Motivating Whistleblowers*

Jeffrey V. Butler† Danila Serra‡ Giancarlo Spagnolo§

August 3, 2018

Abstract

Law-breaking activities within firms are widespread but difficult to uncover, making whistleblowing by employees desirable. We investigate if and how monetary incentives and expectations of social approval or disapproval from the public, and their interactions, affect an employee's decision to blow the whistle when the social damage from the reported misbehavior is more or less salient. Our analysis also has implications for the design and management of firms' internal whistleblowing channels.

JEL Codes: K42, C92, D04.

Key words: Whistleblowing, fraud, rewards, social judgment, experiment.

*We thank seminar participants at LUISS Guido Carli University, Southern Methodist University and the University of California, Merced as well as conference participants at the Bay Area Behavioral and Experimental Workshop, the Texas Experimental Symposium, the North-American ESA 2016 conference, the ASSA 2017 annual meeting and the First Bank of England/CEPR conference on Competition and Regulation in Banking for useful comments and suggestions. We are grateful to the Wallander and Hedelius foundation for financial support.

†University of California Merced, Department of Economics. Email: jbutler6@ucmerced.edu.

‡Southern Methodist University, Department of Economics. Email: dserra@smu.edu.

§Stockholm School of Economics, EIEF, and CEPR. Email: spagnologianca@gmail.com.

1 Introduction

Corporate fraud is widespread around the world. A recent survey of over 6000 organizations across 115 countries (2016 Global Crime Survey)[1] shows that one in three organizations, both worldwide and in the US, experienced fraud in the past 24 months, prevalently in the form of asset misappropriation, cybercrime, corruption, as well as procurement and accounting fraud. About 35% of the surveyed firms reported fraud-related losses exceeding $100,000, and 14% of firms reported losses above $1 million.[2] Dyck, Morse and Zingales (2013) estimated that between 1996 and 2004, about 15% of large[3] publicly traded US corporations engaged in fraud. The estimated expected annual cost of fraud for these firms amounts to a staggering $380 billion.

Due to their informational advantage, employees could potentially play a crucial role in uncovering illegal behavior and initiating internal or external investigations. However, while particular cases of whistleblowing have garnered the attention of the popular press in recent years, from the Enron scandal to the Snowden and Wikileaks-related cases, whistleblowing by employees is actually uncommon. Dyck, Morse and Zingales (2010) analyze 216 securities class action lawsuits filed against large US corporations and find that only about 18% of them were brought forward by an employee. Given the high costs associated with blowing the whistle – ranging from coworkers' disapproval and ostracism to lack of career advancement, job loss and outright harassment (e.g., Miceli and Near, 1994; Rothschild and Miethe, 1999) – this rarity is unsurprising.[4] Psychological costs caused by conflicting moral norms – loyalty toward the firm on the one hand, and fairness or justice concerns on the other – may also make employees reluctant to report wrongdoing taking place within their organization (Gundlach et al. 2003; Liu et al., 2018; Waytz, Dungan and Young, 2013). Fear of media scrutiny or public disapproval might further reduce employees' willingness to blow the whistle. Alternatively, if the expectation is of public approval, media or public scrutiny might actually increase whistleblowing, a possibility we discuss below.

In this paper, we experimentally investigate the effectiveness of different policies that might motivate individuals to report illegal activities taking place within an organization. We focus on both monetary and non-monetary incentives. In particular, we ask whether whistleblowers should be financially rewarded and whether they should be shielded from media scrutiny and social judgment. Moreover, we ask whether different sectors or different kinds of fraud require different policies, depending on whether the social costs generated by fraud are or are not visible and salient to the public – consider Medicare fraud versus insider trading – as suggested by recent legal theory (e.g., Engstrom, 2014b).

Monetary incentives for whistleblowing are the subject of an ongoing and contentious debate, intensified by the financial crisis of 2008. On the one hand, in 2010 the US enacted the Dodd-Frank Act that, among other things, allowed whistleblowers to receive financial bounties for bringing information to the Securities and Exchange Commission (SEC) or the Commodity Futures Trading Commission (CFTC).[5]

[1] https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf

[2] Taking into account that most cases of fraud go undetected and that firms self-selecting into a global crime survey are likely to be "cleaner" than those selecting out, the above numbers undoubtedly underestimate the current state of the corporate world.

[3] "Large" is defined by having assets exceeding $750 million.

[4] Many of these forms of retaliation – including, for example, lack of promotion – are sufficiently opaque to escape whistleblower protection laws, and the Ethics Resource Center (2014) reports a steady increase across time in the percentage of whistleblowers facing retaliation,even when whistleblowing is internal to the firm..

[5] The US is a pioneer in the enactment on laws and provisions that protect and reward whistleblowers. In 1986, the US strengthened provision of the False Claims Act (FCA), originally passed by Congress in 1863 and signed by President Abraham Lincoln to fight government fraud, allowing among other things for the qui tam, or whistleblower, provisions. It allows any individual or non-governmental organization to file an FCA lawsuit on behalf of the US Government and, if successful, to obtain up to 30% of recoveries plus fines. Another early whistleblower reward scheme targeting tax evasion is

On the other hand, across the Atlantic, regulatory agencies remain strongly opposed to financially rewarding whistleblowers,[6] even though US agencies consider them a great success[7] and the available empirical research (Dyck, Morse, and Zingales, 2010) suggests that they are indeed effective motivators of whistleblowing.[8]

The issue of protecting corporate whistleblowers from social judgment has not yet been examined by law-makers or the media, but we think it is an important one to address, given its potential impact on individuals' willingness to report illegal acts.[9] In fact, a vast theoretical and experimental literature has shown that individuals' behavior is highly responsive to the possibility of social observability and judgment (e.g., Andreoni and Bernheim, 2009; Andreoni and Petrie, 2004; Ariely et al., 2009; Benabou and Tirole, 2006; Carpenter and Myer, 2010; Gerber et al., 2008; Linardi and McConnell, 2011; Xiao and Houser, 2011), therefore suggesting that public scrutiny is likely to have a significant effect on whistleblowing. However, should we expect this effect to be positive or negative? The answer may depend on how whistleblowers expect to be judged by the public: will they be seen as snitches or as heroes? This may in turn hinge on how salient the social costs of manager malfeasance are to the public. For instance, in 1971 economist Daniel Ellsberg leaked the Pentagon papers concerning US involvement in Vietnam. He is widely viewed as a hero, which may be in part due to the salience of the (literal, physical) public harm associated with this controversial war. Public opinion is much more divided on Edward Snowden, who is seen by few as a hero and by many as a traitor. Perhaps not coincidentally, the public harm revealed by Snowden is more diffuse, distant and difficult to quantify.

An additional factor that may affect how whistleblowers are (or expect to be) perceived by the public is the presence of financial rewards. If whistleblowers get remunerated for their reporting, this may change (their expectations of) the public judgment of their actions, turning them from *heroes* to *greedy snitches*. In other words, financial rewards may crowd out non-monetary motivations driven by expectations of social approval (Benabou and Tirole, 2006). Therefore, the impact of financial rewards on whistleblowing may be lower, perhaps even turning negative, in the presence of social judgment (Gneezy and Rustichini, 2000b). Studying how financial rewards and expectations of social approval or disapproval interact in incentivizing (or discouraging) whistleblowing is therefore important and is one of the primary aims of our study.

In order to identify the impact of financial rewards and social judgment, and their interaction, on whistleblowing in a controlled setting where we can carefully measure individuals' willingness to report corporate wrongdoing, we employ a novel framed laboratory experiment that simulates the relationships between employees and managers within a firm. In our basic set-up, managers have the chance to engage

the one run by the IRS, which was substantially strengthened in 2006.

[6] In the UK, for example, the Bank of England's Prudential Supervision Authority and the Financial Conduct Authorities –gave a joint, strongly negative response in 2014 to a request for opinion from the UK parliament on financially rewarding whistleblowers, even arguing (incorrectly) that there was no empirical evidence of incentives leading to an increase in the number or quality of disclosures (see Nyrerod and Spagnolo 2018).

[7] The SEC reported in 2015 that they received 4000 tips from whistleblowers, an increase of 30% from 2012, with steady growth since 2011 probably resulting from increased awareness of the law. According to the IRS, their whistleblower program has helped to recover $3 billion since 2007, with $343 million recovered in 2013 and $310 million in 2014 (IRS, 2015).

[8] Dyck, Morse, and Zingales (2010) calculated that in sectors where the False Claim Act does not allow employees to obtain a financial reward, corporate fraud is unveiled by employees in 14% of the cases, while this percentage more than doubles (to 41%) when the False Claim Act can be applied, a highly significant difference. A series of articles published in top law journals (Engstrom, 2012, 2013, 2014a) also show empirically that several concerns about distortions linked to the False Claim Act are not justified in the light of the available data. Evidence on the (rather positive) effects of the whistleblower rewards linked to the Dodd-Frank Act is in Call et al. (2017), and Wilde (2017).

[9] There does not seem to be a consensus on whether the identity of whistleblowers should be safeguarded from the media and, more generally, the public. For instance, in the US, investigations conducted by the Security and Exchange Commission (SEC) protect the identity of whistleblowers, whereas investigations conducted under the False Claim Act expose whistleblowers by requiring them to file a court case.

in law-breaking behavior to benefit themselves and their employees at the expense of other subjects playing the role of members of the public. Employees, who are not victims but rather beneficiaries of the manager's illicit behavior, are given the option of blowing the whistle on their manager. Whistleblowing is costly for the employee and leads to the automatic imposition of a monetary penalty on the manager.

Across treatments, we manipulate the presence of both financial rewards for, and social judgment of, whistleblowers. In particular, in some treatments whistleblowing entails a net monetary cost to the employee, while in other treatments whistleblowing engenders a net financial gain. To test whether non-monetary motives such as aversion to social disapproval or desire for social approval play a role in whistleblowing, in some treatments potential whistleblowers are informed that participants assigned the role of member of the public are allowed to send costless judgmental messages – in the form of smiley or frowny faces – to employees who choose to blow the whistle. To induce variation in employees' expectations of positive or negative public judgment, we also manipulate across treatments whether members of the public are aware of the costs imposed on them by manager malfeasance. This variation also allows us to investigate whether financial rewards and social judgment, and their interaction, have a different impact on whistleblowing, and therefore are more or less desirable when applied to different kinds of fraud or different industries.

Our investigation also offers guidance for firms' internal whistleblowing policies, where top management is interested in finding out about possible misbehavior by lower ranked division managers.[10] In fact, our game could easily be reinterpreted as one where a division manager can illegally enhance his or her unit's performance while putting co-workers in other units (the public) at risk of legal action and reputation loss. In this setting, financial rewards would be wage raises or promotions, and social judgment could be implemented (prevented) by disclosing (protecting) the whistleblower's identity to (from) co-workers.

2 Literature Review

While there exist a number of theoretical economic analyses of whistleblowing (Spagnolo, 2004; Aubert et al., 2006; Friebel and Guriev, 2012; Felli and Hortala-Vallve, 2016; Heyes and Kapur, 2009; and Givati, 2016), empirical studies are rare and typically suffer from fundamental measurement and identification challenges, as only illegal behavior that has been uncovered and only whistles that have been blown can be observed. Consequently, existing studies focus on either the infringements that have been discovered (e.g., Dyck et al., 2010) or use scenario-based survey data (e.g., Feldman and Lobel, 2010). The management literature has employed models and surveys to identify the personality and situational variables predictive of whistle-blowing (e.g. Dozier and Miceli, 1985; Near and Miceli, 1995; Gundlach et a., 2003; Miceli et al., 2012). For instance, recent work by Lui et al. (2018) highlights the importance of employees' identification with the organization – together with the ethical culture in the organization and personality traits – as predictive of whistleblowing.[11]

[10] The Sarbanes-Oxley Act and the new proposed EU Directive on Whistleblowers require firms to establish policies to elicit employees' wrongdoing through internal whistleblowing, which may allow managers to uncover and correct employees' malpractices and reduce the potentially large legal and reputation costs of having malpractices uncovered by regulators. Of course this is a delicate point, as when the malpractice is induced by top management, internal whistleblowing policies may be misused for 'cover-ups' and reduce firms' cost of misbehaving (see e.g. Felli and Hortala-Vallve, 2016).

[11] The management literature recognized early on that whistleblowing may enable organizational leaders to correct practices that may harm the organization and is therefore desirable. However, to our knowledge, management and organizational studies have not considered the importance of pecuniary factors, nor that of social approval or of the visibility of the negative externalities caused by fraud, in incentivizing (or disincentivizing) employees' willingness to blow the whistle against

The measurement and identification issues that make empirical investigations of whistleblowing problematic have led to a recent surge of experimental studies on the factors that may induce employees to blow the whistle against malfeasance. Laboratory experiments are particularly valuable, as they allow researchers to directly observe both wrongdoing and whistleblowing, and to measure responsiveness to changes in incentives in a controlled environment.

One of the first whistleblower experiments is by Reuben and Stephenson (2013), who examine individuals' willingness to report team members after observing them cheat while knowing that blowing the whistle would cause the whole group to be penalized. More recently, Carpenter et al. (2017) experimentally investigate peer reporting within a firm and find that sharing profits with employees may effectively incentivize individuals to blow the whistle against shirking co-workers.

Bartuli et al. (2016) study whistleblowing in an experimental context that is closer to ours, i.e. a setting where: i) the potential whistleblower is an employee that benefits from the wrongdoing of the manager; ii) such wrongdoing generates losses to a third party; and iii) blowing the whistle is costly. However, while we are interested in testing policies aimed at incentivizing whistleblowing, Bartuli et al. (2016) aim to identify personality traits that are more likely to lead to whistleblowing.[12] Similarly, Waytz et al. (2013) use survey questions to investigate the relationship between propensity to blow the whistle and a specific individual trait: the subjective valuation of fairness/justice over loyalty.

The experimental study most closely related to ours is by Schmolke and Utikal (2016), who investigate whistleblowing in a neutrally framed environment where one subject may increase his payoff at the cost of increasing inequality among other players who can then report this behavior to a third subject, the potential whistleblower. Blowing the whistle leads to punishment and redistribution of payoffs to restore initial conditions. The authors study the effects of rewards for, versus fines for not, blowing the whistle and find that even modest monetary rewards increase the probability of whistleblowing. While the experiment has other interesting treatments,[13] it does not investigate the role that expectations of social approval or disapproval may play in the whistleblowing decision, and how they may interact with financial incentives.

More tangentially related to our study is the well-developed literature on whistleblowing in the context of illegal cartel formation among firms. Apesteguia, Dufwenberg and Selten (2007) were the first to study leniency and rewards to whistleblowers in an experiment on illegal cartel formation in the context of static Bertrand competition. Their results suggest that rewarding whistleblowers increases the likelihood of whistleblowing without reducing market prices. In a repeated game version of an analogous experiment, Bigoni et al. (2012) find that offering a monetary reward to the first whistleblower leads to high reporting rates that strongly deter cartel formation as predicted by theory (Spagnolo 2004, 2008). A number of other experimental studies focus on the effectiveness of leniency policies providing amnesty or asymmetric legal treatment to accomplice-witnesses that blow the whistle against collusion without the use of monetary rewards, including Hamaguchi et al. (2009), Hinloopen and Soetevent (2008), Bigoni et al. (2015), and Cotten and Santore (2016).

Somewhat less directly related to our study is also another growing strand of experimental literature that investigates whistleblowing in the context of corrupt transactions between public officials and

observed malfeasance.

[12] They find that employees who are more altruistic and more concerned about ethical issues are more likely to blow the whistle. For survey-based studies of personality and whistleblowing, see also Miceli and Near (1992, 1994) and Feldman and Lobel (2010).

[13] They manipulate whether and how the reporting subject and the enforcing authority are positively or negatively affected by the first subject's decision.

citizens/firms. For instance, Abbink and Wu (2017) simulate both one-shot and repeated transactions between firms and public officials where firms can obtain illegal services through the payment of a bribe. They find that whistleblower amnesty and monetary rewards strongly deter illegal transactions in a one-shot setting, but that deterrence is limited in repeated relationships. Abbink et al. (2014), Buckenmaieret et al. (2017), Schikora (2011) and Serra (2012) find similar results with amnesty alone.[14]

In sum, the existing experimental literature – whether it simulates a firm environment, illegal cartel formation or corrupt transactions – has mainly focused on the effect of financial rewards and/or amnesty on the propensity to report wrongdoing, or on the deterrence effects of whistleblowing on wrongdoing. While we also investigate the effect of financial rewards on whistleblowing, our main contribution to the literature is threefold. First and foremost, we examine how non-monetary motivations in the form of expectations of public approval or disapproval affect the propensity to blow the whistle against somebody that is in a position of power and whose law-breaking benefited the potential whistleblower. This is a largely unexplored question. In fact, while there is a growing literature on how social observability and judgment affect behavior (e.g., Andreoni and Bernheim, 2009; Andreoni and Petrie, 2004; Ariely et al., 2009; Benabou and Tirole, 2006; Gerber et al., 2008; Linardi and McConnell, 2011; Salmon and Serra, 2017; Xiao and Houser, 2011; see also the overview provided by Bursztyn and Jensen, 2017), to the best of our knowledge there are no studies investigating the relationship between whistleblowing and public judgment. This is an important relationship, as the results of our analysis have the potential to inform policy about whether and in what contexts protecting whistleblowers from public scrutiny is desirable. Second, we ask whether different kinds of wrongdoing, possibly taking place in different industries, require different kinds of policies. In particular, we differentiate between cases of fraud generating negative externalities to society that are easily visible to the public and cases of fraud involving social costs that are less transparent or salient to the public, and consider whether the effects of financial and non-financial incentives differ across these contexts. Finally, our study sheds light on whether financial rewards may be less effective if whistleblowers are exposed to public/media scrutiny, i.e., whether they may induce the public to view whistleblowers more as *snitches* than as *heroes*.

3 The Experiment

3.1 Design

The experimental session consists of six stages, as shown in Figure 1. At the beginning of the experiment (Stage 0), participants are randomly assigned either the role of "member of a firm" or the role of "member of the public." Each firm is made of three subjects, and while multiple firms participate in each session, firms operate independently from each other. In other words, there is no interaction between firms and the payoffs of each firm member are determined solely by the actions that take place within their firm. There are 6 participants playing as members of the public, i.e., double the number of the members of any given firm. This is to recreate in the lab the standard case where the "society" that may be negatively affected by corporate fraud is larger than the firm engaging in it.

Following the role assignment stage, the experiment begins and it comprises of 4 active stages (Stages 1 to 4 in Figure 1), only one of which is randomly chosen for payment at the end of the experimental

[14] Breuer (2013) studies the effects of financial rewards for whistleblowers in a laboratory experiment on tax evasion and finds a strong positive effect of rewards on subjects' willingness to blow the whistle on the tax declaration of another subject and little evidence of crowding out of non-monetary motivations.

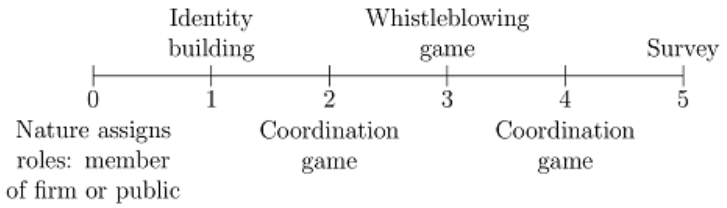

Figure 1: The stages of the experimental session

session. In Stage 5, subjects participate in a brief post-experiment survey.

Since loyalty to the firm and to one's manager is an important feature of work within organizations and a potential obstacle to the decision to report wrongdoing (Waytz et al., 2013), Stage One was designed to induce a sense of identity and social cohesion among each firm's members. In this stage, the three members of each firm engage in a series of team-building tasks with interdependent payoffs to create a sense of "shared fate," a feature which has been shown to induce a common identity (Ashforth and Mael, 1989). The first task is the Kandinsky and Klee painting elicitation module first developed in Tajfel et al. (1971), in which subjects view a series of paintings and guess whether each of them is a Klee or a Kandinsky. Each individual gets credit if at least one member of the firm guesses correctly. The second task consists of a series of addition problems. As before, each member of the firm earns money for each problem that at least one member of the firm solves correctly. The third task involves a series of multiplication problems, each of which involves multiplying two two-digit numbers. Individual payoffs are determined as in the previous team-building tasks. The members of the public engage in the same three tasks but their payoffs are determined exclusively by their own performance. At the end of each task, firm members are informed of their own performance and the overall firm performance, which generates their earnings. Members of the public are informed only of their own performance.

Stage Two consists of a one-shot minimum-effort coordination game aimed at testing whether Stage One resulted in the desired within-firm cohesion. Each member of a firm plays the game with the other two members, while each member of the public plays the game with two other members of the public. Participants choose a level of effort between 110 and 170, with their payoffs being determined by the difference between the minimum effort chosen in the group and their own effort multiplied by 0.75.[15] Subjects are not informed of the outcome of this game and the resulting earnings until the end of the experimental session. If the identity-building task conducted in Stage 1 was successful, we would expect the minimum effort chosen by firm members to be higher than the minimum effort chosen by members of the public (since the latter did not engage in the identity-building task among themselves).

In Stage Three, participants play the Whistleblowing Game. Subjects retain the role of either member of the firm or member of the public. Within each firm, one participant is randomly chosen to be the "manager" and the remaining two participants are assigned the role of "employees."[16] By having

[15] We chose the minimum effort game as we wanted a coordination game with multiple Pareto-ranked pure-strategy Nash equilibria. Chen and Chen (2011) have recently shown that, in the context of the minimum effort game, induced group identity increases the effort levels chosen by the group members, facilitating coordination on the efficient high-effort equilibrium.

[16] We chose the role allocation to be done randomly rather than based on individual performance in stage one as we

two employees of identical status and a manager, we aim to simulate most organizational set-ups where multiple individuals have the same tasks and respond to the same high-ranked supervisor or manager.[17]

The employees engage in a real-effort task consisting of adding two-digit numbers, as in task two of Stage One of the experiment. Each correct answer generates private earnings at a piece rate of 2 ECU and also contributes to a firm fund at a piece rate of 1 ECU.[18] There are a total of 12 problems per employee, resulting in maximum private earnings of 24 ECU per employee and a maximum firm fund of, also, 24 ECU. The firm fund is later distributed back to the manager (one half of the fund) and the employees (one fourth each).

The manager gets a fixed wage of 24 ECU and has the chance to double the firm's fund by engaging in a more difficult real-effort task (multiplying two-digit numbers, as in task three of Stage One of the experiment) and answering at least eight of the 12 problems correctly. Alternatively, the manager can augment the fund by "breaking the law." The manager is informed that breaking the law generates money for the firm but causes a monetary loss of 2 ECU to each of the six members of the public. Our payoff configuration implies that, as in real organizations, the manager always makes more money than the employees,[19] and his or her performance, whether through legal or illegal practices, may add significant value to the firm and therefore benefit the employees.

As before, members of the public are only involved in individual decision-making. They have an initial endowment of 14 ECU and, like the employees, engage in a real-effort task consisting of adding two-digit numbers. The task generates 2 ECU for each correct answer. However, their final earnings also depend on the rule-breaking choice of the managers of the firms in the session, since each manager's wrongdoing causes a loss of 2 ECU to each member of the public. This implies that the total loss suffered by each member of the public ranges from a minimum of 0 (if all managers in the session decide not to break the law) to a maximum of (2xN) if all the managers of the N firms in the session – with N ranging from 2 to 6, depending on session size – decide to break the law.

Note that the decision to break the law would be socially efficient only if the manager were not able to augment the fund by successfully completing the multiplication task and if the firm fund were larger than 12 ECU. If the employees jointly generate a firm fund of 12 ECU, law-breaking behavior by a low-ability manager would generate a firm surplus of 12 ECU while also generating a societal loss of 12 ECU. A high-ability manager's decision to break the law is always socially inefficient. This is because a high-ability manager would always be able to solve the multiplication task correctly, thus generating the firm surplus without any negative externalities on society. In order to keep the manager's decision to break the law comparable across firms and independent of efficiency concerns, we do not reveal the size of the firm fund to the manager before eliciting his or her decision to break the law.

We measure employees' willingness to blow the whistle by using the strategy elicitation method. We ask each employee within a firm whether they would blow the whistle if they found out that their manager broke the law. Blowing the whistle requires the employee to pay a monetary cost of 5 ECU and imposes a monetary penalty of 14 ECU on a law-breaking manager. Whistleblowing confers no direct benefit to members of the public. In particular, it cannot offset the monetary harm imposed on them

wanted to have enough variation in CEOs' decisions to break the law. Since such a decision is likely to be correlated with CEO's ability, role allocation by merit would have likely resulted in low frequency of law breaking.

[17] We also aimed to reduce each employee's competitive feelings and inequality aversion toward the manager.

[18] Experimental currency units (ECU) were exchanged for dollars at the end of the experiment at the rate of 2 ECU per $1, as described below.

[19] Note that the manager's wage equals the private earnings of the employee if the employee is highly productive, i.e. he or she solves all 12 problems correctly. Even in this case, the manager ends up with higher earnings, since he or she receives half of the firm fund versus the one-fourth received by the employee.

by manager malfeasance. Our primary rationale for this design choice is to mirror real-world situations where the public is not fully compensated for the harm caused by firm misbehavior, either because of transactions costs of prosecution, such as delays in adjudication or legal fees, or because it is difficult to perfectly assess precisely who the victims are.

Our choice of the strategy method balanced several considerations. Using the strategy method allows us to observe employees' willingness to blow the whistle even in situations where rule-breaking does not actually occur. This confers two advantages. First, we avoid selection issues which complicate empirical analyses of data when observations are missing non-randomly. A second advantage is sample size. Essentially, had reporting been directly elicited, each employee's whistleblowing decision would have been conditional on the actual occurrence of law-breaking, compromising comparability across employees and resulting in fewer data points. The primary disadvantage of the strategy method is external validity. In real life situations, the decision to blow the whistle may often be "hot" rather "cold" particularly when it is made immediately after the observation of manager wrongdoing, as it is in our experiment. Since the strategy method requires decisions to be made before law-breaking has actually occurred, it may not capture visceral which can affect "hot" decisions in the real world. While we took this disadvantage seriously, our concerns were partially allayed by a recent study examining dozens of studies to compare results using the direct-response method with results using the strategy method which concluded "... in no case do we find that a treatment effect found with the strategy method is not observed with the direct-response method" (Brandts and Charness, 2011). An earlier study by the same authors found no difference in positive (rewarding) or negative (punishing) responses to others' behavior across the direct-response and strategy methods (Brandts and Charness, 2000).

We compute final earnings within a firm by randomly choosing one of the two employees in the firm and implementing the stated whistleblowing decision conditional on the matched manager's behavior. With this design choice, we purposely abstract from the potential presence of collective action problems in the decision to blow the whistle and from the need to control for subject behavior and expectations in such a strategic situation. These aspects have been analyzed in other contexts (see, e.g., Bigoni et al. 2012, 2015) and would have increased complexity and noise in the measuring of the effects we are interested in here.

Stage Four concludes the remunerated portion of the experiment with a minimum-effort coordination game identical to the game subjects played in stage two. We included this stage with the purpose of possibly identifying the effects of the decisions made in the whistleblowing game – i.e., the manager's law-breaking decision and the employees' reporting decisions – on firm cohesion.

In Stage Five, after participating in the experiment, subjects fill out a survey. As part of the survey, all subjects are presented with four actual whistleblowing cases that differ both in the extent to which the negative externalities caused by the illegal behavior are visible to the public and in the presence of financial rewards for whistleblowers. The four cases are the Snowden case, the Enron case, the UBS case and the Tenet case.[20] We chose these cases because the visibility of negative externalities varies substantially across the cases, as do the financial incentives for the whistleblowers involved.

[20] For information on the Snowden case, see: https://www.nytimes.com/2014/01/02/opinion/edward-snowden-whistle-blower.html?_r=0. For the Enron case, see: http://news.bbc.co.uk/2/hi/business/5335214.stm. For the UBS case, see: http://www.wsj.com/articles/SB10000872396390444017504577645412614237708. For information on the Tenet case, see: http://www.corporatecrimereporter.com/news/200/tenet-healthcare-to-pay-514-million-gets-non-prosecution-agreement-two-units-with-no-assets-to-plead-guilty.

3.2 Treatments

We employ three treatment variations by manipulating the presence of financial rewards for whistleblowers, whether whistleblowers are exposed to social judgment, and whether the members of the public are aware of the negative externalities that the manager's illegal actions generate on them.

1. *Reward* vs. *No Reward*: In the No Reward condition, whistleblowing employees bear a cost of 5 ECU, while in the Reward condition an employee that blows the whistle against his or her manager also receives a financial reward of 10 ECU (i.e., whistleblowing results in a net financial gain of 5 ECU for the employee). All participants in the game. i.e., managers, employees and members of the public, are made aware of the cost associated with whistleblowing as well as the financial reward (in the Reward treatment).

2. *Social Judgment* vs. *No Social Judgment*: Under Social Judgment, members of the public are given the chance to send messages of approval or disapproval to whistleblowers. Similarly to Carpenter and Seki, 2011) and Salmon and Serra (2017), these messages take the form of a smiley face, a frowny face or a neutral face. Each member of the public can also choose to send no message at all to whistleblowers.[21] Sending a message comes at no cost to the member of the public and does not lead to any monetary reward or penalty for the whistleblower. Crucially, employees are also informed, before they make their reporting decision, that each member of the public will be able to send one of these messages to an employee who chooses to blow the whistle. By contrast, in the No Social Judgment treatment, the public is informed of whistleblowing but cannot send messages of any kind to the whistleblower.[22]

3. *Visible* vs. *Invisible Externalities*: Under Visible Externalities, all experimental participants (i.e., managers, employees and members of the public) are told that the members of the public will be made aware of the monetary losses they suffer (or could suffer) due to each manager's illegal actions. In other words, all subjects are informed about the exact payoff configuration resulting from the game, i.e., the members of the public know that, in addition of their initial endowment, they earn 2 ECU for each problem they solve correctly, and they lose 2 ECU for each manager that engages in law-breaking. In contrast, under Invisible Externalities the members of the public are informed that managers of firms can engage in wrongdoing, and they are told whether they did or did not at the end of the session, but they do not know that such wrongdoing affects their own earnings negatively. We achieve this by not disclosing to members of the public exactly how much they could earn from each correctly solved problem while they engage in the task. We tell them that they will earn money for the task and will be informed how much they made at the end. Managers and employees are aware that under Invisible Externalities the members of the public do not know about the monetary losses that they may suffer due to managers' law-breaking behavior.

[21] Carpenter nd Seki (2011) were the first to use messages showing unhappy faces to signal social disapproval. Salmon and Serra (2017) expanded on this methodology by allowing participants to signal either approval or disapproval through messages displaying happy, unhappy or neutral faces.

[22] Note that we are not allowing the co-worker or the manager to send messages of approval or disapproval to the whistleblower. We omitted this possibility not because we believe the judgement from direct co-workers to be unimportant or uninteresting (although this would partly the case for the judgement from the reported manager), but to avoid an additional complication and potential source of variation in beliefs that might have confused the answer to the core questions of our paper, i.e., the effects of social judgement from the more distant public, and its interaction with visibility of the externality and financial incentives.

	Invisible Externalities		Visible Externalities		Total	
Treatments	Sessions	Subjects	Sessions	Subjects	Sessions	Subjects
No Rewards & No Judgment	4	63	3	51	7	69
No Rewards & Social Judgment	3	51	3	60	6	75
Rewards & No Judgment	4	75	3	54	7	96
Rewards & Social Judgment	3	48	4	69	7	84
Total	14	234	13	237	27	471

Table 1: Summary of experimental sessions and treatments.

The interactions between our three treatment manipulations generate eight experimental conditions, as shown in Table 1.

3.3 Implementation

We conducted 27 sessions involving 471 participants at the University of California, Santa Barbara's Experimental and Behavioral Economics laboratory (EBEL), as shown in Table 1. Each subject participated in only one session and one treatment. In each session, 6 subjects were randomly assigned the role of members of the public (MPs) and between 6 and 18 subjects were randomly assigned the role of members of a firm, for a total of between 3 and 6 firms per session. Members of each firm made decisions independently from all the other firms participating in a session.

In referring to subject roles, the experimental environment and available actions, we used the same contextual labels we used in Section 3.1 when describing the game. We chose to implement a framed experiment because, as recently discussed in Alekseev, Charness and Gneezy (2016), psychological and social factors may play a significant role in individuals' decisions to engage in and report on unlawful behavior and, in such situations, framing may help subjects more fully understand the decision-making context.[23]

The experiment consisted of an initial role-assignment stage, followed by four active stages plus a survey. Subjects were presented with the instructions for each stage on their computer screen immediately before that stage began. Only one randomly selected stage of the experiment was used for actual payments. Experimental earnings were converted from ECUs to dollars at the exchange rate of \$1 for 2 ECU. The experiment was programmed in z-Tree (Fischbacher, 2007) and subjects were recruited among pre-registered UCSB students using ORSEE (Greiner, 2015). In order to guarantee anonymity, at the beginning of each session subjects were randomly assigned an identification number, which they kept for the duration of the experiment. At no point during the experiment did we ask subjects to reveal their names and, although actual names were used during the payment process for accounting purposes, we informed subjects that we would not register their names and therefore would not be able to link them to the choices made in the experiment. Each session lasted between 60 and 90 minutes, with average earnings of \$29 per subject (including a \$10 show-up fee).

[23] Framing effects have been found in a large set of pro-social games, including public goods games (Andreoni, 1995; Cookson 2000; Rege and Telle 2004; among the others) and dictator games (Eckel and Grossman, 1996; Brañas-Garza, 2007). For a recent study of how frames significantly affect first- and second-order beliefs, see Dufwenberg, Gächter, and Hennig-Schmidt (2011). Alekseev, Charness and Gneezy (2016) provide a recent review of experiments employing either abstract or meaningful frames to present the decision-making setting to the experimental subjects. Their general finding is that "evocative language either does not affect behavior or affects it in a desirable way by evoking the desired emotional response."

3.4 Hypotheses

A wide array of motivations may influence individuals' preferences and decisions. Our reading of the existing whistleblowing literature suggests three which are likely to be particularly important: monetary incentives, personal moral concerns and a preference for social approval. By varying the presence of financial whistlblower rewards, the possibility of social judgement and whether the negative externalities imposed by misbehaving managers are visible to the public, our experimental treatments were designed to manipulate these three motivations in transparent ways. To provide a point of departure in thinking about how these manipulations may affect whistleblowing, in Section A of the Appendix we construct a simple framework explicitly incorporating monetary incentives, moral concerns and preferences for social approval into whistleblowers' utility. By making assumptions about how our treatments affect these three motivations, we use the framework to illustrate how our treatments may affect employee whistleblowing. Our framework allows us to formulate three broad hypotheses.

The simplest case assumes that each of the motivations – financial, moral and social – are independent of one another. Note that this implies financial incentives cannot "crowd out" non-financial incentives in the sense of directly altering the moral or social utility consequences of whistleblowing. Absent such crowding out, our first prediction is straightforward:

Hypothesis 1 *Financial rewards will increase the likelihood that an employee will blow the whistle.*

Whether this hypothesis will be supported or not in the data is not *a priori* obvious. As discussed in the introduction, there is widespread concern, partly related to previous experimental work, that financial rewards might crowd out intrinsic pro-social incentives to blow the whistle, those based on the expectations of social judgement, or both. In principle, we might therefore observe a decrease in the frequency of whistleblowing following the introduction of financial incentives. We return to this possibility below.

As mentioned, the other main focus of our analysis is the role of social judgement. With respect to social judgment, we assume that the public is more likely to perceive whistleblowing as a pro-social act when it is aware of the harm associated with manager misbehavior. Intuitively, when members of the public are aware that they are being harmed by the firm, they are more likely to want the manager to be punished and, consequently, to socially reward the whistleblower for triggering such punishment. If, instead, the public does not feel directly affected by the manager's wrongdoing, it is possible that it will perceive the whistleblower as somebody who decided to run afoul of the widespread moral norm of group loyalty[24] and commit an anti-social act, leading to social disapproval. In other words, the visibility of the negative externalities to the public is likely to affect whistleblowers' beliefs about how they will be perceived and judged by the public if they do blow the whistle, i.e., as heroes if the externalities are visible and as snitches if they are not visible. These assumptions lead to our second hypothesis.

Hypothesis 2 *Allowing for social judgment will decrease whistleblowing in our Invisible Externalities treatments relative to our Visible Externalities treatments.*

[24] In our discussion, we are abstracting from the concerns that individuals may have about the social judgment that they would receive from their fellow firm members. A plausible assumption is that employees prefer to appear loyal to fellow firm members while also wanting to appear pro-social to members of the public, especially if they are subject to public judgment. When the negative externalities caused by fraud are visible to the public, loyalty toward firm members and preferences for social approval from members of the public pull employees in different directions. When the negative externalities are invisible to the public, both motivations steer employees away from blowing the whistle.

Until now, we have been putting aside one of the key questions of our inquiry: whether and how financial incentives and social judgement interact in affecting the decision to blow the whistle. Whether or not a whistleblower receives a financial reward may obviously affect the way the public perceives the whistleblower, or how the whistleblower expects to be perceived and judged by the public, which is what ultimately matters for eliciting whistleblowing. The fact of being paid for blowing the whistle may, for example, reduce the perceived ethical value of the act and induce the public to see the whistleblower more as a snitch than as a hero. We turn now to this question.

For our next hypothesis we focus on our Visible Externalities treatments where we have just argued that social judgment should increase whistleblowing. Now, however, we relax the assumption of motivation independence. This permits a simple form of crowding out. In particular, if potential whistleblowers believe that monetary rewards will directly negatively impact the public's opinion of whistleblowers, then the whistleblower will expect less social approval when there are rewards versus the case with no rewards. This utility offset will only be a factor in the case where the public is allowed to voice their approval or disapproval, i.e., where social judgment is possible. These observations lead us to the following general hypothesis.

Hypothesis 3 *In our Visible Externalities treatments, social judgment will be less effective at inducing additional whistleblowing with monetary rewards than without rewards.*

A potential consequence of crowding out is that financial incentives may be less effective, or even counterproductive, at eliciting additional whistleblowing when whistleblowers are subject to social judgment than if crowding out were not present or if social judgment were not possible.

Finally, we note that our data might allow us to address how different environments ultimately affect manager malfeasance by altering the likelihood that illegal activity is detected and punished. We do not have the richness of variation in incentives for managers that we have for employees – in particular, the manager is never directly exposed to social judgment. This makes our hypothesis with regard to manager behavior straightforward.

Hypothesis 4 *Managers will be less likely to break the law in treatments where whistleblowing is more likely.*

4 Results

We start by assessing the extent to which we were able to create social ties between members of the same firm in the Stage One tasks that preceded the whistleblowing game. As a measure of the resulting within-firm cohesion, we use the minimum effort chosen by members of a firm in the coordination game in Stage Two that followed our team-building tasks. A comparison of the average minimum effort chosen by members of a firm and the average minimum effort chosen by members of the public, who did not engage in team-building tasks,[25] provides strong evidence of induced firm cohesion. The minimum effort chosen by members of firms is significantly higher than the minimum effort chosen by members of the public (123.70 vs. 119.26; two-sided t-test $p - value$ of 0.001).[26] This finding suggests that we were

[25] As explained in Section 3.1, during the team-building stage of the experiment (Stage One) members of the public engaged in the same tasks as the members of a firm, but their payoffs were determined solely by their individual performance in these tasks.

[26] In the game, each member of a three-person group had to choose an effort level in the [110, 170] range, with payoffs being determined by: [minimum effort in the group – 0.75*(own effort)].

successful in generating social cohesion and, possibly, in-group loyalty among members of a firm.[27]

In what follows, we present and discuss the core results of the paper: the effects of our treatments on employees' willingness to blow the whistle against their manager (Section 4.1). We then present our findings on managers' law-breaking behavior across treatments (Section 4.2). We conclude by describing the members of the public's approval or disapproval of whistleblowers in the Social Judgment treatments (Section 4.3).

4.1 The decision to blow the whistle

Overall, about 33% of employees decided to blow the whistle against their law-breaking managers. There is considerable variation across treatments, with the percentage of whistleblowers ranging from 10% to 59%, as shown in Figure 2 and Table 2. Since the Visible Externalities and the Invisible Externalities treatments simulate different types of illegal actions or different industries where the damages generated by fraud to the public are either more or less difficult to identify, we present the results obtained under the two settings separately. A number of important results emerge from Figure 2 and Table 2. First, the presence of financial rewards seems to generally and substantially increase the prevalence of whistleblowing. This holds both when whistleblowers are subject to social judgment and when they are not. The sole exception, to which we return towards the end of this section, is that financial rewards are ineffective when the externalities caused by fraud are visible to the public and whistleblowers are shielded from social approval or disapproval.

Second, whistleblowing is prevalent even when financially costly and it varies substantially with contextual variables having no direct earnings consequences. This implies that we are justified in our desire to take into account non-pecuniary motivations when setting policies with regard to whistleblowing. The observed behavior is mostly consistent with the idea that individuals care about social approval. Indeed, the data shows that the possibility of social judgment has a different effect on whistleblowing depending on the visibility to the public of the costs imposed on them by manager malfeasance. In particular, when the negative externalities caused by fraud are not visible to the public, the possibility of social judgment, through expectations of social disapproval, tends to decrease employees' willingness to blow the whistle, whereas when negative externalities are visible to the public, social judgment generally increases whistleblowing, possibly because whistleblowers expect to receive messages of social approval. This is discernible from Figure 2 and Table 2 by considering all pairwise comparisons of the forms (–, –, No Judgment) and (–, –, Social Judgment). The lone exception occurs when negative externalities are visible to the public but there are no whistleblower rewards.

Moving beyond simple pairwise comparisons, in Table 3 we report estimates from a linear probability model where the dependent variable is a dummy equal to 1 if the employee is willing to blow the whistle and 0 otherwise. Table 5 in the Appendix reports estimates generated by probit regressions. In the first two columns, we split our data by the visibility of negative externalities for clarity, as behavior was substantially different across this dimension. In column 3, we pool our data across all treatments and

[27] Recall that subjects played the same minimum effort task in Stage Four, following the whistleblowing game, as we aimed to test whether the occurrence of whistleblowing would affect firm cohesion. However, the low occurrence of actual whistleblowing in the game prevented us from conducting such analysis. This is because whistleblowing could only occur if the manager broke the law and if the employee randomly chosen (with a 50% chance) to determine payoffs was willing to blow the whistle. In practice, this occurred in only 3 out of 103 cases/firms. A simple comparison of the Stage 2 and Stage 4 minimum effort tasks shows a decline in the minimum effort observed both within firms (123.70 vs. 121.77, one-tailed p=0.080) and within members of the public (119.26 vs.116.67, one-tailed p=0.072).

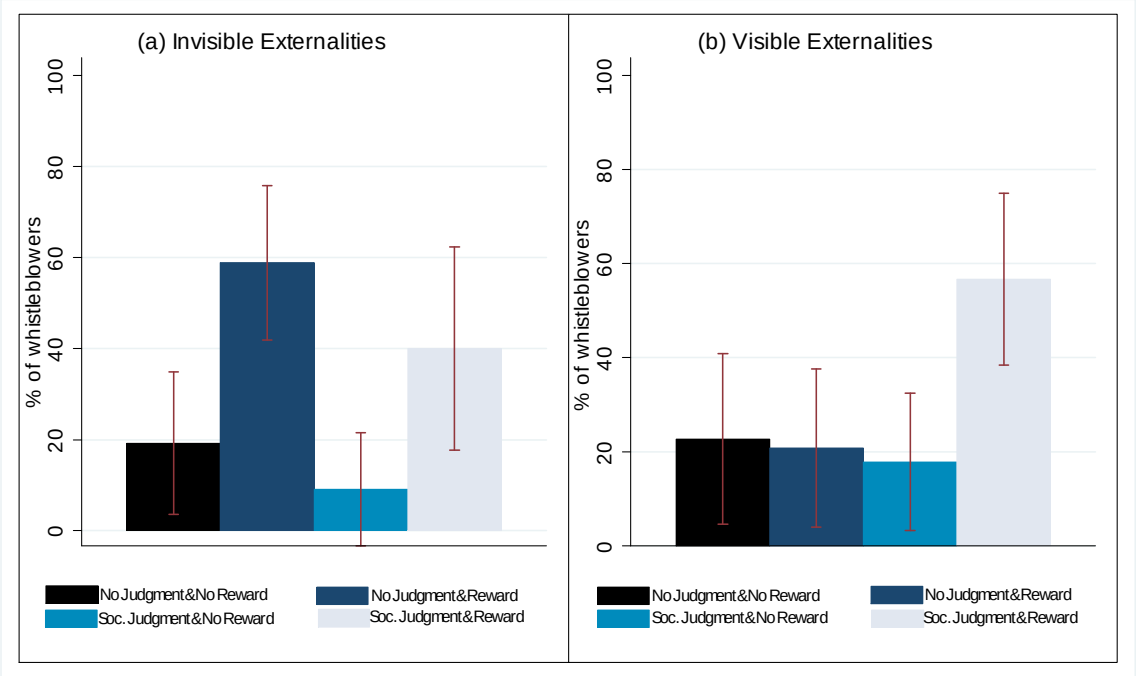

Figure 2: Whistleblowing across treatments.

	No Judgment & No Reward	No Judgment & Reward	Social Judgment & No Reward	Social Judgment & Reward
Invisible Externalities	19.23%	58.82%	9.09%	40.00%
H_0: Reward = No Reward	p-value = 0.002 (0.003) if Judgment=0		p-value = 0.019 (0.030) if Judgment=1	
H_0: No Judgment = Judgment	p-value =0.321 (0.429) if Rewards=0		p-value = 0.181 (0.260) if Rewards=1	
Visible Externalities	22.73%	20.83%	17.86%	56.67%
H_0: Reward = No Reward	p-value = 0.876 (1.000) if Judgment=0		p-value = 0.002 (0.003) if Judgment=1	
H_0: No Judgment = Judgment	p-value = 0.669 (0.732) if Rewards=0		p-value = 0.008 (0.012) if Rewards=1	

Note: P-values are generated by Chi-square tests. P-values from Fisher exact tests in parentheses.

Table 2: Whistleblowing under different treatments

	Dep. Variable: Dummy equal to 1 if employee is willing to blow the whistle, 0 otherwise					
	Invisible Ext.	Visible Ext.	All	All	All	All
Rewards	0.36***	0.21**	0.36***	0.35***	0.40***	0.36***
	(0.064)	(0.079)	(0.063)	(0.063)	(0.075)	(0.072)
Social Judgment	-0.15*	0.16*	-0.15**	-0.16**	-0.10	-0.14
	(0.068)	(0.079)	(0.067)	(0.065)	(0.072)	(0.082)
Visible Externalities			-0.10	-0.11	0.03	0.00
			(0.091)	(0.090)	(0.071)	(0.077)
Visible x Reward			-0.15	-0.15	-0.41***	-0.36***
			(0.100)	(0.099)	(0.093)	(0.097)
Visible x Social Judgment			0.31***	0.32***	0.05	0.10
			(0.103)	(0.112)	(0.091)	(0.113)
Social Judgment x Reward					-0.09	-0.03
					(0.129)	(0.131)
Judgment x Reward x Visible					0.49***	0.40**
					(0.156)	(0.174)
Firm performance/Own performance				-0.17**		-0.15**
				(0.070)		(0.071)
Constant	0.21***	0.11	0.21***	0.26	0.19***	0.19
	(0.052)	(0.077)	(0.051)	(0.313)	(0.058)	(0.321)
Controls	No	No	No	Yes	No	Yes
Observations	102	104	206	206	206	206

Controls are: gender, economics major, number of firms in the session, ratio between firm performance and own performance in team building task, and effort chosen in minimum effort task. We report the only control variable that is significant: the ratio between firm performance and own performance in the team building task. Robust standard errors, clustered at the session level, in parentheses; *** p<0.01, * * p<0.05, * p<0.1.

Table 3: Treatment effects

include interaction terms between the Reward and Social Judgment treatment dummies and the Visible Externality dummy. In column 4, we additionally include a set of control variables that include gender, whether the subject is an economics major and the number of firms in the session. In order to proxy for employees' loyalty to the firm, our set of controls also includes the ratio between firm performance and own performance in stage one of the experiment and the effort level chosen in the minimum effort game of stage two. The former variable captures the extent to which each employee may feel "indebted" to the other firm members for the earnings accumulated during the team-building stage, while the latter variable is a measure of firm cohesion, plausibly capturing trust and cooperation among firm members. In the final two columns, for completeness we include all interactions between our treatment dummies as well as a triple interaction term.

The first pattern that becomes apparent in Table 3 is that rewards have a substantial and statistically significant main effect on whistleblowing. The marginal effect of financial rewards is to increase the prevalence of whistleblowing by about 36 percentage points when the negative externalities of fraud are not visible to the public. Even when the externalities are visible to the public (columns 2), the estimated marginal effect is positive and large in magnitude (21 percentage points). This is confirmed by the estimates in columns 3 and 4. Table 6 in the Appendix reports the estimated marginal effects of financial rewards under the different treatment conditions, as generated by the regression analysis displayed in column 4 of Table 3.

The estimates displayed in columns 5 and 6 of Table 3 confirm that rewards are effective under invisible externalities and absent social judgment (first row), and no less effective in the presence of social judgment (insignificant coefficient of "Reward x Social Judgment" in row 6). Moreover, the combination of rewards with social judgment and visibility of negative externalities caused by fraud further increases the likelihood of whistleblowing (positive and significant coefficient on the triple interaction in row 7). The only case where financial rewards are less effective is when the negative externalities are visible to the public in the absence of social judgment (negative and significant coefficient of "Visible x Rewards" in row 4). We will return to the possible interpretation of this result at the end of this section. In order to facilitate the interpretation of the estimates reported in columns 5 and 6 of Table 3, we report the estimated marginal effects of financial rewards in all treatment conditions in the first row of Table 7 in Appendix.

Our first result follows.

Result 1 *Financial rewards generally increase employees' willingness to blow the whistle against a law-breaking manager.*

From Table 3 we can also reconfirm our impression of how the visibility of public harm interacts with social judgment to affect whistleblowing. In particular, either by considering Visible and Invisible treatments separately (columns 1 and 2) or by pooling the data and inspecting the estimated interactions between treatments, we can see that the possibility of social judgment substantially and significantly decreases whistleblowing when the public is unaware of the costs imposed on them by manager malfeasance. When these externalities are clear to the public, on the other hand, the possibility of social judgment tends to increase whistleblowing. The marginal effects of social judgment generated by the estimates in column 4 of Table 3 – reported in the second row of Table 6 in Appendix – show that the possibility of social judgment decreases whistleblowing by 16 percentage points if the externalities are not visible to the public, while increasing it by the same percentage points if the externalities are visible to the public.

Separating out the effects of all our treatment manipulations by interacting the different treatment dummies in columns 5 and 6 of Table 3 confirms our main social judgment results, yet it also shows that the possibility of social judgment increases whistleblowing when the externalities are visible only in the presence of financial rewards (insignificant coefficient of "Visible x Social Judgment" in row 5; significant and positive coefficient of the triple interaction in row 7).[28] These observations lead to our second result.

Result 2 *(a) When the negative externalities generated by fraud are not visible to the public, the possibility of social judgment decreases whistleblowing.*

(b) When the negative externalities generated by fraud are visible to the public, the possibility of social judgment either has no impact or increases whistleblowing.

Result 2 suggests that individuals in our experiment directly value social (dis)approval and that they expect social approval to be more likely when the public realizes manager malfeasance directly harms them and whistleblowers are financially rewarded. On the other hand, when whistleblowing might be interpreted as disloyalty toward the firm, an anti-social act, the result suggests employees expect public scrutiny to entail social disapproval.

[28] See also the marginal effects displayed in the second row of Table 7. While social judgment seems to have a negative effect on whistleblowing under invisible externalities, both with or without rewards (the coefficients are close to conventional levels of statistical significance) and a positive effect under visible externalities and rewards, the impact is null under visible externalities and no rewards.

Next, we consider our third hypothesis: whether financial incentives crowd out the salutary effect of social judgment on whistleblowing in our Visible Externalities treatments, which can be examined in various ways. First, consider the raw data as presented in Figure 3, Panel b. The effect of social judgment without rewards is the difference in heights between the third bar and first bar, while the effect of social judgment with rewards would be the height difference between the fourth and second bars. Crowding out would be consistent with the latter difference being smaller in magnitude than the former difference. However, this is clearly not the case. The figure suggests the effect of social judgment without rewards is essentially zero,[29] while the effect of social judgment with rewards is to increase whistleblowing by about 40 percentage points. More formally, the effect of rewards on the effect of social judgment in the Visible Externalities treatments can be seen in Table 3. Confirming appearances from Figure 3, the positive and significant coefficient in Column 6 on the triple interaction "Judgment x Reward x Visible" also suggests that rewards increase the effect of social judgment on whistleblowing by about 40 percentage points. All together, our data provide little support for Hypothesis 3.

Result 3 *Financial rewards do not weaken the effect of Social Judgment, i.e., we find no evidence of crowding-out of non-pecuniary motivations linked to expectations of social (dis)approval.*

Interestingly, however, we do find evidence for a different type of crowding out. The negative and significant interaction between Visible Externalities and Reward in columns 5 and 6 of Table 3 suggests that, absent social judgment, rewards are less effective in industries or cases of fraud where the public feels that it is directly affected by managers' law-breaking behavior. This pattern is also apparent in Figure 2, when comparing the first two bars in the left panel to the same two bars in the right panel: rewards strongly increase whistleblowing when whistleblowing is not subject to social judgment in the invisible externalities case, but have no effect when externalities are visible. Thus, in the absence of social judgment, externality visibility alters the effect of financial rewards.

Since we did not design our experiment to focus on this type of crowding out, we can only speculate about the underlying mechanism. One possibility is that individuals' *intrinsic* motivations associated with whistleblowing are higher when the externalities are visible to the public; in this case, the introduction of financial rewards, absent public scrutiny, crowds out these motivations, resulting in an overall null effect of rewards. Another possibility is that the moral environment is more complex than we have been assuming and that, for example, whistleblowers learn about their own motivations through their actions – they "self-signal", in the terminology of Benabou and Tirole (2006). In this setting, when the whistleblower knows that the public is not aware of the costs imposed on them, blowing the whistle simply expresses a preference for justice or fairness – punishing the manager for bad behavior. When the whistleblower knows the public is aware of the harm imposed on them, motivations become more difficult to disentangle and, in particular, the "choosing sides" aspect – i.e., empathizing more with the public than with the in-group (firm) – becomes more salient. Abstaining from whistleblowing would then become a self-signal about loyalty to the firm, made stronger by forgoing financial rewards, which might generate the pattern observed in the data. In Appendix Section B we discuss more formally how the pattern can be explained in our simple theoretical framework.

[29] To see how this null result could be explained in the context of our theoretical framework, please refer to the Appendix, Section B.

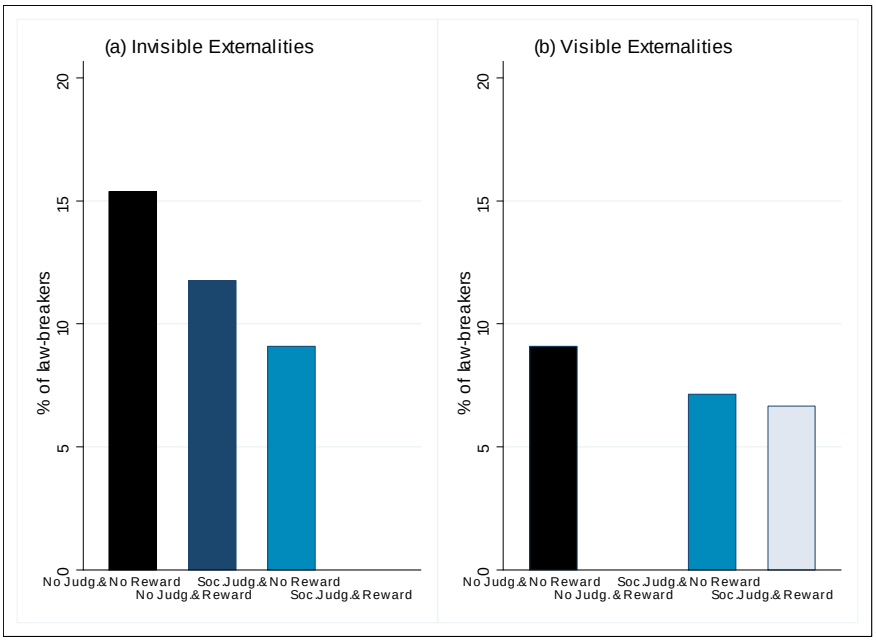

Figure 3: Manager law-breaking across treatments

4.2 Manager's Law-Breaking Behavior

Our experiment was primarily meant to investigate employees' decision to blow the whistle against their manager. As a consequence, our sample of managers is quite small, with a total of 103 observations. Overall, about 8% of managers decided to break the law to double the firm fund at the expense of the members of the public. The occurrence of law-breaking varies across treatments, as shown in Figure 4. Table 8 in Appendix reports the law-breaking statistics by treatment and the p-values generated by Chi-square tests. A clear pattern we see in the data is the reduction in managers' illegal behavior when there exist financial rewards for whistleblowers, suggesting that managers expect rewards to increase employees' willingness to report wrongdoing and that, consequently, whistleblower rewards may have substantial preventive or deterrent effects on corporate crime. However, the small sample size and generally low frequency of law-breaking preclude us from investigating further the underlying pattern.

Result 4 *There are no statistically significant differences in manager law-breaking across treatments, although the amount of law breaking appears lower in the presence of rewards for whistleblowers.*

We can investigate another determinant of manager behavior, however. Regression analysis[30] provides evidence of the impact of the manager's skills on the probability of breaking the law. In particular, the manager's performance in the stage one multiplication task is negatively related to the probability the manager breaks the law. This finding seems in line with Baloria et al. (2015), who document that the companies that lobbied against the whistleblower rewards provision in the Dodd-Frank Act were precisely the less well run companies with weaker compliance programs and poorer governance structures (e.g., less separation between Chairman and CEO). These are also the firms for which whistleblower rewards are

[30] The corresponding table is not reported here but is available from the authors upon request. The estimates also show that none of the treatments had a signficant impact on the manager's decision to break the law.

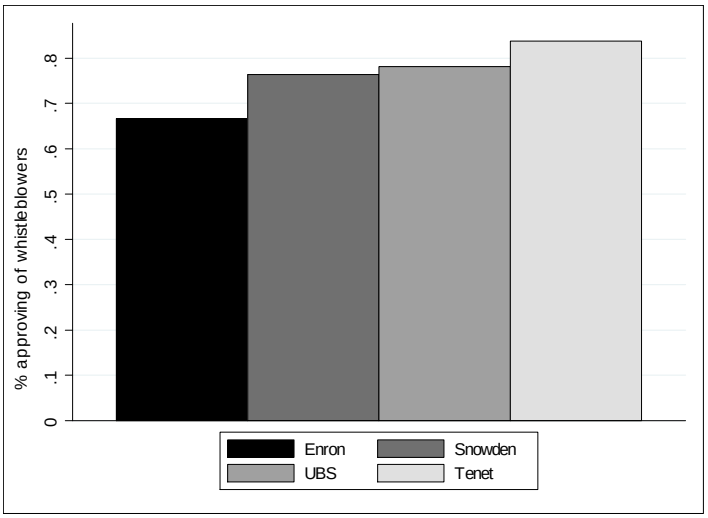

Figure 4: Social judgment of four whistleblowing cases (survey)

perceived by the market to be more needed and more likely to have positive effects in terms of improving management/governance and protecting shareholders.

4.3 Social Judgment of Whistleblowers

In this section, we investigate the social judgment of whistleblowers under different conditions. We start by analyzing individual answers to post-experiment survey questions eliciting opinions on the social appropriateness or inappropriateness of actual whistleblowing cases. As part of our post-experiment survey, all study participants were presented with four actual whistleblowing cases – the Snowden case, the Enron case, the UBS case and the Tenet case – and asked to evaluate the social appropriateness of blowing the whistle in each case. As discussed in Section 3.1, we chose these cases because they vary in the visibility of the negative externalities that illegal behavior caused to the public and in the presence of financial rewards for the whistleblower. The social costs of the unlawful actions unmasked by the whistleblower are clearly visible in the Snowden (national security) and the Tenet (health care) cases, less visible in the UBS (tax evasion) case and even less visible in the Enron (earnings management) case. Moreover, financial rewards were present in the UBS and Tenet cases and not in the Enron and Snowden cases.[31]

Figure 3 reports the percentages of survey participants stating that the decision made by the whistleblower is socially acceptable. The social acceptability of whistleblowing is lowest in the Enron case and highest in the Tenet case ($p = 0.000$). Pairwise comparisons between cases suggest that both the visibility of the externalities and the presence of financial rewards may be responsible for the observed increase in the social acceptability of the whistleblowing act. Naturally, this is only suggestive evidence. Next, we analyze the messages that the members of the public sent to whistleblowers in our social judgment treatments.

[31] In order to minimize ordering effects, the four cases were presented in the above order, but not one after the other. Subjects were first presented with the Snowden case and were then asked a number of unrelated questions collecting demographics and attitudinal preferences, they then saw the Enron case, followed by more unrelated questions. The UBS case came afterwards, followed by more questions before the appearance of the Tenet case. For each whistleblowing scenario, we provided a summary of the case and we asked subjects to rank the appropriateness of the whistleblower's decision.

	Invisible Externalities				Visible Externalities			
	Happy	Neutral	No message	Unhappy	Happy	Neutral	No message	Unhappy
No Reward	72.22%	5.56%	11.11%	11.11%	38.89%	33.33%	22.22%	5.56%
Reward	77.78%	5.56%	16.27%	0%	62.50%	16.67%	20.83%	0%

Table 4: Percentages of Members of the Public that sent each kind of message to whistleblowers

Overall, across all treatments, 18% of members of the public decided to send no message to the whistleblowers, 63% sent a message of approval, 4% sent a message of disapproval, and the remaining 15% sent a neutral message. Table 4 reports the percentages of members of the public that sent each possible message, or no message, under the different treatment manipulations. The presence of rewards seems to induce the members of the public to (not)send (frowny)smiley faces to whistleblowers, especially when the externalities are visible to the public ($p = 0.13$). This is in line with our finding of the positive impact of rewards on whistleblowing, and suggests that potential whistleblowers correctly anticipated that the presence of financial rewards would not negatively affect the judgment that members of the public would have of them. A plausible interpretation of this result is that rewards signal to the public the "right thing to do," as suggested by legal theories on the expressive role of the law (e.g. Sunstein, 1996).

Contrary to our expectations, the visibility of the social cost of fraud does not lead to increased social approval of whistleblowers. If anything, when the public is aware of the monetary loss that they suffer because of manager rule-breaking, they are less likely to send smiley faces to whistleblowers ($p = 0.044$ in the no rewards treatments and $p = 0.289$ in the rewards treatments). This is a puzzling finding that may either indicate that members of the public somehow held the whistleblower responsible for their monetary loss – as they saw him/her as a member of the fraudulent firm – or that the members of the public see the messages as tools to express their general feelings, i.e., either satisfaction or dissatisfaction with the outcomes of the experiment. While this is less than ideal, as we would rather have the members of the public view and use the messages as instruments to express approval or disapproval of whistleblowers, the analysis of the public's messaging behavior is only tangentially relevant to our investigation. In fact, what matters for our research question is the whisteblowers' anticipation of the messages of approval or disapproval that they would receive under the different treatment manipulations.

5 Conclusion

Our study contributes to the policy debate and growing literature on the motivations and incentives for employees to blow the whistle on corporate fraud. Despite being splashed across the covers of popular journals in recent years, whistleblowing is rare and the vast majority of white-collar crime remains undetected and unpunished (Dyck et al., 2013). In this paper, we examined two policies that may motivate employees to blow the whistle on white-collar crime: the use of financial rewards, and the protection (exposure) of whistleblowers from (to) public scrutiny and social judgment. We also examined the interaction between these two sources of whistleblowing incentives and tested whether financial rewards may crowd out non-pecuniary motivations linked to expectations of social approval. Finally, we asked whether different policies should be used for different cases of fraud or different industries, depending on whether the public feels directly affected by the negative externalities generated by the illegal activities undertaken within the organization, as discussed in the legal debate.

We employed a specially designed laboratory experiment that allowed us to observe willingness to break the law, willingness to blow the whistle on rule breaking, and public reaction to whistleblowing. Crucially, in our setting, manager wrongdoing caused financial losses to "real" third parties, while potential whistleblowers did not take part in the illegal activities but benefited from them, and whistleblowing was costly.

We found strong evidence of the effectiveness of financial rewards on whistleblowing. We did not find evidence of crowding out of non-pecuniary motivations driven by a preference for social approval. Our data also show that financial rewards are more effective when the whistleblower is subject to social judgment than when he/she is not. Our findings with respect to the relationship between whistleblowing and public scrutiny show that the possibility of social judgment may act as either an incentive for, or a deterrent against, blowing the whistle. Social judgment acts as a deterrent when the public does not feel directly affected by the negative externalities caused by corporate fraud, and may act as an incentive when the opposite holds. This suggests that, in order to maximize whistleblowing, industries and corresponding cases of fraud should be classified based on the perceived negative effects they have on the public and different policies should be adopted, either protecting or exposing whistleblowers. Overall, our results confirm previous research on the effectiveness of financial rewards for whistleblowing and provide novel insights about the interaction between financial incentives and whistleblowers' concerns about social judgment.

References

Abbink, K. & K. Wu. (2017). "Reward Self-Reporting to Deter Corruption: An Experiment on Mitigating Collusive Bribery." Journal of Economic Behavior and Organization, 133: 256-272.

Abbink, K., U. Dasgupta, L. Ghanghadaran & T. Jain. (2014). "Letting the Briber Go Free: An Experiment on Mitigating Harassment Bribes." Journal of Public Economics 111: 17-28..

Alekseev, A., Charness, G., and U. Gneezy, (2016). "Experimental Methods: When and Why Contextual Instructions are Important." Journal of Economic Behavior and Organization http://dx.doi.org/10.1016/j.jebo.2016.12.0

Andreoni, J., & Bernheim, B. D., (2009). "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." Econometrica, 77(5), 1607-1636.

Andreoni, J., & Petrie, R. (2004). "Public goods experiments without confidentiality: a glimpse into fund-raising." Journal of Public Economics, 88(7), 1605-1623.

Andreoni, J. (1995). "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments." The Quarterly Journal of Economics, vol. 110, no. 1: pp. 1–21.

Apesteguia J., M. Dufwenberg and R. Selten, (2007). "Blowing the Whistle." Economic Theory, vol. 31: 143-166.

Ariely, D., Bracha, A., & Meier, S., (2009). "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." The American Economic Review, 99(1): 544-555.

Ashforth, B.E., and F. Mael (1989). "Social identity theory and the organization." Academy of management review, 14(1): 20-39.

Aubert, C., Kovacic, W., and Rey P. (2006). "The Impact of Leniency and Whistleblowing Programs on Cartels." International Journal of Industrial Organization, 24: 1241-1266.

Baloria, V. Marquardt, C. Wiedman, C. (2015). "A Lobbying Approach to Evaluating the Whistleblower Provisions of the Dodd-Frank Reform Act of 2010." Unpublished working paper. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract id=1923310.

Bank of England Prudential Regulations Authority & Financial Conduct Authority (2014). "Financial Incentives for Whistleblowers." Available at: https://www.fca.org.uk/news/financial-incentives-for-whistleblowers.

Bartuli, J., Djawadi, B., Fahr, R. (2016). "Business Ethics in Organizations: An Experimental Examination of Whistleblowing and Personality." IZA Discussion Paper No.10190.

Bénabou, R., and Tirole, J. (2006). "Incentives and prosocial behavior." The American Economic Review, 96(5): 1652-1678.

Bigoni, M., Fridolfsson, S., Le Coq, C.and Spagnolo, G. (2012). "Fines, Leniency and Rewards in Antitrust." RAND Journal of Economics, 43(2): 368-390.

Bigoni, M., Fridolfsson, S., LeCoq, C., and Spagnolo, G. (2015). "Trust, Leniency and Deterrence." Journal of Law, Economics and Organization 31(4): 663-689.

Bolton, G.E., and A. Ockenfels. (2000). "ERC: A theory of equity, reciprocity, and competition." American Economic Review, 90(1): 166-193.

Branas-Garza, P. (2007). "Promoting helping behavior with framing in dictator games." Journal of Economic Psychology, Elsevier, vol. 28(4), pages 477-486.

Brandts, J. and G. Charness, (2000). "Hot vs. cold: Sequential responses and preference stability in experimental games." Experimental Economics, 2(3): 227-238.

Brandts, J. and G. Charness, (2011). "The strategy versus the direct-response method: A first survey of experimental comparisons." Experimental Economics, 14(3): 375-398.

Breuer, L. (2013). "Tax Compliance and Whistleblowing: The Role of Incentives." The Bonn Journal of Economics, Vol 2, No. 2: 7-44.

Buckenmaier, J., Dimant E. and L. Mittone (2017). "Experimental Evidence on Tax Evasion, Corruption and Incentives to Blow the Whistle." Working paper.

Bursztyn, L., & Jensen, R. (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." Annual Review of Economics, 9, 131-153.

Call, A.C., G.S. Martin, N.Y. Sharp, and J.H. Wilde. (2017). "Whistleblowers and Outcomes of Financial Misrepresentation Enforcement Actions." Journal of Accounting Research, forthcoming.

Carpenter, J., and Myers, C. K. (2008). "Why volunteer? Evidence on the role of altruism, image, and incentives." Journal of Public Economics 94(11): 911-920.

Carpenter, J., Robbett, A., & Akbar, P. A. (2017). "Profit Sharing and Peer Reporting." Management Science: https://doi.org/10.1287/mnsc.2017.2831

Carson, T., Verdu, M., and Wokutch, R. (2008). "Whistle-Blowing for Profit: An Ethical Analysis of the Federal False Claims Act." Journal of Business Ethics, Vol. 77(3): 361-376.

Charness, G., and Rabin, M. (2002). "Understanding social preferences with simple tests." The Quarterly Journal of Economics, 117(3): 817-869.

Chen, R., and Chen, Y. (2011). "The potential of social identity for equilibrium selection." American Economic Review 101.6: 2562-89.

Cookson, R. (2000). "Framing effects in public goods experiments." Experimental Economics 3.1: 55-79.

Cotten S. and Santore R. (2016). "Whistleblowers, Amnesty, and Managerial Fraud: An Experimental Investigation." Unpublished manuscript.

Dozier, J. B., & Miceli, M. P. (1985). "Potential predictors of whistle-blowing: A prosocial behavior perspective." Academy of Management Review, 10(4), 823-836.

Dufwenberg, M., Gächter, S and Hennig-Schmidt, H. (2011). "The framing of games and the psychology of play." Games and Economic Behavior 73.2 (2011): 459-478.

Dyck, A., Morse, A. and Zingales, L. (2010). "Who Blows the Whistle on Corporate Fraud?" The Journal of Finance, 65: 2213-53.

Dyck, A., Morse, A. and Zingales, L. (2013). "How pervasive is corporate fraud?" Rotman School of Management Working Paper 2222608.

Eckel, C. and Grossman, P.J. (1996). "Altruism in anonymous dictator games." Games and Economic Behavior 16: 181-191.

Engstrom, D.F. (2012). "Harnessing the Private Attorney General: Evidence from Qui Tam Litigation." Columbia Law Review, 112: 1244-1325.

Engstrom, D.F. (2013). "Public Regulation of Private Enforcement: Empirical Analysis of DOJ Oversight of Qui Tam Litigation Under the False Claims Act." Northwestern University Law Review, 107, 1689-1756.

Engstrom, D.F. (2014a). "Private Enforcement's Pathways: Lessons from Qui Tam Litigation." Columbia Law Review, 115: 1913-2006.

Engstrom, D.F. (2014b). "Whither Whistleblowing? Bounty Regimes, Regulatory Context, and the Challenge of Optimal Design." Theoretical Inquiries L.15 605.

Ethics Resource Center (2014). "National Business Ethics Survey of the U.S. Workforce." Available at: http://www.ethics.org/ecihome/research/nbes, accessed 31/10/2016.

Fehr, E., and K.M. Schmidt (1999). "A theory of fairness, competition, and cooperation." The Quarterly Journal of Economics, 114(3): 817-868.

Fehr, E., & U. Fischbacher (2004). "Third-party punishment and social norms." Evolution and human behavior, 25(2): 63-87.

Fehr, E., and S. Gächter (2002). "Do Incentive Contracts Undermine Voluntary Cooperation?" Working Paper 34, Institute for Empirical Research in Economics, University of Zurich.

Fehr, E., and J.A. List (2004). "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOs." Journal of the European Economic Association, 2(5): 743–71.

Feldman, Y., Lobel, O. (2010). "The Incentives Matrix: The Comparative Effectiveness of Rewards, Liabilities, Duties and Protections for Reporting Illegality." Texas Law Review, Vol 87

Felli, L. and R., Hortala-Vallve (2016). "Collusion, Blackmail and Whistle-Blowing." Quarterly Journal of Political Science, 11(3): 279-312.

Fischbacher, U. (2007). "z-Tree: Zurich toolbox for ready-made economic experiments." Experimental Economics 10.2: 171-178.

Frey, B.S., and F. Oberholzer-Gee (1997). "The cost of price incentives: An empirical analysis of motivation crowding-out." The American Economic Review 87.4: 746-755.

Friebel, G. and S. Guriev (2012). "Whistle-Blowing and Incentives in Firms," Journal of Economics & Management Strategy, 21(4): 1007-1027.

GAO-11-619 (2011). U.S. Gov't Accountability Office Report 11-619, Criminal Cartel Enforcement: Stakeholder Views on Impact of 2004 Antitrust Reform Are Mixed, but Support Whistleblower Protection, available at http://www.gao.gov/new.items/d11619.pdf.

Gerber, A. S., Green, D. P., & Larimer, C. W., (2008). "Social pressure and voter turnout: Evidence from a large-scale field experiment." American Political Science Review, 102(01): 33-48.

Givati, Y. (216). "A Theory of Whistleblower Rewards." The Journal of Legal Studies 45:1, 43-72.

Gneezy, U., Meier, S., and P. Rey-Biel (2011). "When and why incentives (don't) work to modify behavior." The Journal of Economic Perspectives 25.4: 191-209.

Gneezy, U., and A. Rustichini. (2000a). "Pay Enough or Don't Pay At All." Quarterly Journal of Economics, 115(3): 791–810.

Gneezy, U., and A. Rustichini. (2000b). "A Fine Is a Price." Journal of Legal Studies, 29(1): 1–18.

Greiner, B. (2015). "Subject pool recruitment procedures: organizing experiments with ORSEE." Journal of the Economic Science Association 1.1: 114-125.

Gundlach, M. J., Douglas, S. C. and M. J. Martinko (2003). "The decision to blow the whistle: A social information processing framework." Academy of management Review 28.1: 107-123.

Hamaguchi, Y., Kawagoe, T. and A. Shibata, (2009). "Group Size Effects on Cartel Formation and the Enforcement Power of Leniency Programs." International Journal of Industrial Organization 27(2): 145-165.

Heyes, A. and S. Kapur, (2009). "An Economic Model of Whistle-Blower Policy." Journal of Law, Economics, & Organization, 25(1): 157-182.

Hinloopen, J. and A. Soetevent, (2008). "Laboratory Evidence on the effectiveness of corporate leniency programs." RAND Journal of Economics, 39(2): 607-616.

IRS (2015), IRS Whistleblower Program, Fiscal Year 2015, Annual Report to the Congress.

Linardi, S., & McConnell, M. A., (2011). "No excuses for good behavior: Volunteering and the social environment." Journal of Public Economics, 95(5), 445-454.

Liu, Y., Zhao, S., Li, R. et al. (2018). "The relationship between organizational identification and internal whistle-blowing: the joint moderating effects of perceived ethical climate and proactive personality." Review of Managerial Science 12: 113-134.

Miceli, M. P. and Janet P. Near (1992). "Blowing the Whistle: The Organizational and Legal Implications for Companies and Employees." Lexington Books.

Miceli, M. P. and Janet P. Near (1994). "Relationship among Value Congruence, Perceived Victimization, and Retaliation against Whistle-blowers." Journal of Management 20(4), 773-794, Lexington Books.

Nyrerod, T and Spagnolo, G. (2018). "Myths and Numbers on Whistleblower Rewards." Dp 12957, London, Centre for Economic Policy Research. https://www.cepr.org/active/publications/discussion_papers/dp.php?dpno

Rege, M., and K. Telle (2004). "The impact of social approval and framing on cooperation in public good situations." Journal of Public Economics 88.7: 1625-1644.

Reuben, E., and M. Stephenson (2013). "Nobody likes a rat: On the willingness to report lies and the consequences thereof." Journal of Economic Behavior & Organization 93 (2013): 384-391.

Rothschild, J., and T. Miethe (1999). "Whistle-Blower Disclosures and Management Retaliation." Work and Occupations, 26(1): 107-128.

Salmon, T. C., & Serra, D. (2017). "Corruption, social judgment and culture: An experiment." Journal of Economic Behavior & Organization, 142: 64-78.

Schmolke, K.U. and V. Utikal (2016). "Whistleblowing: Incentives and Situational Determinants." FAU - Discussion Papers in Economics, No. 09/2016. Available at SSRN: https://ssrn.com/abstract=2820475 or http://dx.doi.org/10.2139/ssrn.2820475.

Serra, D. (2012). "Combining top-down and bottom-up accountability: evidence from a bribery experiment." Journal of Law, Economics, and Organization, 28(3), 569-587

Shikora, J. (2011). "Bringing good and bad Whistle-blowers to the Lab." Munich Discussion Paper No. 2011- 4, online: http://epub.ub.uni-muenchen.de/12161/.

Spagnolo, G., (2004). "Divide et Impera: Optimal Leniency Programs." CEPR Discussion Papers 4840.

Spagnolo, G., (2008). "Leniency and Whistleblowers in Antitrust." Ch.12 of P.Buccirossi (Ed.) Handbook of Antitrust Economics, 2008, M.I.T. Press.

Sunstein, Cass R. (1996). "On the expressive function of law." University of Pennsylvania law review 144(5): 2021-2053.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). "Social categorization and intergroup behaviour." European Journal of Social Psychology, 1(2), 149-178.

Titmuss, R.M. (1970). The Gift Relationship: From Human Blood to Social Policy (1970). Reprinted by the New Press, ISBN 1-56584-403-3 (reissued with new chapters 1997, John Ashton & Ann Oakley, LSE Books)

Waytz, A., Dungan, J., and L. Young (2013). "The whistleblower's dilemma and the fairness-loyalty tradeoff." Journal of Experimental Social Psychology, Vol. 49, pp. 1027-1033.

Wilde, J.H. (2017). "The Deterrent Effect of Employee Whistleblowing on Firms' Financial Misreporting and Tax Aggressiveness." The Accounting Review, In-Press.

Xiao, E., and D. Houser (2011). "Punish in public." Journal of Public Economics, 95(7), 1006-1017.

APPENDIX

In this section we construct an ad-hoc framework which incorporates two forms of non-pecuniary concerns likely to play a role in whistleblowing. In this framework, an individual's overall utility is a function of money, personal moral concerns and social approval. We assume no necessary relationship between social approval and morality. Actions that garner social approval may not be moral, according to an individual's own moral standards, and *vice-versa*.

Each employee has two possible actions: blow the whistle, which we denoted by w, or remaining silent, $\neg w$. Employees also have subjective beliefs about which of these actions is moral given the context, c, and which will garner social approval. Beliefs in our framework take a particularly simple form: with probability $p(c)$, the employee believes that w is the moral action in the current context; and with the remaining probability, $1 - p(c)$, s/he believes that $\neg w$ is the moral action. Similarly, beliefs about the probability that w ($\neg w$) will garner social approval are given by $q(c)$ $(1 - q(c))$.

We assume that overall utility is composed of three components: money utility, moral utility and social (approval) utility. For the money utility component, we assume risk neutrality so that $u(\$x) = x$. For the moral utility component, taking the moral (immoral) action raises (lowers) utility by one unit, so that in expectation the moral utility of w is given by $p(c) \times 1 + (1 - p(c)) \times (-1) = 2p(c) - 1$. The social utility component takes a similar form, so that in expectation social utility of w is $2q(c) - 1$.

There is heterogeneity in the population of employees with regard to how much weight moral and social concerns carry in decision making. We represent these weights with idiosyncratic parameters $\beta_i \geq 0$ and $\gamma_i \geq 0$, with cumulative distribution functions M, S. The associated density functions, m, s, have finite support on the non-negative reals. Employee i's overall utility from blowing the whistle or remaining silent can be written as:

$$\begin{aligned} U_i(w, p, q, c) &= x(w) + \beta_i(2p(c) - 1) + \gamma_i(2q(c) - 1), \\ U_i(\neg w, p, q, c) &= x(\neg w) + \beta_i(1 - 2p(c)) + \gamma_i(1 - 2q(c)). \end{aligned}$$

In these equations we write p, q as functions of the context, c, which we manipulate with our experimental treatments. We assume that social approval is more likely when the public knows they are being harmed and that, in contexts without the possibility of social judgment, social utility does not enter into decision-making. Rather than introducing an indicator function for the presence of social judgment in this latter case, for notational simplicity we simply set $q = \frac{1}{2}$ when social judgment is not possible.

$$\begin{aligned} q(\text{Visible Externalities}) &> q(\text{Invisible Externalities}), \\ q(\text{No Judgment}) &= \tfrac{1}{2}. \end{aligned}$$

We assume that by creating firm loyalty through the Identity Building stage of our experiment, that in each context (at least some) employees will believe that whistleblowing is the immoral action:

$$p(c) < \tfrac{1}{2} \text{ for each c, for some employees.}$$

A Whistleblowing in our framework

In the discussion that follows, we suppress the dependence of p, q on context in our notation for ease of exposition. We also ignore knife-edge cases of exact indifference between whistleblowing and not.

A.1 Baseline case: no direct interactions among utility components

To begin familiarizing ourselves with the framework while investigating what it might say about whistleblowing, we begin with a very simple case. Here we assume that there are no direct interactions among the three components of utility – monetary, moral and social. This rules out, e.g., any direct effect of monetary rewards on the morality of whistleblowing (p) or the likelihood of social approval (q).

We can characterize the likelihood of whistleblowing as the probability that the utility of whistleblowing exceeds the utility of remaining silent. Using our expressions for the utility of each action above, this probability is governed by the threshold:

$$x(w) - x(\neg w) > 2[\beta_i(1-2p) + \gamma_i(1-2q)]. \tag{1}$$

Notice that p and q, the probabilities that whistleblowing is moral and will induce social approval, respectively enter with a negative sign into the RHS of the inequality. Thus, the RHS can be interpreted as the moral and social costs of whistleblowing, while the LHS represents the net monetary benefit of whistleblowing. An employee therefore blows the whistle whenever the net monetary benefit exceeds the subjective non-monetary costs of whistleblowing.

From Equation 1, it is clear that monetary rewards for whistleblowers strictly increase the LHS of the equation without directly affecting the RHS. Thus, for a given p, q, M and S, (weakly) more employees will find whistleblowing to be the utility-maximizing action. This leads us to our first observation: absent any direct effect of rewards on moral or social concerns, an increase in monetary whistleblower rewards (weakly) increases the likelihood of whistleblowing in the population. The amount of an increase in whistleblowing, if any, will depend on the distribution of decision weights given to moral and social concerns, M and S.

Observation 1 Increasing monetary rewards (weakly) increases the likelihood of whistleblowing.

Next, we focus on the RHS of 1. Notice that the RHS is strictly increasing in $1-p$, the subjective likelihood that whistleblowing is immoral, as well as $1-q$, the subjective likelihood of social disapproval. Consequently, contextual factors which increase p or q will reduce the RHS. For a given level of monetary rewards, then, an increase in either p or q will weakly increase the population probability of blowing the whistle. The level of the increase will depend again on the distribution of the moral and social decision weights in the population.

We have assumed above that one contextual factor that increases the subjective likelihood of social approval is the visibility of externalities. Suppose we make the additional assumptions that $q(\text{Visible Externalities}) > \frac{1}{2} > q(\text{Invisible Externalities})$. In words these assumptions say that the public (on average) approves of whistleblowing when it knows about the harm imposed upon it by manager malfeasance; and when it does not know about this harm it disapproves. These assumptions tell us that when social concerns enter the employee's utility function – in our Social Judgment treatments – these concerns increase (decrease) the RHS of Equation 1 when when externalities are invisible (visible). Our second observation is therefore that social judgment will weakly increase the prevalence of whistleblowing in the population when externalities are visible compared to the case with no social judgment; and, compared to the no judgement case, when externalities are invisible social judgment will weakly decrease the prevalence of whistleblowing.[32]

[32] Notice that absent the additional assumptions about q crossing the threshold of $\frac{1}{2}$, we would still have an implication

Observation 2 Social judgment weakly (decreases) increases whistleblowing when externalities are (invisible) visible.

A.2 Relaxing the no direct interactions assumption

Our first two observations above relied on the assumption that there were no interactions among the three components of utility. To see the complications inherent in relaxing this assumption, consider assuming a negative interaction between morality and social approval. Suppose we assume that whenever p increases q decreases and *vice-versa*. Then even though it is reasonable to also assume that visible externalities increase the likelihood of social approval, we could rationalize any observed change in whistleblowing with an ancillary assumption about the strength of the negative effect of q on p. If social judgment with visible externalities actually decreases whistleblowing, we could rationalize this with a more-than-offsetting decrease in the morality of whistleblowing.

Consequently, we are cautious in relying on interactions among monetary and non-monetary components, focusing only on the most straightforward interaction in this section. Specifically, we examine a particular reduced-form way of incorporating "crowding out." A frequently raised concern is that monetary incentives may crowd out non-monetary incentives. An obvious and straightforward way to capture this concern in our framework is to assume a negative relationship between the monetary gain from whistleblowing, $x(w)$, and either the moral or social utility components of utility. While the analysis would be essentially the same for either of these components, the justification is more straightforward for the latter. We also restrict attention to the Visible Externalities case, where social judgment should generally increase the prevalence of whistleblowing.

To incorporate crowding out, we assume that there is a negative relationship between monetary rewards and q, the employee's belief that whistleblowing will induce social approval. Thus higher monetary rewards reduce the employees' expectations of social approval. This could be because the employee believes the public is more likely to attribute whistleblowing to greed when there are rewards than if there were no rewards, and greed is socially disapproved of.

For concreteness fix the monetary consequences of whistleblowing at 5, with rewards or -5, without rewards. Let $p = \frac{1}{4}$ and the distribution of β_i be degenerate with point mass 6. Let γ_i be uniformly distributed on $[0, 10]$. Denote by q the likelihood of social approval without rewards and by $q' < q$ the likelihood of social approval with rewards. In particular, assume that without rewards whistleblowing is certainly met with social approval ($q = 1$), but that when whistleblowers receive financial rewards social approval is only slightly more likely than disapproval: $q' = \frac{53}{100}$.

We can compute the effectiveness of social approval without rewards as the difference in the probability of whistleblowing with and without social judgment. Given our assumptions, without social judgment and without rewards, whistleblowing never occurs: $-5 < 2\beta_i(1-2p) = 6$. But with social judgement, whistleblowing occurs whenever:

$$-5 > 2[\beta_i(1-2p) + \gamma_i(1-2q)] = 6 - 2\gamma_i.$$

Consequently, employees with $\gamma_i > frac112$ blow the whistle. The effectiveness of social judgment is therefore to increase the probability of whistleblowing by $\frac{9}{20}$.

of the visibility of externalities. However, we would only know that social judgment would be less effective at inducing additional whistleblowing in the population when externalities are invisible than when externalities are visible.

We can repeat the same exercise with rewards. Again, without social judgment whistleblowing never occurs because $5 < 2\beta_i(1-2p) = 6$. With social judgment, however, employees blow the whistle whenever:

$$5 > 6 + 2\gamma_i(1 - 2q')] = 6 - \gamma_i \frac{12}{100}.$$

This inequality is satisfied for $\gamma_i > 8\frac{1}{3}$. The effectiveness of social judgment when there are rewards is therefore to increase the probability of whistleblowing by $\frac{5}{30}$.

Since $\frac{5}{30}$ is much smaller than $\frac{9}{20}$, we see how the addition of monetary rewards in the case where rewards directly reduce the likelihood of social approval may crowd out the effectiveness of social judgment.

Observation 3 Crowding out can be explained by a negative interaction between rewards and social judgment in our framework.

It is also important to notice that crowding out in the sense just mentioned is not a necessary consequence of a negative interaction between rewards and social approval in our framework. Reducing the effectiveness of social judgment requires a confluence of features that depend on the size of monetary rewards, the distribution of the social utility parameter γ, and the strength of the negative interaction. For instance, if we keep all the parameters of our example above but use a slightly higher $q' = \frac{6}{10}$, then the effectiveness of social judgment with rewards can be calculated to be $\frac{3}{4}$, which is greater than $\frac{9}{20}$, the effect without rewards.

Rather than considering all other possible interactions and their implications for whistleblowing, we now turn directly to relating our framework to our experimental results.

B Relating our experimental results to our framework

In this section, we relate our framework to our experimental results. We begin by describing how we believe our treatments relate to the parameters of our framework.

We assume that the distribution of β_i and γ_i in the population was fixed and not affected by our treatments. Through our identity-building exercise we attempted to create a moral tension between being disloyal to one's firm (low p) and exposing illegal activity (high p). We assume the overall effect is to make whistleblowing on average immoral for at least some employees.

The primary parameters we sought to manipulate were the likelihood of social approval, q, and monetary rewards. We implemented two levels of monetary rewards for whistleblowing, one positive and one negative. In our No Rewards treatments, the monetary consequences of blowing the whistle were $x(w) = -\$5$, while not blowing the whistle yielded $x(\neg w) = \$0$. In our Rewards treatments, $x(w) = \$5$ while $x(\neg w) = \$0$. We attempted to manipulate employees beliefs about q indirectly by varying whether the public was aware of the harm done to them by the managers' illegal behavior. In particular, we interpret our Visible Externalities treatments as the employees' subjective belief they will receive social approval, q, as the public is more likely to view whistleblowers as acting on their behalf. On the other hand, we assume that in our Invisible Externalities treatments the employee may believe that the public may view whistleblowing as an act of disloyalty, thus decreasing q.

Given this description, we first note than many of our results are clearly consistent with the simplest form of our framework – without interactions among the components of utility. We observed above

that in this simplest case monetary rewards generally increase whistleblowing (Observation 1 above and Result 1 in the text). Also in this case, we have shown that our framework is consistent with social judgment either increasing or decreasing whistleblowing according to the likelihood of social approval or disapproval (Observation 2, Result 2). Turning to the case with interactions among utility components, we also observed the absence of evidence for crowding out (Result 3) is consistent with our framework – as would have been evidence for crowding out (Observation 3). Finally, the evidence we find for crowding out in our No Judgment treatments, mentioned at the end of Section 4.1 in the paper, is also consistent with our framework as noted above.

To be more concrete, we provide a numerical example illustrating how our framework can be made consistent with a particular pattern in our data.

B.1 A contrived numerical example

In this numerical example, we demonstrate that our simple framework, without interactions among motivations, can accommodate the pattern of social judgment being more effective at inducing whistleblowing when combined with monetary rewards than without monetary rewards. Assume for simplicity that whistleblowing is certainly immoral but likely to garner social approval: $p = 0, q = \frac{4}{5}$. That is, whistleblowing is more likely to induce social approval than disapproval, which is plausible in out Visible Externalities treatments. Suppose that our remaining experimental factors affect neither p nor q. Finally, assume that that the moral concern distribution is degenerate with $Prob(\beta_i = 3) = m(3) = 1$, while social approval concerns are non-degenerate, being distributed uniformly on the interval $[0, 9]$.

The net monetary utility from whistleblowing is -5 in our "No Rewards" treatments, while in our "Rewards" treatments, this net monetary utility is $+5$. Given these parameters, in our "No Rewards" treatments social judgment is completely ineffective at increasing whistleblowing. No employee is willing to blow the whistle without social judgment:

$$-5 < 2[\beta_i(1 - 2p) = 2 \times 3 = 6.$$

With social judgment, it is also the case that no employees are willing to blow the whistle. For all $\gamma_i \in [0, 9]$:

$$-5 < 2[3 + \gamma_i(1 - 2q)] = 6 - \frac{3}{5}\gamma_i.$$

In our "Rewards" treatments, however, social judgment *is* effective at inducing whistleblowing. We construct the same calculations as above. The only thing that has changed is the net monetary utility of whistleblowing, which is now 5 rather than -5. Without social judgment, we still have no whistleblowing, since

$$5 < 2[\beta_i(1 - 2p)] = 2 \times 3 \times 1 = 6.$$

With social judgment, however, for a substantial range of γ_i's the net monetary utility of whistleblowing now outweighs the moral and social costs. We can calculate the probability of whistleblowing to be:

$$Prob(5 > 2[3 + \gamma_i(1 - 2 \times \frac{4}{5})]) = 1 - S(\frac{5}{6}) \approx 90.7\%.$$

All together, in our numerical example social judgment is completely ineffective at inducing whistleblowing without monetary rewards. In stark contrast, adding monetary rewards makes social judgment almost completely effective, increasing the population propensity to blow the whistle by over 90 percentage points.

We now turn to a pattern that is not transparently consistent with our simple framework and explain some ways the framework could be extended to incorporate the pattern.

B.2 A puzzling pattern

A pattern in our data that appears at first glance puzzling occurs in the context of no social judgment. When there is no possibility of social judgment, our data suggest that monetary rewards are ineffective at inducing whistleblowing in our Visible Externalities treatments, but quite effective when externalities are not visible. This is puzzling because, since there is no social judgment, the social approval component of utility should not matter, but this is precisely the only component that should be affected by the visibility of externalities. Consequently, to explain this pattern in our framework we would need a plausible story relying on the moral component of overall utility and allow for some type direct effect of our treatments on our framework parameters, or interaction among them, that we have not yet considered. We provide two plausible examples of how our framework could be plausibly extended to incorporate even this puzzling pattern.

One such, hopefully plausible but necessarily ad hoc, story relies on allowing p to vary with the visibility of externalities. Suppose that

$$p(\text{Visible Externalities}) < p(\text{Invisible Externalities}).$$

This could be the case because when it is clear that the employee knows the public knows it is being harmed, choosing to blow the whistle might feel more like choosing sides between the firm and the public than when the public is uninformed. If this were the case, then we could easily construct an example consistent the puzzling pattern. For instance, reconsider the parameters from our numerical example above, except now suppose that:

$$\begin{aligned} p(\text{Visible Externalities}) &= 0, \\ p(\text{Invisible Externalities}) &= \tfrac{1}{6}. \end{aligned}$$

When externalities are visible, the moral cost of whistleblowing is

$$max\{-5, 5\} < 2[\beta_i(1 - 2p(\text{Visible Externalities})] = 6.$$

In words, the net monetary benefit – either 5 or -5 – never outweighs the moral cost of whistleblowing. Consequently, monetary rewards are completely ineffective at inducing whistleblowing.

However, when externalities are not visible:

$$-5 < 2[\beta_i(1 - 2p(\text{Visible Externalities}) = 4 < 5.$$

While no employees blow the whistle without rewards ($-5 < 4$), all employees blow the whistle when there are whistleblower rewards ($5 > 4$). Thus, when there is no possibility of social judgment, rewards

are completely effective in treatments when externalities are invisible but completely ineffective when externalities are visible.

Another plausible story can be constructed if we allow the morality of whistleblowing, p, to depend directly on both rewards and the visibility of externalities. In particular, assume that the monetary rewards generally lower p, which can be a reduced-form way to capture one form of "crowding out." Suppose also that visible externalities generally increase the morality of whistleblowing. Finally suppose that when externalities are not visible, whistleblowing is already maximally immoral, so that $p = 0$ irrespective of rewards. The following values of p capture this story:

$$\begin{aligned} p(\text{Visible Externalities, Rewards}) &= 0, \\ p(\text{Visible Externalities, No Rewards}) &= \tfrac{3}{4}, \\ p(\text{Invisible Externalities, Rewards}) &= 0, \\ p(\text{Invisible Externalities, No Rewards}) &= 0. \end{aligned}$$

Assume β_i is no longer degenerate, but rather is distributed uniformly on $[0, 3]$. Consider first the case with invisible externalities. When there are no rewards, no employee blows the whistle since the monetary benefit of whistleblowing is always negative but the moral cost is always positive. When there are rewards, employees with $\beta_i < \frac{5}{2}$ blow the whistle. This occurs with probability $\frac{5}{6}$. Therefore, the overall effectiveness of rewards when there is no social judgment and when externalities are invisible is increase the probability of whistleblowing by $\frac{5}{6}$.

Contrast this with the visible externalities case. With rewards, $p = 0$ again so that the condition for employee whistleblowing is identical to the one we just calculated – $\beta_i < \frac{5}{2}$. Consequently, with rewards whistleblowing occurs with probability $\frac{5}{6}$. Without rewards, whistleblowing is more likely to be moral than immoral: $p = \frac{3}{4}$. The condition for whistleblowing in this case boils down to $\beta_i < 5$, which occurs with probability one in our example. Consequently, when there is no social judgment but externalities are visible rewards increase the probability of whistleblowing by only $\frac{1}{6}$. In this sense, rewards are less effective at inducing additional whistleblowing when externalities are visible, just as we observe in our data.

APPENDIX TABLES

	Dep. Variable: Dummy equal to 1 if employee is willing to blow the whistle, 0 otherwise					
	Invisible Ext.	Visible Ext.	All	All	All	All
Rewards	1.09***	0.61***	1.09***	1.07***	1.09***	1.01***
	(0.203)	(0.229)	(0.199)	(0.208)	(0.239)	(0.228)
Social Judgment	-0.47**	0.47**	-0.47**	-0.47**	-0.47	-0.54
	(0.214)	(0.230)	(0.210)	(0.208)	(0.329)	(0.360)
Visible Externalities			-0.25	-0.27	0.12	0.05
			(0.304)	(0.314)	(0.249)	(0.267)
Visible x Reward			-0.48	-0.49	-1.16***	-1.07***
			(0.300)	(0.312)	(0.301)	(0.303)
Visible x Social Judgment			0.94***	0.94***	0.29	0.38
			(0.309)	(0.339)	(0.382)	(0.444)
Social Judgment x Reward					-0.01	0.15
					(0.427)	(0.450)
Judgment x Reward x Visible					1.16**	0.99*
					(0.505)	(0.571)
Firm performance/Own performance				-0.58**		-0.57*
				(0.275)		(0.294)
Constant	-0.87***	-1.11***	-0.87***	-0.57	-0.87***	-0.67
	(0.184)	(0.250)	(0.181)	(1.041)	(0.207)	(1.093)
Controls	No	No	No	Yes	No	Yes
Observations	102	104	206	206	206	206

Note: We report estimates from probit regressions. Controls are: gender, economics major, number of firms in the session, ratio between firm performance and own performance in team building task, and effort chosen in minimum effort task. We report the only control variable that is significant: the ratio between firm performance and own performance in the team building task. Robust standard errors, clustered at the session level, in parentheses; *** p<0.01, * * p<0.05, * p<0.1.

Table 5: Probit regressions

	Under Invisible Externalities	Under Visible Externalities
The Effect of Rewards	0.35***	0.20**
	(0.000)	(0.015)
The Effect of Social judgment	-0.16**	0.16*
	(0.023)	(0.078)

Note: We report the marginal effects corresponding to the linear combinations of the estimated coefficients displayed in column 4 of Table 3. p-values in parentheses.

Table 6: Estimated marginal effects

	Under Invisible Externalities		Under Visible Externalities	
	With No Judgment	With Soc. Judgment	With No Judgment	With Soc. Judgment
The Effect of Rewards	0.36***	0.33***	-0.003	0.40**
	(0.000)	(0.007)	(0.959)	(0.017)
	With No Reward	With Reward	With No Reward	With Reward
The Effect of Social judgment	-0.14	-0.16	-0.03	0.37**
	(0.106)	(0.123)	(0.627)	(0.027)

Note: We report the marginal effects corresponding to the linear combinations of the estimated coefficients displayed in column 6 of Table 3. p-values in parentheses.

Table 7: Estimated marginal effects by sub-treatment

	No Judgment & No Reward	No Judgment & Reward	Social Judgment & No Reward	Social Judgment & Reward
Invisible Externalities	15.38%	11.76%	9.09%	0.00%
H_0: Reward = No Reward	p-value = 0.773 (1.000) if if Judgment=0		p-value = 0.329 (1.000) if if Judgment=1	
H_0: No Judgment = Soc. Judgment	p-value =0.642 (1.000) if Rewards=0		p-value = 0.260 (0.516) if Rewards=1	
Visible Externalities	9.09%	0.00%	7.14%	6.67%
H_0: Reward = No Reward	p-value = 0.286 (0.478) if if Judgment=0		p-value = 0.968 (1.000) if if Judgment=1	
H_0: No Judgment = Soc. Judgment	p-value = 0.859 (1.000) if Rewards=0		p-value = 0.362 (1.000) if Rewards=1	

Note: We report the marginal effects corresponding to the linear combinations of the estimated coefficients displayed in column 8 of Table 3. P-values are generated by Chi-square tests. P-values from Fisher exact tests in parentheses. The decline observed when the externalities are visible is also not statistically significant.

Table 8: Manager's law-breaking behavior