prosecutor learns whether the employee asks him to conduct an investigation or not. The prosecutor then decides on whether to conduct an investigation or not.

• The employer learns whether the employee asked the prosecutor to conduct an investigation or not. The employer decides whether he dismisses the employee or not. [In P-R only: However, dismissal is only possible in case that the employee did not ask the prosecutor to conduct an investigation.]

• All participants learn their individual payoffs from the current round, and the decisions leading to these payoffs.

• Behavior in a given round does not affect earnings in upcoming rounds.

5. Total earnings for the first part of the experiment

At the end of both parts of the experiment, three rounds out of the total of 30 rounds will be selected randomly and independently from each other. The points that you have earned in these three rounds will be summed up and exchanged into EURO. The exchange rate is 1 EURO = 15 points. The resulting payoff plus the show-up fee of 12 EURO plus your earnings from the second part of the experiment will then constitute your overall payoff from the experiment.
C  Overview: Number of Observations

Table 6: Number of Observations Across Treatments and Conditions

(a) Number of Observations in Figure 2 (Dismissal)

<table>
<thead>
<tr>
<th></th>
<th>NoP</th>
<th>P-R</th>
<th>P-RI</th>
<th>P-RIM</th>
<th>RL</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-employee + Report</td>
<td>30</td>
<td>30</td>
<td>22</td>
<td>24</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>L-employee + No Report</td>
<td>30</td>
<td>26</td>
<td>20</td>
<td>24</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>H-employee + Report</td>
<td>29</td>
<td>30</td>
<td>22</td>
<td>23</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>H-employee + No Report</td>
<td>30</td>
<td>29</td>
<td>21</td>
<td>23</td>
<td>22</td>
<td>21</td>
</tr>
</tbody>
</table>

(b) Number of Observations in Figure 6 (Reporting)

<table>
<thead>
<tr>
<th></th>
<th>NoP</th>
<th>P-R</th>
<th>P-RI</th>
<th>P-RIM</th>
<th>RL</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-employee</td>
<td>45</td>
<td>45</td>
<td>33</td>
<td>36</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>H-employee</td>
<td>45</td>
<td>45</td>
<td>33</td>
<td>36</td>
<td>33</td>
<td>33</td>
</tr>
</tbody>
</table>

(c) Number of Observations in Figure 7(a) (Investigations)

<table>
<thead>
<tr>
<th></th>
<th>NoP</th>
<th>P-R</th>
<th>P-RI</th>
<th>P-RIM</th>
<th>RL</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Report</td>
<td>45</td>
<td>45</td>
<td>33</td>
<td>36</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>No Report</td>
<td>45</td>
<td>45</td>
<td>33</td>
<td>36</td>
<td>31</td>
<td>33</td>
</tr>
</tbody>
</table>

(d) Number of Observations in Figure 7(b) (Misbehavior)

<table>
<thead>
<tr>
<th></th>
<th>NoP</th>
<th>P-R</th>
<th>P-RI</th>
<th>P-RIM</th>
<th>RL</th>
<th>IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-employee</td>
<td>30</td>
<td>30</td>
<td>22</td>
<td>24</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>H-employee</td>
<td>30</td>
<td>30</td>
<td>22</td>
<td>24</td>
<td>22</td>
<td>22</td>
</tr>
</tbody>
</table>

Notes: As discussed in Section 5 above, in each session of the experiment, each subject played 30 periods in a given treatment, but possibly in different roles. Hence, our unit of observation are averages on the subject-level. Therefore, the number of observations in each treatment also depends on role assignments. As for panel (a), Figure 2 only exhibits treatments NoP and P-R. For the other treatments, obtaining protection no longer depends on the reporting decision only. As discussed in Section 5.3, the results remain strongly in line with Prediction D and hence are not reported in the main text. In panel (b), since we used the strategy method to elicit the employees’ reporting decision, the number of observations in Figure 6 does not vary across misbehavior decisions. The number of observations for Figures 3 and 4 are not reported here separately, as these two figures re-appear in Figures 6 and 7, and hence are already included in panels (b) to (d).