[Preliminary Version. Do Not Circulate or Post Online.]

Optimal Law Enforcement with Ordered Leniency*

Claudia M. Landeo† and Kathryn E. Spier‡

March 28, 2018

Abstract

This paper studies the design of enforcement policies to detect and deter harmful short-term activities committed by groups of injurers. With an ordered-leniency policy, the degree of leniency granted to an injurer who self-reports depends on his or her position in the self-reporting queue. By creating a "race to the courthouse," ordered-leniency policies lead to faster detection and stronger deterrence of illegal activities. The socially-optimal level of deterrence can be obtained at zero cost when the externalities associated with the harmful activities are not too high. Without leniency for self-reporting, the enforcement cost is strictly positive and there is underdeterrence of harmful activities relative to the first-best level. Hence, ordered-leniency policies are welfare improving. Our findings for environments with groups of injurers complement Kaplow and Shavell's (1994) results for single-injurer environments. Experimental evidence provides support for our theory.

KEYWORDS: Law Enforcement; Leniency; Self-Reporting; Ordered Leniency; Harmful Externalities; White-Collar Crime; Securities Fraud; Insider Trading; Market Manipulation; Whistleblowers; Non-Cooperative Games; Prisoners' Dilemma; Coordination Games; Risk Dominance; Pareto Dominance; Experiments

JEL Categories: C72, C90, D86, K10, L23

*We acknowledge financial support from the National Science Foundation (NSF Grant SES-1155761). We thank Susan Norton for administrative assistance, Tim Yuan for programming the software used in the experimental section of this study, and the Harvard Decision Science Laboratory for the assistance in conducting the experimental sessions.

†University of Alberta, Department of Economics. Henry Marshall Tory Building 7-25, Edmonton, AB T6G 2H4. Canada. landeo@ualberta.ca, tel. 780-492-2553.

‡Harvard Law School and NBER. 1575 Massachusetts Ave., Cambridge, MA 02138. kspier@law.harvard.edu, tel. 617-496-0019.

1 Introduction

Illegal activities are often committed by groups of people working together rather than by individuals working alone. Common examples in the corporate setting include insider trading and market manipulation schemes. In 2011, the FBI reported 726 corporate fraud cases, several of which involved losses to public investors that individually exceeded $1 billion, and 343 securities fraud cases involving more than 120,000 victims and approximately $8 billion in losses (FBI, 2012). More generally, illegal activities committed by groups of wrongdoers impose considerable costs on society. To combat illegal group activities, law enforcement agencies often grant leniency to wrongdoers who come forward and self-report.

In a typical leniency program, wrongdoers who self-report early face lower sanctions than those who self-report later.[1] For instance, in 2014, the Securities and Exchange Commission (SEC) brought insider trading charges against Christopher Saridakis, a top executive at GSI Commerce, and several co-conspirators for providing tips to family and friends in advance of eBay's acquisition of GSI. Saridakis paid a penalty equal to twice the amount of his tippees' profits,[2] and was imprisoned after pleading guilty to criminal charges. One of Saridakis' co-conspirators was forced to disgorge his own profits and paid a penalty equal to three times his own profits and all of the profits of his own tippees. In contrast, a co-conspirator who aided the prosecution paid a reduced penalty equal to one half of his profits, while another co-conspirator who cooperated early paid no penalty at all (Ceresney, 2015).[3]

This paper studies the design of enforcement policies to detect and deter illegal short-term activities committed by groups of injurers.[4] We focus on a class of mechanisms where the amount of leniency granted to an injurer depends on his or her position in a self-reporting queue. The earlier an injurer reports the act, the higher his or her position in the self-reporting queue. We call these mechanisms "ordered-leniency policies." Ordered-leniency policies that give greater leniency

[1] An example of such a program is the Securities and Exchange Commission's Cooperation Program.

[2] In insider trading cases, the term "tipper" refers to a person who has broken his fiduciary duty by revealing inside information. The term "tippee" refers to a person who knowingly uses inside information to make a trade.

[3] See also *SEC v. Saridakis and Gardner,* Civil Action No. 14 2397 (U.S. District Court Eastern District of Pennsylvania 2014). For another interesting insider-trading case involving leniency for early cooperation, see *SEC v. Wrangell* (2012), https://www.sec.gov/litigation/complaints/2012/comp-pr2012-193-wrangell.pdf.

[4] Illegal *short-term* activities do not involve an ongoing relationship among group members. They are sometimes referred to as illegal "occasional" activities. See Buccirossi and Spagnolo (2006). In game-theoretic terms, they correspond to one-shot strategic environments. Leniency programs have been also applied to illegal *long-term* activities such as cartels. For a recent survey of this literature, see Spagnolo and Marvão (2016).

to early cooperators create a so-called "race to the courthouse."[5] This leads to better detection and to stronger deterrence. Although our paper is motivated by insider trading and securities fraud, our analysis applies to any kind of harmful short-term activity committed by a group of wrongdoers. To the best of our knowledge, there are no previous theoretical or experimental analyses of ordered-leniency policies for short-term group activities.

In our model, the enforcement agency commits to an enforcement policy involving investigation efforts, a sanction, and a degree of leniency for injurers who self-report. Next, given the enforcement policy, the potential injurers decide whether to participate in a harmful group act. If the act is committed, then the injurers decide *whether* and *when* to report themselves to the authorities. The decision of an injurer to self-report hinges on the likelihood of detection if he remains silent, which itself depends on both the enforcement efforts of the agency and the self-reporting decision of the other injurer. Specifically, the likelihood that an injurer will be detected and sanctioned is assumed to be increasing in the number of injurers who self-report and in the enforcement efforts of the agency. Hence, negative externalities are present in the self-reporting stage. Depending on the detection probabilities and the degree of leniency, the self-reporting stage might resemble a prisoners' dilemma game or a coordination game.

We demonstrate that the ordered-leniency policy that creates maximal deterrence imposes the highest possible sanction on injurers who fail to self-report but are caught nonetheless, and discounts the sanction for the first injurer to self-report. Depending on the parameter values, the second injurer to self-report may receive lenient treatment as well (albeit to a lesser degree). Granting leniency to the second injurer is particularly valuable when the inculpatory evidence provided by the first injurer alone is insufficient to convict the second injurer with certainty. Granting leniency only to the first injurer to self-report or to both the first and second injurer creates a race to the courthouse where both injurers promptly report the act. This strengthens deterrence.

The degree of leniency for those who self-report depends critically on the equilibrium refinement that applies in the self-reporting stage. With the Pareto-dominance refinement, the optimal leniency policy is strong in the sense that it grants larger discounts to injurers who self-report. Strong leniency creates a prisoners' dilemma strategic environment where self-reporting is a dominant strategy. With the risk-dominance refinement (Harsanyi and Selten, 1988), the optimal

[5]The expression "race to the courthouse" typically refers to the first-to-file legal rule that provides superior rights to the first action filed in civil litigation cases. In our environment, earlier reporting raises the chances of being the first in the self-reporting queue.

leniency policy is mild in the sense that it gives smaller discounts to injurers who self-report. With mild leniency, the self-reporting stage is a coordination game. Although self-reporting is not a dominant strategy for the injurers, a mild leniency policy can in theory be highly effective if the injurers are sufficiently distrustful of each other. We show that ordered-leniency policies generate a race-to-the-courthouse effect where, in equilibrium, self-reporting occurs immediately.

Importantly, we show that ordered-leniency policies are welfare improving whenever the set of possible fines is bounded from above. Without leniency for self-reporting, for any bounded set of fines, the enforcement agency's efforts must be strictly positive and there will be underdeterrence of harmful activities relative to the first-best level. Holding the fine and the costs of enforcement fixed, an ordered-leniency policy will increase the expected fine, thus raising level of deterrence and increasing social welfare. Indeed, our analysis demonstrates that the socially-optimal level of deterrence can be obtained at zero cost when the externalities associated with the harmful activities are not too high.

We provide experimental evidence regarding the effects of ordered-leniency policies. Given that there are multiple equilibria in the self-reporting stage, and the optimal ordered-leniency policy depends on the equilibrium refinement applied, it is appropriate to use experimental economics methods. Three leniency environments are studied: No Leniency, where no penalty reductions are granted; Strong Leniency, where the first to report receives a strong reduction in the penalty; and Mild Leniency, where the first to report receives a mild reduction in the penalty. Our findings suggest that ordered-leniency policies are effective detection mechanisms. Importantly, we provide empirical evidence of a "race-to-the-courthouse" effect of ordered-leniency policies. In particular, our results indicate that the implementation of either Strong or Mild ordered-leniency policies increases the likelihood of self-reporting by one or both injurers. Our findings under Mild Leniency suggest that the parties' behaviors are aligned with the risk-dominance refinement. Our experimental results also indicate that some subjects systematically underestimate the likelihood and severity of sanctions when making their decisions about committing the harmful act. These findings might suggest the presence of self-serving bias on subjects' beliefs about getting the first position in the self-reporting queue. As a result, the deterrence power of ordered-leniency policies are weakened, and harmful acts are committed more frequently than predicted.

Finally, we explore several extensions. First, we extend our theoretical framework to groups of injurers with multiple members. Attention is restricted to coalition-proof Nash equilibria (Bernheim et al., 1987). We show that the highest level of deterrence is achieved when all injurers who commit the act later self report, and receive successive discounts for self-reporting based on their

positions in the self-reporting queue. In general, the leniency for the first injurer to report may not be full, and the leniency for the last injurer to report may not be zero. We demonstrate that the race-to-the-courthouse effect is robust to the number of members in the group of injurers.

Our paper contributes to the literature on the control of harmful externalities by presenting the first formal analysis of optimal enforcement policies with ordered leniency for harmful short-term activities conducted by a group of wrongdoers, and by offering experimental evidence of the effectiveness of ordered-leniency policies as detection mechanisms.[6] Our work is related to several strands of literature. The closest to our work are the studies on enforcement and self-reporting.

Kaplow and Shavell (1994) study self-reporting in a model of probabilistic law enforcement. In the context of a single injurer, they demonstrate that self-reporting of harmful acts can lower the ex post costs of investigation without compromising deterrence. This can be accomplished by allowing those who self-report to pay a sanction equal to or slightly less than the expected sanction they would face if they did not report the act. Given that investigatory efforts do not need to be allocated to identify the injurers who self-report, the enforcement agency can economize on its ex post enforcement efforts.[7] In contrast, in our model, the costs of enforcement are sunk ex ante, before the act is committed and before the injurers have the opportunity to self-report. Holding the costs of enforcement fixed, ordered-leniency policies create a race-to-the-courthouse among multiple injurers. This raises the expected fine and strengthens deterrence. Thus, our findings for groups of injurers complement Kaplow and Shavell's (1994) results for single-injurer environments.

Feess and Walzl (2004) study optimal enforcement with self-reporting for illegal activities committed by two-member criminal teams. The focus of their paper are the consequences of cooperation in the self-reporting stage on enforcement. They show that maximal deterrence can be reached at virtually no cost when injurers decide non-cooperatively whether to self-report or decide cooperatively with an exogenous probability of cooperation. Our analysis differs in several respects. First, the mechanism studied by Feess and Walzl (2004) grants leniency only when *exactly one* injurer self-reports. If both injurers report, then neither receives any sanction

[6] In seminal work, Becker (1968) demonstrates that a very small probability of detection coupled with a very high sanction can deter crime at essentially zero cost. Polinsky and Shavell (1984) show that when injurers have limited assets and sanctions are bounded above, then the optimal enforcement policy involves investigation costs and deterrence falls short of the first-best level.

[7] Self-reporting is also socially valuable because early detection of harmful activities might minimize further social costs (Malik, 1993; Innes, 1999). Self-reporting has also been studied in the context of pollution (Livernois and McKenna, 1999) and tax evasion (Andreoni, 1991; and Malik and Schwab, 1991).

reduction. Then, ordered-leniency policies are not studied and hence, a race-to-the-courthouse effect cannot be elicited in their environment. Second, the authors assume that the Pareto-dominance refinement applies in case of multiplicity of equilibria, and hence, leniency policies under the risk-dominance refinement are not investigated.[8] Buccirossi and Spagnolo (2006) study the effects of leniency policies on sequential bilateral short-term illegal activities. They find that moderate leniency policies (i.e., policies involving a reduction in the sanction but not a reward for self-reporting) might have the perverse effect of providing an effective governance mechanism for illegal short-term activities that otherwise will not be implemented due to a hold-up problem. Optimal enforcement policies are not studied.

Another strand of literature related to our paper is that on plea bargaining, where an individual has the option to plead guilty in exchange for a reduced sentence. In models with a single defendant, Landes (1971) demonstrates that plea bargaining agreements reduce prosecutorial costs and Grossman and Katz (1983) find that plea bargaining might produce insurance and screening effects.[9] Kobayashi (1992) studies plea bargaining using a model with two defendants where the acceptance of a plea agreement by one defendant raises the probability of conviction of the other, the probability of conviction of the more culpable defendant is higher than the probability of conviction of the less culpable defendant, and the identities of the defendants are known by the prosecutor. He finds that the plea bargaining policy that maximizes deterrence involves a lower penalty for the most culpable defendant. None of these papers consider ordered-leniency policies.[10]

Our paper is also related to the literature on enforcement of competition policy and leniency programs for illegal long-term activities committed by criminal groups. Motta and Polo (2003) find that, when the enforcement authority has limited resources and hence, is unable to prevent collusion ex-ante, leniency policies enhance welfare by increasing the likelihood of cartel cessation and shortening investigation. Spagnolo (2005) demonstrates that leniency policies undermine internal trust by increasing individual incentives to defect. As a result, these policies destabilize cartels. Optimal leniency policies reward the first party to report with the fines paid by all other parties. When fines and rewards are sufficiently high, the first best is obtained at a zero cost.[11] Bigoni et al. (2012) provide experimental evidence of the effects of leniency and rewards

[8] In addition, Feess and Walzl's (2004) social welfare analysis focuses on minimizing harm to victims and does not include the injurers' private benefits, and environments with multiple injurers are not investigated.

[9] Negative effects might occur if innocent defendants are more risk-averse than guilty defendants, and innocent defendants might be induced to plead guilty. See also Reinganum (1988).

[10] See also Kraakman (1986) and Arlen and Kraakman (1997) for seminal work on third-party enforcement.

[11] Chen and Rey (2013) extend Spagnolo (2005) by considering not only pre-investigation leniency but also post-

on enforcement and the stability of collusion in repeated-game environments. They find that leniency enhances deterrence but contributes to the stabilization of surviving cartels. Prices fall to the competitive levels when rewards are provided to whistleblowers.

Our work shares some features with studies on contract design in the presence of externalities among contract recipients. In the context of exclusionary vertical restraints, Rasmusen et al. (1991) and Segal and Whinston (2000) demonstrate that, when there are economies of scale in production, incumbent monopolists can design profitable exclusive-dealing contracts by exploiting the negative externalities among the buyers. Landeo and Spier (2009, 2012) provide experimental evidence of the exclusionary power of these types of contracts.[12]

The rest of the paper is organized as follows. Section 2 introduces the model setup. Section 3 presents the equilibrium analysis of the injurers' decisions about committing the act, self-reporting, and the time to report. Section 4 constructs the optimal enforcement policies with and without leniency for self-reporting. We show that the optimal ordered-leniency policy always creates superior incentives, and we identify necessary and sufficient conditions for achieving the first-best outcome. Section 5 presents experimental evidence of the effects of ordered-leniency policies. Section 6 extends our benchmark model to groups of injurers with multiple members, stochastic detection rates, and asymmetric benefits from committing a harmful act across injurers, and demonstrates that the main insights derived from our benchmark model and their implications for the design of optimal enforcement policies are robust. Section 7 concludes. Formal proofs are presented in the Appendix.

2 Model Setup

Our framework involves three risk-neutral players: Two identical representative potential injurers and an enforcement agency.[13] We assume that the potential injurers seek to maximize their private benefits from committing a harmful act. The enforcement agency seeks to maximize social

investigation leniency. Harrington (2013) investigates the incentives to apply for leniency when each cartel member has private information about the likelihood of conviction without self-reporting and leniency is granted only to the first cartel member to self-report. Feess and Walzl (2010) study leniency policies when one cartel member might provide stronger evidence than the other member. See Livernois and McKenna (1999) for a model of pollution regulation and self-reporting in a repeated-game environment.

[12]See Landeo and Spier (2015) and Che and Yoo (2001) for applications to incentive contracts for teams, and Kornhauser and Revesz (1994) and Spier (1994) for applications to civil litigation under joint and several liability.

[13]Later, we consider groups with more than two members.

welfare. Social welfare includes the aggregation of the benefits to the injurers. It also includes the social costs: The harm inflicted on others (externalities associated with the harmful activities) and the cost of enforcement. We assume that the enforcement agency cannot costlessly identify the parties responsible for committing the harmful act, and that the set of fines or monetary sanctions is bounded. Without loss of generality, we abstract from time discounting.

The timing of the game is as follows. First, the enforcement agency publicly commits to an enforcement policy with ordered leniency to detect and prevent harmful short-term activities committed by groups of injurers. The enforcement policy components are (f, r_1, r_2, e). (1) $f \in (0, \bar{f}]$ denotes a fine or monetary sanction (measured per injurer).[14] The maximal fine, $\bar{f}$, can be greater than, lower than, or equal to the harm inflicted on others (measured per injurer), h.[15] (2) $r_1, r_2 \in [0, 1]$ denote the leniency multipliers that correspond to the first and second positions in the self-reporting queue, respectively.[16] The discount for position i in the reporting queue is then $1 - r_i$, $i = 1, 2$. Thus, we study *ordered-leniency policies* where the first injurer to report pays $r_1 f$, regardless of whether a second injurer reports, and the second injurer to report pays $r_2 f$. (3) $e \in [0, 1)$ denotes the enforcement agency's effort (investigation effort), which, as we will describe below, determines the probability that harmful acts are detected. We let $c(e)$ be the cost of enforcement or investigation (measured per injurer), and assume that $c(0) = 0$, $c'(0) = 0$, $c'(e) \geq 0$, $c''(e) > 0$, and $lim_{e \to 1} c'(e) = \infty$.[17]

Second, after observing the enforcement policy, the potential injurers play a two-stage game. In Stage 1, the potential injurers simultaneously and independently decide whether to participate in a harmful activity. The act is committed if and only if both injurers decide to participate. The benefit for each injurer is $b \in [0, \infty)$, which is distributed according to probability density function $g(b)$ and cumulative distribution function $G(b)$, common knowledge. The realization of b is revealed to both potential injurers before they make their decisions regarding committing the act.[18] If the act is committed, Stage 2 starts; otherwise, the game ends. In Stage 2, the injurers simultaneously and independently decide *whether* and *when* to report the harmful act to the enforcement agency. Specifically, each injurer can choose to report the act at any time t in an interval $[0, 1]$ where $t = 0$ represents immediate reporting and $t > 0$ represents delayed reporting.

Third, the injurers (parties responsible for causing harm), if detected, are accurately identified

[14] $\bar{f}$ can be interpreted as the potential injurer's wealth. When the fine is above $\bar{f}$, the injurer is judgment-proof.
[15] In contrast, Kaplow and Shavell (1994) assume that the maximal fine is greater than or equal to the harm.
[16] Multipliers $(r_1, r_2) = (1, 1)$ imply that the enforcement policy does not grant leniency for self-reporting.
[17] These assumptions ensure an interior solution for the social welfare maximization problem.
[18] Note that committing the act is socially desirable if and only if the benefits, b, exceed the social harm, h.

by the enforcement agency and sanctioned. The probabilities of detection and the sanctions are as follows. Absent any self-reporting by the injurers, harmful acts are detected with probability p_0 and each injurer pays a fine f. If one injurer reports the act, then the injurer who reports pays fine $r_1 f$ and the silent accomplice is accurately detected and fully sanctioned (i.e., pays a fine f) with probability p_1. If both injurers report the act, then the first to report pays fine $r_1 f$ and the second to report pays fine $r_2 f$. If the two injurers report at exactly the same time, then an equally-weighted coin flip determines who obtains the first and second positions in the self-reporting queue. Finally, we assume that p_0 and p_1 depend on the enforcement agency's effort, $e \in [0, 1)$, and p_1 also depends on the exogenous strength of inculpatory evidence, $\pi \in (0, 1)$. Specifically, $p_0(e) = e$ and $p_1(e, \pi) = e + (1 - e)\pi$.[19] Then, $0 \leq p_0(e) < p_1(e, \pi) < 1$.

The equilibrium concept is subgame-perfect Nash equilibrium. Our focus is on pure-strategy equilibria that survive the elimination of weakly-dominated strategies. When multiple pure-strategy equilibria arise, we present separate equilibrium analyses for the Pareto-dominance and risk-dominance refinements (Harsanyi and Selten, 1988).

The first-best outcome is used as a benchmark in the welfare analysis of ordered-leniency policies. The first best is defined as the social welfare outcome of an environment in which the enforcement agency can costlessly identify the parties responsible for committing the harmful act (and their private benefits) and decide which acts to prohibit. Then, in the first-best outcome, the cost of effort is zero and acts are committed if and only if $b > h$.[20]

We proceed backwards and begin our analysis with the injurers' decisions. We then analyze the optimal enforcement policy with ordered leniency and conduct social welfare analysis.

3 Injurers' Decisions: Equilibrium Characterization

We begin by characterizing the equilibrium behavior of the injurers in Stage 2, the self-reporting stage. Next, we study the potential injurers' decisions regarding committing the act in Stage 1.

[19] This specification may be derived from first principles. Suppose that absent self reporting by either injurer, detection is the outcome of a single Bernoulli trial with success probability $p_0 = e$. When one injurer self reports and another does not, there is a second independent Bernoulli trial that succeeds in detecting the non-reporting injurer with probability π. Then $p_1 = e + (1 - e)\pi$ is the probability that the silent injurer is detected.

[20] In practice, of course, the enforcement agency cannot costly identify the injurers. Hence, to detect and deter harmful acts, the enforcement agency needs to spend resources on detection and implement leniency programs for self-reporting.

3.1 Decision to Report the Act and Time to Report

Recall that when both potential injurers decide to participate in the harmful act in Stage 1, the act is committed and Stage 2 occurs. In Stage 2, the injurers simultaneously and independently decide *whether* and *when* to report the harmful act to the enforcement authority. Specifically, an injurer who decides to report the act also needs to choose the time of his or her report, $t \in [0, 1]$.

We first analyze the length of time taken by the injurers to report the harmful act. The analysis presented here is general in the sense that it allows r_1 to be greater than, equal to, or lower than r_2. In later sections, we verify that optimal enforcement policies with ordered leniency require $r_1 > r_2$. Lemma 1 characterizes the equilibrium report time.

Lemma 1: *If $r_1 < r_2$, then an injurer who reports the act will do so immediately, $t = 0$. If $r_1 > r_2$, then an injurer who reports the act will delay reporting until the last moment, $t = 1$. If $r_1 = r_2$, then an injurer who reports the act may do so at any time, $t \in [0, 1]$.*

Lemma 1 follows from the elimination of weakly-dominated strategies. Suppose that $r_1 < r_2$, so the first injurer to report the act receives a larger penalty reduction than the second injurer to report. Intuitively, $r_1 < r_2$ generates an incentive to minimize the time to report in order to secure the first position in the self-reporting queue, i.e., "a race to the courthouse." If injurer $k(k = 1, 2)$ believes that injurer $j(j = 1, 2, j \neq k)$ will not report at all, then injurer i is just as well off reporting immediately as delaying until some later time. However, if injurer k believes that there is a non-zero chance that injurer j will report at time $t = 0$, then injurer k is strictly better off reporting immediately as well. In other words, late reporting is a weakly-dominated strategy. If instead $r_1 > r_2$, then the second injurer to report receives a larger penalty reduction than the second injurer to report. In this case, early reporting is a weakly-dominated strategy.[21] Importantly, Lemma 1 implies that if both injurers report the harmful act, and if $r_1 \neq r_2$, then both injurers are equally likely to get the first position or the second position in the self-reporting queue.[22]

Second, we study the injurers' decisions about whether to report the act. The strategic-form

[21] If an injurer believes that there is a non-zero chance that the other injurer will report the act at $t = 1$, then the injurer strictly prefers to wait until $t = 1$ to report as well. If $r_1 = r_2$, then there is no advantage to being first or second, and the injurers are indifferent about the reporting time.

[22] When $r_1 < r_2$, self-reporting occurs immediately at $t = 0$, and when $r_1 > r_2$ self-reporting occurs at $t = 1$. By assumption, when the two injurers report at exactly the same time, an equally-weighted coin flip determines who obtains the first position in the self-reporting queue.

Figure 1: Strategic-Form Representation of the Self-Reporting Subgame (Expected Payoffs)

	No Report (NR)	Report (R)
No Report (NR)	$b - p_0 f$, $b - p_0 f$	$b - p_1 f$, $b - r_1 f$
Report (R)	$b - r_1 f$, $b - p_1 f$	$b - \left(\frac{r_1+r_2}{2}\right) f$, $b - \left(\frac{r_1+r_2}{2}\right) f$

representation of the self-reporting subgame is presented in Figure 1. If neither injurer self-reports, then the act is detected with probability p_0 and each injurer receives a payoff of $b - p_0 f$. If one injurer self-reports but the other does not, then the injurer who self-reports pays $r_1 f$ with certainty and the silent accomplice pays $p_1 f$ in expectation giving payoffs $b - r_1 f$ and $b - p_1 f$, respectively. Finally, if both injurers self-report, then they are equally likely to get the first and second positions in the self-reporting queue. So, each injurer receives an expected payoff of $b - \left(\frac{r_1+r_2}{2}\right) f$.[23] Lemma 2 characterizes the pure-strategy Nash equilibria of the self-reporting subgame.

Lemma 2. *Take the benefit b, the fine f, and the detection probabilities, p_0 and p_1, as fixed. The pure-strategy Nash equilibria of the self-reporting subgame are as follows.*

1. *$r_1 \leq p_0$ and $\frac{r_1+r_2}{2} \leq p_1$: There is a unique pure-strategy Nash equilibrium where both injurers self-report, (R, R).*

2. *$r_1 \leq p_0$ and $\frac{r_1+r_2}{2} > p_1$: There are two pure-strategy Nash equilibria where exactly one injurer self-reports, (R, NR) and (NR, R).*

3. *$r_1 > p_0$ and $\frac{r_1+r_2}{2} \leq p_1$: There are two pure-strategy Nash equilibria, one where both injurers self-report and one where neither injurer self-reports. (R, R) Pareto dominates (NR, NR) if and only if $\frac{r_1+r_2}{2} \leq p_0$. (R, R) risk dominates (NR, NR) if and only if $\frac{3r_1+r_2}{4} \leq \frac{p_0+p_1}{2}$.*

4. *$r_1 > p_0$ and $\frac{r_1+r_2}{2} > p_1$: There is a unique pure-strategy Nash equilibrium where neither injurer self-reports, (NR, NR).*

In Case 1 of Lemma 2, self-reporting is a weakly-dominant strategy for both injurers. So, (R, R) is the unique Nash equilibrium that survives the elimination of weakly-dominated strategies.[24] When the expected sanction for self-reporting is not too small, $\left(\frac{r_1+r_2}{2}\right) f > p_0 f$, then the injurers

[23] If $r_1 = r_2$, different reporting times would lead to the same expected payoffs.

[24] More specifically, the second Nash equilibrium where both injurers decide not to report, (NR, NR) does not survive the elimination of weakly-dominated strategies.

are jointly worse off self-reporting than they are remaining silent and the self-reporting subgame resembles a prisoners' dilemma environment.[25] In Case 2, there are two pure-strategy Nash equilibria, (R, NR) and (NR, R), where exactly one injurer reports the act and the other does not.[26] In Case 3, both (NR, NR) and (R, R) are Nash equilibria. If one injurer believes that the other will remain silent then he will remain silent as well, since the expected penalty associated with remaining silent, $p_0 f$, is smaller than the penalty from being the only injurer to report, $r_1 f$. But if he believes that the other injurer will report, then he is better off reporting too since paying $\left(\frac{r_1+r_2}{2}\right)f$ on average is better than paying $p_1 f$. Thus, the self-reporting subgame in Case 3 is a coordination game. Finally, in Case 4, no-reporting is a strictly-dominant strategy for both injurers. So, (NR, NR) is the unique Nash equilibrium.

The set of Nash equilibria associated with Case 2, (R, NR) and (NR, R), cannot be narrowed with either the Pareto-dominance or the risk-dominance refinements (Harsanyi and Selten, 1988). In contrast, the two pure-strategy Nash equilibria that arise in Case 3, (R, R) and (NR, NR), may be ranked using standard equilibrium refinements. When $\frac{r_1+r_2}{2} \leq p_0$, the expected sanction is lower when both injurers report committing the act. So, (R, R) is the Pareto-dominant Nash equilibrium if and only if $\frac{r_1+r_2}{2} \leq p_0$. When $\frac{3r_1+r_2}{4} \leq \frac{p_0+p_1}{2}$, an injurer would prefer to self-report when there is a fifty-percent chance that the other injurer will also report. Thus, (R, R) is the risk-dominant Nash equilibrium if and only if $\frac{3r_1+r_2}{4} \leq \frac{p_0+p_1}{2}$.

3.2 Decision to Commit the Act

In Stage 1, the potential injurers simultaneously and independently decide whether to participate in the harmful activity.[27] If *both* potential injurers decide to participate in the activity, then the act is committed. The payoff for each injurer is equal to the payoff that corresponds to the Nash equilibrium of the self-reporting subgame shown in Figure 1. If one or both potential injurers decide not to participate, then the act is not committed. The game ends and the payoff for each potential injurer is zero.

A potential injurer's decision about whether to participate in the harmful activity in Stage 1 depends on his private benefit from committing the act and the expected fine (which is determined

[25] If $\left(\frac{r_1+r_2}{2}\right)f < p_0 f$, self-reporting is jointly efficient for the injurers and the game is not a prisoners' dilemma.

[26] Without loss of generality, we assume that, when indifferent, the injurers decide to self-report. This assumption allows us to eliminate the potential Nash equilibrium where both injurers decide not to report, (NR, NR).

[27] Our findings also hold in environments in which the injurers jointly decide whether to commit an act, but binding agreements between the injurers regarding their reporting choices in Stage 2 are not allowed.

in the Stage 2 continuation game). The b-value that equals the expected fine represents the "deterrence threshold" and is denoted by $\hat{b}$. When the individual benefit of committing the act, b, is greater than the deterrence threshold, $\hat{b}$, then participating in the activity is a weakly-dominant strategy. In particular, if injurer k ($k = 1, 2$) believes that injurer j ($j = 1, 2$, $j \neq k$) will participate in the activity with non-zero probability, then injurer k strictly prefers to participate in the act.[28] Conversely, when b is smaller than the deterrence threshold, $\hat{b}$, the injurer will choose not to participate in the activity.[29] Finally, when b is exactly equal to the deterrence threshold, $\hat{b}$, then the injurer is indifferent between participating and not participating in the act and, without loss of generality, we assume that the injurer does not participate in the act. The deterrence thresholds are constructed using Lemma 2 above. Lemma 3 characterizes the equilibrium decisions in Stage 1. Cases 1–4 correspond to Cases 1–4 included in Lemma 2.

Lemma 3. *Take the fine f, and the detection probabilities, p_0 and p_1, as fixed. Each potential injurer will decide to participate in the activity under the following conditions.*

1. *$r_1 \leq p_0$ and $\frac{r_1+r_2}{2} \leq p_1$: The injurer decides to participate if and only if $b > \hat{b} = \left(\frac{r_1+r_2}{2}\right)f$.*
2. *$r_1 \leq p_0$ and $\frac{r_1+r_2}{2} > p_1$: The injurer decides to participate if and only if $b > \hat{b} = \left(\frac{r_1+p_1}{2}\right)f$.*
3. *$r_1 > p_0$ and $\frac{r_1+r_2}{2} \leq p_1$: If $\frac{r_1+r_2}{2} \leq p_0$ (Pareto Dominance) or $\frac{3r_1+r_2}{4} \leq \frac{p_0+p_1}{2}$ (Risk Dominance), the injurer decides to participate if and only if $b > \hat{b} = \left(\frac{r_1+r_2}{2}\right)f$. If $\frac{r_1+r_2}{2} > p_0$ (Pareto Dominance) or $\frac{3r_1+r_2}{4} > \frac{p_0+p_1}{2}$ (Risk Dominance), the injurer decides to participate if and only if $b > \hat{b} = p_0 f$.*
4. *$r_1 > p_0$ and $\frac{r_1+r_2}{2} > p_1$: The injurer decides to participate if and only if $b > \hat{b} = p_0 f$.*

In Case 1, since both injurers self-report in the unique Nash equilibrium, the deterrence threshold is $\hat{b} = \left(\frac{r_1+r_2}{2}\right)f$. Then, each potential injurer will participate in the harmful act when $b > \hat{b} = \left(\frac{r_1+r_2}{2}\right)f$. In Case 2, where multiple equilibria arise, (R, NR) and (NR, R), our refinements do not eliminate either one and we assume that the deterrence threshold is the average fine, $\hat{b} = \left(\frac{r_1+p_1}{2}\right)f$.[30] Then, each potential injurer will participate in the harmful activity when $b > \hat{b} = \left(\frac{r_1+p_1}{2}\right)f$. In Case 3, the equilibrium refinement will determine which of the two outcomes

[28] When b is greater than the deterrence threshold, then not participating is a weakly-dominated strategy.

[29] Participating is a weakly-dominated strategy in this scenario. If injurer k believes that there is a non-zero chance that injurer j will participate in the act, then injurer k strictly prefers not to participate.

[30] Given that neither the Pareto-dominance nor risk-dominance refinements reduce the set of equilibrium outcomes, it is reasonable to assume that neither the enforcement agency nor the players themselves can predict which

is obtained, (R, R) and (NR, NR), and so the deterrence threshold is either $\hat{b} = \left(\frac{r_1+r_2}{2}\right)f$ or $\hat{b} = p_0 f$. Then, each potential injurer will participate when $b > \hat{b} = \left(\frac{r_1+r_2}{2}\right)f$ or $b > \hat{b} = p_0 f$, depending on the equilibrium. Finally, in Case 4, since neither injurer self-reports in equilibrium, the deterrence threshold is $\hat{b} = p_0 f$. Then, each injurer will participate when $b > \hat{b} = p_0 f$.

Our results suggest that ordered-leniency policies have the potential to create significant social-welfare benefits. Without any opportunities to self-report, the expected fine for each injurer would be capped at $p_0\bar{f}$. Through a leniency program that grants a reduced fine to the first injurer to report the harmful act, $r_1 = p_0 - \varepsilon$ for example, the enforcement agency can induce at least one of the two injurers to come forward and report the act. When one injurer self-reports, the likelihood of catching the silent accomplice rises from p_0 to p_1. With a well-designed enforcement policy with ordered leniency, the enforcement agency can exploit negative externalities between the injurers in the self-reporting subgame to deter a broader range of harmful acts and to economize on enforcement efforts.

4 Optimal Enforcement Policies

This section characterizes the optimal enforcement policies with and without leniency. First, we identify the optimal enforcement policy in the absence of leniency for self-reporting and show that it involves positive enforcement cots, maximal fines, and underdeterrence relative to the first-best level. Second, we consider enforcement policies with ordered leniency. We prove that policies that offer leniency for self-reporting are superior to the optimal enforcement policy without leniency. Holding the enforcement costs fixed, deterrence can be improved with ordered leniency for self-reporting. We then highlight several key features of optimal ordered-leniency policies. Finally, we demonstrate that the first-best outcome can be achieved with an ordered-leniency policy when the externality from the harmful activities, h, is not too high.

outcome will occur, and hence, they assign an equal weight to each outcome. This assumption is intuitive and empirically relevant but much stronger than necessary. All that is required for the results that follow is that the deterrence threshold in Case 2 is strictly smaller than $p_1 f$. This would be true if the players, at the time that they are committing the act, put a non-zero chance on both (R, NR) and (NR, R). We will see that the enforcement agency can implement a deterrence threshold of $p_1 f$ in a setting where self-reporting is a dominant strategy for both players as in Case 1. Thus, for several reasons, the enforcement agency would eschew enforcement policies associated with Case 2.

4.1 Optimal Enforcement Policy without Leniency

Consider an environment where leniency for self-reporting is not granted, so the leniency multipliers are $(r_1, r_2) = (1, 1)$.[31] According to Lemma 2 (Case 4), there is a unique pure-strategy Nash equilibrium where neither injurer self-reports.[32] The probability that the injurers are detected and fined is $p_0 = e$. Then, each injurer faces an expected fine ef, and so each will commit the act if and only if $b > \hat{b} = ef$ (Lemma 3, Case 4). Social welfare is the aggregation of the benefits to the individuals who commit the act minus the social costs associated with the act (the harm inflicted on others, h, and the cost of enforcement $c(e)$).[33] Normalizing the size of the population of injurers to unity, the social welfare function can be written as:

$$W = \int_{ef}^{\infty} (b - h)\, g(b) db - c(e). \tag{1}$$

Next, we identify the optimal fine, f, and the optimal detection probability (optimal enforcement effort), e,[34] that maximize social welfare in the no-leniency environment. Consider first the optimal fine f. It is easy to show that the optimal fine will be maximal, $f = \bar{f}$. To see why, suppose that the optimal $e > 0$ and that the optimal fine is less than maximal, $f < \bar{f}$. By raising the fine slightly while at the same time lowering the probability of detection so as to keep the product ef constant, the same level of deterrence can achieved but at a lower cost than $c(e)$.

Consider now the optimal detection probability (optimal enforcement effort), e. Substitute $\bar{f}$ into the social welfare function and differentiate it with respect to e. The first-order condition is given by:

$$(h - e\bar{f})\bar{f} g(e\bar{f}) - c'(e) = 0. \tag{2}$$

The first term is the incremental social benefit of increased deterrence. When the probability of detection is raised, the acts that were previously exactly on the margin between committing and not committing the act (those with private benefits $b = e\bar{f}$) are now deterred. The social benefit of deterring these marginal acts is $h - e\bar{f}$.[35] The volume of additional cases that are deterred when e is raised is $\bar{f} g(e\bar{f})$, which depends upon the height of the probability density function

[31] The environment without leniency is a special case of enforcement with ordered leniency for self-reporting.

[32] If an injurer reports, he pays f (irrespective of the decision of his accomplice; if an injurer remains silent, he pays $p_0 f$ (if his accomplice does not report the act) or $p_1 f$ (if his accomplice reports the act). Then, self-reporting is a strictly-dominated strategy.

[33] The fines are simply transfers from the injurers to the enforcement agency, and therefore are not included in the social welfare function.

[34] By assumption, the detection probability when no injurer self-reports $p_0 = e$.

[35] If $h - e\bar{f} < 0$, there will be a destruction of social value when deterring the marginal act.

when evaluated at $e\bar{f}$. The second term, $c'(e)$, is the incremental social cost associated with the higher detection probability. It is easy to verify that the optimal e will be always positive. Taking the fine $\bar{f}$ as fixed and starting with $e = 0$, the incremental social value of raising the probability is positive (since harmful acts with very small benefits will no longer be committed) while the incremental social cost is negligible since, by assumption, $c'(0) = 0$. Hence, the enforcement cost, $c(e)$, will be also positive.

Using equation (2) and rearranging terms, we find that under an enforcement policy with no-leniency, the optimal deterrence threshold (optimal expected fine), $\hat{b}$, satisfies:

$$\hat{b} = e\bar{f} = h - \frac{c'(e)}{\bar{f}g(e\bar{f})}. \tag{3}$$

There may be multiple solutions to this equation. However, under our assumptions on $c(e)$, all of the solutions involve $e > 0$. Then, the optimal enforcement policy has a deterrence threshold $\hat{b} \in (0, h)$.[36] It is interesting to compare the optimal enforcement policy without leniency to a social-welfare benchmark. Without leniency for self-reporting, the first-best outcome is not achievable. Since $\hat{b} < h$, the optimal enforcement policy without leniency has positive enforcement costs and a deterrence threshold that is strictly smaller than the first-best level. Proposition 1 outlines our findings.

Proposition 1. *For any bounded set of fines, an enforcement policy without leniency for self-reporting cannot implement the first-best outcome. The optimal enforcement policy has a maximal fine, a positive enforcement cost, and underdeterrence.*

As has been emphasized in the literature on control of harmful externalities (Polinsky and Shavell, 1984), the failure to implement the first-best outcome with an enforcement policy without leniency for self-reporting is a consequence of having a maximal fine, $\bar{f}$.[37] If the set of fines were instead unbounded, the enforcement agency could get arbitrarily close to the first-best outcome with an extremely high fine coupled with an arbitrarily small probability of detection (Becker, 1968).

[36] Our results regarding optimal enforcement without leniency policies for groups of injurers are aligned with Kaplow and Shavell's (1994) finding on enforcement without self-reporting in single-injurer environments.

[37] Intuitively, having a maximal fine implies that increasing deterrence is expensive. When the benefit to the injurer, b, is very close to social harm, h, then the social benefit of increasing the expected fine is very small (because $b - h$ is negative but small). Since $c'(e) > 0$, increasing the fine leads to a first-order increase in costs.

4.2 Optimal Enforcement Policy with Ordered Leniency

This section characterizes the optimal enforcement policy with ordered leniency. First, we show that for any bounded set of fines, there exists an enforcement policy with ordered leniency that is strictly superior the optimal enforcement policy without leniency described in the previous section. Second, we take the agency's enforcement effort, e, and the corresponding probabilities of detection, p_0 and p_1, as fixed and identify the fine, f, and the leniency multipliers, r_1 and r_2, that generate maximal deterrence (i.e., the highest expected fine). Third, we demonstrate that the first-best outcome may be achieved with ordered-leniency policies at a zero cost when the externalities associated with the harmful activities are not too high.

4.2.1 Superiority of Ordered Leniency

We will show that enforcement policies with ordered leniency for self-reporting always outperform enforcement policies without leniency for self-reporting. As shown in the previous section, without leniency, the optimal enforcement policy has strictly positive enforcement costs, maximal fines, and underdeterrence of harmful activities relative to the first-best level. With ordered leniency, and holding enforcement efforts fixed, the enforcement agency can raise the expected fines and achieve a higher level of deterrence.

Proposition 2. *For any bounded set of fines, there exists an enforcement policy with ordered leniency that is strictly superior to the optimal enforcement policy without leniency for self-reporting.*

The proof of Proposition 2, which is omitted, is straightforward. Intuitive explanation follows. When there is no leniency for self-reporting, $(r_1, r_2) = (1, 1)$, the injurers do not self-report and the deterrence threshold is $\hat{b} = p_0\bar{f} < h$. There is underdeterrence relative to the first-best level, and too many harmful acts are committed. Consider now an ordered-leniency policy $(r_1, r_2) = (p_0 - v, p_0 + 2v)$, where $0 < v < p_0$ is a small positive number. With these leniency multipliers, self-reporting is a strictly dominant strategy for both injurers (Case 1 of Lemma 3). Moreover, the expected fine with ordered leniency is higher than the optimal expected fine without leniency, $p_0\bar{f} < (p_0 + v/2)\bar{f} < h$. Holding the level of enforcement effort and the probabilities of detection fixed, the ordered leniency policy $(r_1, r_2) = (p_0 - v, p_0 + 2v)$ raises the deterrence threshold closer to the first-best level and increases social welfare. More generally, given any optimal enforcement policy without leniency, one can always construct an enforcement policy with ordered leniency that is strictly superior.

Importantly, our findings regarding the superiority of enforcement policies with ordered leniency for groups of injurers complement Kaplow and Shavell's (1994) results for single-injurer environments. In Kaplow and Shavell (1994), self-reporting allows the enforcement agency to save the ex post costs of investigating and prosecuting the defendant. In their model, leniency is valuable because it lowers the ex post investigation costs without significantly weakening deterrence. In our model, holding ex ante enforcement efforts fixed, ordered-leniency policies create a race to the courthouse and increase the detection rate. This increases the expected sanctions and strengthens deterrence.

4.2.2 Maximal Deterrence with Ordered Leniency

Taking the enforcement effort, e, and the corresponding probabilities of detection, p_0 and p_1, as fixed, we now characterize the fine, f, and leniency multipliers, (r_1, r_2), that create the highest possible deterrence (i.e., highest expected fine). We will demonstrate that the fine should be set at the maximal level, $\bar{f}$, and that the ordered-leniency policies that implement maximal deterrence give greater leniency to the first injurer to report and induce immediate self-reporting by both injurers. Importantly, we will show that the optimal leniency multipliers will be different for the Pareto-dominance and risk-dominance refinements. Leniency will be stronger (smaller multipliers) under the Pareto-dominance refinement, and leniency will be milder (larger multipliers) under the risk-dominance refinement.

Denote (r_1^S, r_2^S) and (r_1^M, r_2^M) as the leniency multipliers for the Pareto- and risk-dominance refinements, respectively, and $\hat{b}^S$ and $\hat{b}^M$ as the corresponding deterrence thresholds (expected fines). The superscript S refers to "Strong Leniency" and the superscript M refers to "Mild Leniency." Proposition 3 characterizes the fine and leniency multipliers that create maximal deterrence for groups of potential injurers.

Proposition 3. *Take the enforcement effort e as fixed. Maximal deterrence is obtained with a maximal fine, $f = \bar{f}$, and the following leniency multipliers:*[38]

1. *If $p_1 \leq \frac{1+p_0}{2}$, then $(r_1^S, r_2^S) = (r_1^M, r_2^M) = (p_1 - \Delta, p_1 + \Delta)$ where $\Delta \in [p_1 - p_0, \min\{p_1, 1 - p_1\}]$. The injurers commit the act and self-report at time $t = 0$ if $b > \hat{b}^S = \hat{b}^M = p_1 \bar{f}$, and do not commit the act otherwise.*

[38] When $p_1 \leq \frac{1+p_0}{2}$, the leniency multipliers are not unique.

2. *If* $p_1 > \frac{1+p_0}{2}$, *then* $(r_1^S, r_2^S) = (p_0, 1)$ *and* $(r_1^M, r_2^M) = \left(\frac{2(p_0+p_1)-1}{3}, 1\right)$. *The injurers commit the act and self-report at time* $t = 0$ *if* $b > \hat{b}^S = \left(\frac{1+p_0}{2}\right)\bar{f}$ *(Pareto Dominance) and* $b > \hat{b}^M = \left(\frac{1+p_0+p_1}{3}\right)\bar{f}$ *(Risk Dominance), where* $\hat{b}^S < \hat{b}^M$, *and do not commit the act otherwise.*

Proposition 3 provides fundamental implications for the optimal design of enforcement policies with ordered leniency. The formal analysis is presented in the Appendix. An intuitive discussion of the main insights follows.

Remark 1. *The Fine Is Maximal.*

The highest deterrence is obtained by imposing the maximal fine, $f = \bar{f}$. This follows from the fact that the equilibria of the self-reporting subgame described in Lemmas 2 and 3 do not depend on the level of the fine, f.

Remark 2. *Both Injurers Self-Report.*

Maximal deterrence is achieved when both injurers self-report. It is obvious that a leniency policy where at least one injurer self-reports creates stronger deterrence than a policy where no injurer self-reports. By offering $(r_1, r_2) = (p_0, 1)$, at least one injurer self-reports and the expected fine rises above $p_0\bar{f}$ (the expected fine if neither reports). More specifically, if $p_1 \geq \frac{1+p_0}{2}$, then we are in Case 1 of Lemma 2 where both injurers self-report, and the expected fine is $\left(\frac{1+p_0}{2}\right)\bar{f} > p_0\bar{f}$. On the other hand, if $p_1 < \frac{1+p_0}{2}$, then we are in Case 2 of Lemmas 2 and 3 where exactly one injurer self-reports and the expected fine is $\left(\frac{p_0+p_1}{2}\right)\bar{f} > p_0\bar{f}$. In this latter case, where only one injurer self-reports, deterrence will be even stronger if leniency is granted to the second injurer as well. When $(r_1, r_2) = (p_0, 2p_1 - p_0)$, both injurers self-report and the expected fine rises to $p_1\bar{f}$.[39]

Remark 3. *The First Injurer to Self-Report Always Receives More Lenient Treatment.*

Suppose that $p_1 \geq \frac{1+p_0}{2}$ and $(r_1, r_2) = (p_0, 1)$. We are in Case 1 of Lemma 2, where both injurers self-report. Rewarding the first injurer creates a proverbial race to the courthouse between the two injurers, and the expected fine is $\left(\frac{1+p_0}{2}\right)\bar{f} > p_0\bar{f}$.[40]. If the multipliers were reversed, so $(r_1, r_2) = (1, p_0)$ (i.e., the second to report gets the more lenient treatment), then neither injurer would self-report and the expected fine would be $p_0\bar{f}$, the same as in the absence of a leniency policy.[41] Giving more leniency to the first injurer to report the act increases deterrence.

[39] According to Proposition 3 Case 2, this is an optimal policy ($\Delta = p_1 - p_0$).

[40] If $p_1 \geq \frac{1+p_0}{2}$ then only one injurer would self-report, and the expected fine is still strictly higher than $p_0\bar{f}$

[41] More generally, given an ordered-leniency policy with $r_1 > r_2$, there exists an ordered-leniency policy with $r_1' < r_2'$ that creates stronger deterrence.

Remark 4. *The Second Injurer to Self-Report May Also Receive Leniency.*

When the strength of the inculpatory evidence is weak then the second injurer to report the act receives leniency, too. To see why, suppose that $p_1 < \frac{1+p_0}{2}$. If leniency is granted only to the first injurer, $(r_1, r_2) = (p_0, 1)$, we are in Case 2 of Lemmas 2 and 3 where only one injurer reports the act and the other remains silent, and the deterrence threshold is $\left(\frac{p_0+p_1}{2}\right)\bar{f}$. Now suppose instead that the agency gives partial leniency to the second injurer too, $(r_1, r_2) = (p_0, 2p_1 - p_0)$. With these leniency multipliers, there is a race to the courthouse, both injurers self-report, and the deterrence threshold rises to $p_1\bar{f}$.[42] Deterrence is stronger when the second injurer also receives leniency.[43]

Remark 5. *Stronger Deterrence Is Obtained with Risk Dominance.*

Proposition 3 implies that the deterrence threshold never lower, and may be higher, when the risk-dominance refinement is applied in the self-reporting subgame.[44] In the first part of Proposition 3, when $p_1 \leq \frac{1+p_0}{2}$, leniency multipliers are the same under the Pareto-dominance and risk-dominance refinements, and so the two equilibrium refinements lead to the same deterrence threshold, $\hat{b}^S = \hat{b}^M = p_1\bar{f}$. In the second part of Proposition 3, when $p_1 > \frac{1+p_0}{2}$, the optimal leniency multipliers under the two equilibrium refinements diverge. Suppose that the enforcement agency chooses the mild leniency policy, $(r_1^M, r_2^M) = \left(\frac{2(p_0+p_1)-1}{3}, 1\right)$.[45] Notice that $r_1^M > p_0$,

[42] When $p_1 > 1/2$, maximal deterrence can be achieved by granting leniency to just the first injurer to report, $(r_1, r_2) = (2p_1 - 1, 1)$. With these multipliers, both injurers self-report and the expected fine is $p_1\bar{f}$. When $p_1 < 1/2$, however, $2p_1 - 1$ is a negative number. Some degree of leniency must be granted to the second injurer, too.

[43] Note that, when viewed from an ex post perspective, the second injurer is worse off when he self-reports. Since $r_2^i > p_1$ for $i = S, M$, the second injurer would be better off remaining silent and paying $p_1\bar{f}$ in expectation than self-reporting and paying $r_2^i\bar{f}$. The reason why the second injurer is willing to self-report is because when the injurer is making the important decision about whether or not to self-report, the injurer does not know whether he will obtain the first position or the second position in the self-reporting queue. Hence, a race-to-the-courthouse effect will be always observed in equilibrium when ordered-leniency policies are implemented.

[44] As demonstrated in the Appendix (proof of Proposition 3), the leniency multipliers under Pareto dominance, (r_1^S, r_2^S), satisfy the conditions stated in Case 1 of Lemma 2. When $p_1 \leq \frac{1+p_0}{2}$, the leniency multipliers under risk dominance, (r_1^M, r_2^M), satisfy either the conditions stated in Case 1 of Lemma 2 or the conditions stated in Case 3 of Lemma 2 (both provide the same level of deterrence); when $p_1 > \frac{1+p_0}{2}$, the leniency multiplier under risk dominance, (r_1^M, r_2^M), satisfy the conditions stated in Case 3 of Lemma 2.

[45] Under these leniency multipliers, the environment corresponds to Case 3 of Lemma 2, where the self-reporting subgame is a coordination game with two Nash equilibria (R, R) and (NR, NR). When risk-dominance is applied, maximal deterrence is achieved.

so neither self-reporting nor no-reporting are dominant strategies. When the risk-dominance refinement is applied in the self-reporting subgame, both injurers self-report and the deterrence threshold is $\hat{b}^M = \left(\frac{1+p_0+p_1}{3}\right)\bar{f} > p_0\bar{f}$. When the Pareto-dominance refinement is applied in the self-reporting subgame, neither injurer self-reports and the deterrence threshold is $p_0\bar{f}$. Then, when the Pareto-dominance refinement is applied in the self-reporting subgame, the enforcement agency must lower the multipliers to $(r_1^S, r_2^S) = (p_0, 1)$ to transform the self-reporting subgame into a prisoner's dilemma.[46] The resulting deterrence threshold is $\hat{b}^S = \left(\frac{1+p_0}{2}\right)\bar{f} < \hat{b}^M$. Hence, when Pareto-dominance is applied in the self-reporting subgame, the deterrence threshold is smaller and the incentives to engage in the harmful activity rise.

4.2.3 Optimal Enforcement Effort with Ordered Leniency

This section characterizes the optimal enforcement effort e when ordered-leniency policies are implemented. Remember that Proposition 3 identifies the leniency multipliers and fine that create maximal deterrence (i.e., the highest expected fine), and that superscripts S and M denote the leniency policies under the Pareto- and risk-dominance refinements, respectively.

The next lemma, which follows from Proposition 3, will be used in the analysis of the optimal enforcement effort e when ordered-leniency policies are implemented. Recall that $p_0 = e$ and $p_1 = e + (1-e)\pi$, where $\pi \in (0,1)$ represents the exogenous strength of inculpatory evidence. Then, $p_1 \leq \frac{1+p_0}{2}$ holds if and only if $\pi \leq \frac{1}{2}$, and $p_1 > \frac{1+p_0}{2}$ holds if and only if $\pi > \frac{1}{2}$. In other words, Cases 1 and 2 of Lemma 4 correspond to Cases 1 and 2 of Proposition 3.[47]

Lemma 4. *The ordered-leniency multipliers (r_1^S, r_2^S) and (r_1^M, r_2^M), characterized in Proposition 3, yield corresponding expected fines $\hat{b}^S(e,\pi)$ and $\hat{b}^M(e,\pi)$ for the injurers. These functions, which are continuous and piecewise differentiable, satisfy:*

1. *If $\pi \leq \frac{1}{2}$, then $\hat{b}^S(e,\pi) = \hat{b}^M(e,\pi) = [\pi + (1-\pi)e]\,\bar{f}$ and $0 < \frac{\partial \hat{b}^i(e,\pi)}{\partial e} < \bar{f}$ for $i = S, M$.*
2. *If $\pi > \frac{1}{2}$, then $\hat{b}^S(e,\pi) = \left(\frac{1+e}{2}\right)\bar{f}$ and $\hat{b}^M(e,\pi) = \left[\frac{(1+\pi)+(2-\pi)e}{3}\right]\bar{f}$. Furthermore, $\hat{b}^S(e,\pi) < \hat{b}^M(e,\pi)$ and $0 < \frac{\partial \hat{b}^M(e,\pi)}{\partial e} < \frac{\partial \hat{b}^S(e,\pi)}{\partial e} < \bar{f}$.*

We now describe the circumstances under which ordered-leniency policies can achieve the first-best outcome. In the first-best outcome, the injurers commit the act if and only if the benefit

[46] This new strategic environment corresponds to Case 1 of Lemma 2, where (R, R) is the unique Nash equilibrium.

[47] Consider Case 1 of Proposition 3, where $p_1 \leq \frac{1+p_0}{2}$. Substituting $p_0 = e$ and $p_1 = e + (1-e)\pi$ into this expression, we find that $p_1 \leq \frac{1+p_0}{2}$ holds if and only if $\pi \leq \frac{1}{2}$. Similarly logic applies to Case 2 of Proposition 3.

exceeds the social harm, $b > h$ and no effort is spent on enforcement, $e = 0$. In this benchmark, $p_0 = 0$ and $p_1 = \pi$.

When $\pi \leq \frac{1}{2}$, we are in Case 1 of Proposition 3 and Lemma 4. With no enforcement effort, $e = 0$, the maximal deterrence is obtained with a maximal fine $\bar{f}$ and leniency multipliers $(r_1^S, r_2^S) = (r_1^M, r_2^M) = (0, 2\pi)$. With these multipliers, the injurers are deterred from committing the act when $b \leq \hat{b}^S = \hat{b}^M = \pi\bar{f}$. Note that if the level of harm is less than the deterrence threshold, $h < \pi\bar{f}$, then there would be overdeterrence relative to the first-best level. However, this may be easily solved by reducing the fine below its maximal level, granting additional leniency to the injurers, or both. When the expected fine is exactly equal to the social harm, h, then the injurers will commit the act if and only if $b > h$, as desired. When the level of harm exceeds the deterrence threshold, $h > \pi\bar{f}$, then there is underdeterrence relative to the first-best level. In this case, deterrence can be improved by spending resources on enforcement. Taken together, when $\pi \leq \frac{1}{2}$, the first-best outcome is achieved at zero cost if and only if the harm is not too high, $h \leq \pi\bar{f}$.

When $\pi > \frac{1}{2}$, we are in Case 2 of Proposition 3 and Lemma 4. Suppose the enforcement efforts are zero, $e = 0$. If the Pareto-dominance refinement is applied to the self-reporting subgame, then the multipliers that create maximal deterrence are $(r_1^S, r_2^S) = (0, 1)$ and the associated deterrence threshold is $\hat{b}^S = \left(\frac{1}{2}\right)\bar{f}$. If the level of harm is below this threshold, $h < \left(\frac{1}{2}\right)\bar{f}$, then the first-best outcome may be obtained by lowering the fine, lowering the leniency multiplier for the second injurer, or both. If the risk-dominance refinement applies, then the leniency multipliers that create the maximal deterrence are $(r_1^M, r_2^M) = \left(\frac{2\pi-1}{3}, 1\right)$ and the associated deterrence threshold is $\hat{b}^M = \left(\frac{1+\pi}{3}\right)\bar{f}$. Applying the same logic as before, when $h < \left(\frac{1+\pi}{3}\right)\bar{f}$, the first-best outcome can be obtained by lowering the fine, lowering the leniency multipliers, or both. Hence, when $\pi > \frac{1}{2}$, the first-best outcome is achieved at zero cost if and only if the harm is not too high, $h \leq \left(\frac{1}{2}\right)\bar{f}$ (Pareto Dominance) and $h \leq \left(\frac{1+\pi}{3}\right)\bar{f}$ (Risk Dominance).

Proposition 4 establishes the necessary and sufficient conditions under which the enforcement agency can implement the first-best outcome with an ordered-leniency policy at a zero cost, and describes the second-best enforcement policy when the first-best outcome cannot be achieved.

Proposition 4. *For any bounded set of fines, an optimal enforcement policy with ordered leniency for self-reporting can implement the first-best outcome at zero cost if and only if* $h \leq \hat{b}^S(0, \pi) = \min\{\pi, \frac{1}{2}\}\bar{f}$ *under the Pareto-dominance refinement, and* $h \leq \hat{b}^M(0, \pi) = \min\{\pi, \frac{1+\pi}{3}\}\bar{f}$ *under the risk-dominance refinement. When* $h > \hat{b}^i(0, \pi), i = S, M$*, the second-best enforcement policy involves a maximal fine, positive enforcement costs, and underdeterrence relative to the first best.*

Intuitively, when the externalities associated with the harmful activities, h, are not too high and enforcement policies with ordered leniency for self-reporting are implemented, injurers are induced to report their harmful acts immediately without affecting their incentives to refrain from committing the acts relative to the first-best outcome. Importantly, the enforcement agency does not need to spend resources on enforcement, $c(e) = 0$. When the externalities associated with the harmful activities, h, are relatively high, the first-best outcome cannot be not obtained. The enforcement agency must spend resources to detect the harmful activities and too many harmful activities will be committed.

Taken together, our previous findings provide a social welfare rationale for the current use of ordered-leniency policies in the real-world. First, holding the enforcement costs fixed, we proved that an enforcement policy with ordered leniency is strictly superior to the optimal enforcement policy without leniency (Proposition 2). Second, we showed that ordered-leniency policies that generate maximal deterrence give successively larger discounts to injurers who secure higher positions in the self-reporting queue, creating a so-called "race to the courthouse" where all injurers report the act immediately (Proposition 3). Third, we demonstrated that socially-optimal level of deterrence can be obtained at zero cost when the externalities associated with the harmful activities are not too high (Proposition 4).

5 Experimental Evidence

This section reports the results from a series of experiments with human subjects. We investigate whether the behavior of the subjects follows the theoretical predictions regarding the effects of ordered-leniency policies on reporting of harmful acts (self-reporting of the act by one or both injurers, and report time) and deterrence (decision not to commit the act by one or both potential injurers). We study three leniency environments: No Leniency (N), where no fine reduction for self-reporting is applied; Strong Leniency (S), where the fine reduction for self-reporting is large, and hence, the expected fine is small; and, Mild Leniency (M), where the fine reduction for self-reporting is small, and hence, the expected fine is high.

5.1 Model Parameterization and Theoretical Predictions

Consider the following parameterization of the model. Across leniency environments, the parameter values are as follows: $b \in [200, 1600]$; $f = 900$; $p_0 = 0.4$, $p_1 = 0.9$, and $t \in [0, 90]$ (measured in

Figure 2: Strategic-Form Representation of the Self-Reporting Subgame (Expected Payoffs)

No Leniency (N)

	NR	R
NR	$b-360$, $b-360$	$b-810$, $b-900$
R	$b-900$, $b-810$	$b-900$, $b-900$

Strong Leniency (S)

	NR	R
NR	$b-360$, $b-360$	$b-810$, $b-300$
R	$b-300$, $b-810$	$b-600$, $b-600$

Mild Leniency (M)

	NR	R
NR	$b-360$, $b-360$	$b-810$, $b-420$
R	$b-420$, $b-810$	$b-660$, $b-660$

seconds). We consider two sets of leniency multipliers: $r_1^S = 0.333$ and $r_2^S = 1$, in case of Strong Leniency; and, $r_1^M = 0.466$ and $r_2^M = 1$, in case of Mild Leniency.[48] The strategic-form representation of the self-reporting subgame for the No Leniency, Strong Leniency, and Mild Leniency environments under these parameters is presented in Figure 2.

The equilibrium predictions are as follows. (1) No Leniency: Both potential injurers decide to commit the act in Stage 1 when $b > 360$; if the act is committed, both injurers decide not to report the act in Stage 2. (2) Strong Leniency: Both potential injurers decide to commit the act in Stage 1 when $b > 600$; if the act is committed, both injurers decide to report the act immediately in Stage 2. (3) Mild Leniency with Pareto-dominance refinement: Both potential injurers decide to commit the act in Stage 1 when $b > 360$; if the act is committed, both injurers decide not to report the act in Stage 2. (4) Mild Leniency with risk-dominance refinement: Both potential injurers decide to commit the act in Stage 1 when $b > 660$; if the act is committed, both injurers decide to report the act immediately in Stage 2. In sum, the highest individual benefit b to induce the injurers not to commit the act in Stage 1, the "deterrence threshold," is $b = 360$, $b = 600$ or $b = 660$.[49] As described, the specific deterrence threshold depends on the leniency environment

[48] By Proposition 3 (Case 2, $p_1 > \frac{1+p_0}{2}$), the leniency multipliers that generate maximal deterrence for these parameter values are $(r_1^S, r_2^S) = (0.400, 1.00)$ and $(r_1^M, r_2^M) = (0.533, 1.00)$. To break indifference, we deduct $\varepsilon = 0.067$ from r_1^S and r_1^M.

[49] We assume that in case of indifference, the potential injurer decides not to commit the act.

Table 1 – Theoretical Point Predictions

	Deterrence Rate	Report Rates[a]		No-Report Rate[a]	Report Time[b]
		(R, R)	(R, NR)/(NR, R)	(NR, NR)	
No Leniency (N)					
$b \in [200, 360]$	1	0	0	1	–
$b \in (360, 1600]$	0	0	0	1	–
Strong Leniency (S)					
$b \in [200, 600]$	1	1	0	0	0
$b \in (600, 1600]$	0	1	0	0	0
Mild Leniency (M)					
• Pareto Dominance					
$b \in [200, 360]$	1	0	0	1	–
$b \in (360, 1600]$	0	0	0	1	–
• Risk Dominance					
$b \in [200, 660]$	1	1	0	0	0
$b \in (660, 1600]$	0	1	0	0	0

Notes: [a]Report and no-report rates conditional on committing the act; [b]report time in seconds.

and the equilibrium refinement adopted. Table 1 outlines the theoretical point predictions.

5.2 Experimental Design

Next, we present a description of the laboratory implementation of our theoretical environments.

Experimental Conditions

Procedural regularity is accomplished by developing a software program that allows the subjects to play the game by using networked personal computers. The software, constructed using the Java programming language, consists of 3 versions of the game, reflecting the three experimental conditions: No Leniency (N), Strong Leniency (S) and Mild Leniency (M).[50] To ensure control and replicability, a free-context environment is implemented. Specifically, neutral labels are used to denote the subjects' roles: Players B1 and B2 (potential injurers 1 and 2, respectively). The "Act" is described as an economic decision involving potential benefits (associated with Stage 1) and potential losses (associated with Stage 2).[51] The players' choices are also labeled in a neutral

[50]The use of a JAVA software especially designed for this study allows us to have flexibility in the design of randomization processes and the design of user-friendly screens.

[51]Please see the Appendix for a sample of the instructions (Mild Leniency condition).

way: Decision whether "To Agree to Jointly Commit the Act" or "Not to Agree to Jointly Commit the Act;" and, decision whether "To Report the Act" or "Not to Report the Act." The game includes 5 practice matches and one actual match. The practice matches allow the subjects to experiment with the different options and hence, learn about the consequences of their choices. Only the actual match is considered in the subject's payment.

The benchmark game corresponds to the Strong Leniency condition (S). Subjects play the role of Player B1 or Player B2. The roles of Players B1 and B2 are similar. Each match involves two stages. In Stage 1, each player independently decides whether to commit the act. The players have 90 seconds to make their decisions in Stage 1. After the decisions are made, both players are informed about the other player's decision. If both players agree to commit the act, Stage 2 starts. Otherwise, the game ends. In Stage 2, each player independently decides whether to report the act. The players have 90 seconds to decide whether to report the act and submit their reports. When both players decide to report at the same time, the computer randomly assigns the first position in the self-reporting queue to each player with equal probability. After the decisions are made, both players are informed about the decision of the other player and the payoffs for both players, and the game ends. The payoffs reflect the Strong Leniency policy (S): The first to self-report receives a fine reduction. It is worth noting that our lab implementation also allows us to collect data on the report time. These data are used to assess whether ordered-leniency policies exhibit a "race to the courthouse" effect, i.e., whether immediate report is observed.

Variations of the benchmark game satisfy the other experimental conditions. In the Mild Leniency condition (M) and No Leniency condition (N), the subjects play a similar game. The only difference across conditions refers to the players' payoffs. Specifically, in the Mild Leniency condition (M), the first to report receives a fine reduction, which is lower than the one granted in case of the Strong Leniency condition (S). In the No Leniency (N) condition, the first to report does not receive a fine reduction.

Each experimental condition includes four 24-subject sessions. To achieve *independent observations in the actual match,* we use the following role and pairing procedure per session: (1) The total number of subjects are randomly assigned to one of the following two groups, Group 1 and Group 2; (2) half of the subjects in each group is assigned the role of Player B1 and the other half is assigned the role of Player B2; (3) for each practice match, Players B1 from Group 1 are randomly paired with Players B2 from Group 2, and Players B2 from Group 1 are randomly paired with Players B1 from Group 2; (4) for the actual match, Players B1 and B2 from Group 1 are randomly paired, and Players B1 and B2 from Group 2 are randomly paired. The same protocol

Table 2 – Experimental Conditions

b-Value Segments	No Leniency (N)	Strong Leniency (S)	Mild Leniency (M)
$b \in [200, 360]$	4	4	4
$b \in (360, 600]$	11	11	11
$b \in (600, 660]$	11	11	11
$b \in (660, 1600]$	22	22	22
Total Number of Pairs	48	48	48

for pair formation is applied across sessions and conditions. As a result, for each session of 24 subjects, 12 independent observations (pairs) are obtained. Hence, 48 independent observations per condition and 144 independent observations in total are obtained.

Table 2 summarizes the information regarding the experimental conditions and observations per b-value segment for the actual match. The theoretical deterrence thresholds guide the design of the distribution of b-values. To ensure *comparability across conditions*, we randomly predetermine the b-values used in the actual match of each of the four sessions of a condition, and apply these values to each condition.[52] For each condition, the total number of b-values for the actual match is equal to 48 (12 values per session; 4 sessions per condition).[53] To ensure *consistency across sessions and conditions*, we randomly predetermine the b-values for each of the five practice matches, and apply these values across sessions and conditions.[54]

Experimental Sessions

We ran twelve 80-minute sessions of 24 subjects each (four sessions per condition; 96 subjects per condition; 288 subjects in total) at Harvard University. Each session was conducted by two research assistants at the Harvard Decision Science Laboratory. Subjects were recruited using the

[52]The chosen distribution of b-values has the following features: Four b-value segments are considered, $[200, 360]$, $(360, 600]$, $(600, 660]$, and $(660, 1600]$; the segments include 8, 23, 23 and 46% of the total b-values, respectively; for each segment, the specific b-values are randomly chosen (equally likely values).

[53]The adopted distribution of b-values allows us: (1) To collect a sufficiently high number of observations to perform statistical analysis of deterrence across conditions, and across relevant b-value segments within each condition; and, (2) to collect a sufficiently high number of observations in which Stage 2 occurs with certainty in equilibrium across conditions ($b > 660$) to perform statistical analysis of detection across conditions.

[54]For each practice match, at least one b-value pertains to each of the four b-value segments; and, the majority of b-values pertain to the last b-value segment, which in theory, elicits a self-reporting stage with certainty. Hence, the distribution of b-values ensures that the subjects will get enough experience regarding the self-reporting stage.

lab's Sona computer program and the lab's subject pool. Subjects were allowed to participate in one experimental session only, and received information only about the game version that they were assigned to play. The participant pool included undergraduate and graduate students from Harvard, Boston and Northeastern universities, from a wide variety of fields of study. A laboratory currency called the "token" (29 tokens = 1 U.S. dollar) was used in our experiment. To avoid negative payoffs, each subject received an initial endowment equal to 700 tokens.[55] The show-up fee was equal to $10. The average game earnings was equal to $32. Then, the average total payment was equal to $42 (average game earnings plus participation fee) for an 80-minute session.

At the beginning of each session, written instructions were provided to the subjects (see the Appendix). The instructions about the game and the software were verbally presented by the experimenter to create common knowledge. Specifically, subjects were informed: (1) about the game structure, possible choices, and payoffs; (2) about the random process of allocating roles; (3) about the randomness and anonymity of the process of forming pairs;[56] (4) about the token/dollar equivalence, and that they would receive the dollar equivalent of the tokens they held at the end of the session. Finally, subjects were asked to complete a set of exercises to ensure their ability to read the information tables. The answers to the exercises were read aloud by the research assistants. Questions about the written instruction and questions about the exercises were answered by the research assistants privately and before the beginning of the practice matches. The rest of the session was entirely played using computer terminals and the software designed for this experiment. After the actual match, subjects were required to fill out a short questionnaire with general demographic questions. At the end of each experimental session, subjects privately received their monetary payoffs in cash.

5.3 Qualitative Hypotheses

The qualitative hypotheses are presented below. Cooper et al. (1990) suggest that risk-dominance is generally the equilibrium selection criterion chosen by subjects in the lab when there are multiple equilibria.[57] Then, we might expect that the majority of subjects will apply the risk-dominance

[55]Note that the minimum possible b-value was equal to 200 tokens and the maximum possible fine was equal to 900 tokens. Then, the minimum possible match payoff was equal to 0 tokens.

[56]In particular, subjects were informed that they would not play with the same partner in any practice match; and, that they would not play with any of their previous partners in the actual match.

[57]Landeo and Spier (2009) offer important evidence in favor of the risk-dominance refinement in contractual settings with multiple equilibria. Burton and Sefton (2004) provide powerful evidence of the role of riskiness in the choice of a strategy. See Ochs (1995) for a survey of seminal work on coordination games.

refinement in our experiment. Hence, the hypotheses related to Mild Leniency are constructed under the risk-dominance refinement.

Hypothesis 1. *Strong Leniency increases the rate of self-reporting by both injurers, with respect to No Leniency. No Leniency increases the rate of no-reporting by both injurers, with respect to Strong Leniency.*

According to our theory, Strong Leniency induces both injurers to self-report, i.e., (R, R) is the unique N.E. of the reporting subgame. In contrast, No-Leniency induces both injurers to no-report, i.e., (NR, NR) is the unique N.E. of the reporting subgame.

Hypothesis 2. *Mild Leniency increases the rate of self-reporting by both injurers with respect to No Leniency. No Leniency increases the rate of no-reporting by both injurers, with respect to Mild Leniency.*

According to our theoretical predictions under the risk-dominance refinement, Mild Leniency induces both injurers to self-report, i.e., (R, R) is chosen by both injurers. If the Pareto-dominance refinement is applied instead, Mild Leniency induces both injurers to no-report, i.e., (NR, NR) is chosen by both injurers.

Hypothesis 3. *Mild Leniency and Strong Leniency exhibit the same rate of self-reporting by both injurers. Mild Leniency and Strong Leniency exhibit the same zero rate of no-reporting by both injurers.*

According to our theoretical predictions under the risk-dominance refinement, Mild Leniency induces both injurers to self-report, i.e., (R, R) is chosen by both injurers. Importantly, both the Mild and Strong Leniency conditions exhibit a zero rate of no-reporting by both injurers. If the Pareto-dominance refinement is applied instead, Mild Leniency induces both injurers to no-report, i.e., (NR, NR) is chosen by both injurers.

Hypothesis 4. *Strong and Mild Leniency exhibit a "race to the courthouse" effect – self-reporting occurs immediately.*

Under optimal ordered-leniency policies, earlier reporting implies higher penalty reduction. According to our theory, when the risk-dominance refinement is applied, Strong and Mild Leniency generate a "race to the courthouse" between the two members of the group of injurers. As a result,

in equilibrium, both injurers will self-report immediately. In the lab, subjects have between zero and 90 seconds to decide to report the act and submit their reports. Then, it is expected that "immediate report" will occur at a time slightly later than zero seconds.

Hypothesis 5. *Within each leniency environment, the deterrence rate is lower when the benefits from the harmful act are greater than the deterrence threshold.*

According to our theory, the benefits associated with the harmful act incentivize the potential injurers to commit the act. In equilibrium, the act is committed only when the benefits are higher than the expected fine (i.e., when the benefits are higher than the deterrence threshold).

Hypothesis 6. *When the benefits from the harmful act are not greater than 660, Mild Leniency increases the deterrence rate, with respect to Strong Leniency; and, Strong and Mild Leniency increase the deterrence rate, with respect to No Leniency.*

According to our theoretical predictions, only benefits greater than the expected fine (i.e., benefits greater than the deterrence threshold) will induce potential injurers to commit the harmful act. Under the risk-dominance refinement, the expected fine for Mild Leniency is higher than the expected fine for No Leniency (660 v. 360), and higher than the expected fine for Strong Leniency (660 v. 600). Then, in theory, Mild Leniency will exhibit the highest deterrence rate and No Leniency will exhibit the lowest deterrence rate. If the Pareto-dominance refinement is applied instead, the expected fine for Mild and No Leniency will be the same (360). Then, there will not be differences in deterrence rates between these two policies.

It is worth noting that, in theory, Strong and Mild Leniency (when the risk-dominance refinement is applied) have the property of incentivizing both injurers to be the first to report. As a result, both injurers will report immediately, and hence, they will be equally likely to get the first position in the self-reporting queue (i.e., the chance of each injurer to get the first position will be equal to 50%).[58] In the lab, however, some subjects might exhibit cognitive biases, such as self-serving bias (Babcock et al., 1995),[59] and hence believe that their chances to be the first to report are greater than the chances of the other injurers (i.e., they might believe that their

[58]Remember that, in the Strong Leniency environment, the expected fine under (R, R) and equal likelihoods of getting the first or second position in the self-reporting queue is equal to .50(300) + .50(900) = 600; and, in the Mild Leniency environment, the expected fine under (R, R) and equal likelihoods of getting the first or second position in the self-reporting queue is equal to .50(420) + .50(900) = 660.

[59]Self-serving bias is attributed to motivated reasoning, i.e., a propensity to reason in a way that supports the individual's subjectively favored beliefs by attending only to some available information (Kunda, 1990, 1987). See

chances to get the first position in the self-reporting queue are greater than 50%). In the limiting case, some subjects might believe that they will always be the first to report. Then, under Strong Leniency, they might consider a fine equal to 300 instead of an expected fine equal to 600 when making their decision about committing the act.[60] Similarly, under Mild Leniency, some subjects might consider a fine equal to 420 instead of an expected fine equal to 660 when making their decision about committing the act.[61] As a result, we might observe deterrence rates lower than those predicted by the theory.[62]

5.4 Results

This section discusses our experimental findings. Given our experimental design, the collected 144 observations (pairs) are independent. Then, it is appropriate to use non-parametric statistical analysis. Specifically, our analysis involves the use of the Fisher-exact, Wilcoxon signed-rank, and Wilcoxon rank-sum (Mann Whitney) tests.

Table 3 outlines our findings about the effect of ordered-leniency policies on self-reporting. The Report and No-Report rates are conditional on committing the act. Consider first the effect of the implementation of a Strong Leniency policy on self-reporting (first two lines of Table 3). In theory, self-reporting by both injurers is the unique N.E. under Strong Leniency, and no-reporting by both injurers is the unique N.E. under No Leniency. Our findings suggest that the (R, R) rate is significantly higher under Strong Leniency (.79 v. .00, for Strong and No Leniency, respectively,

Babcock et al. (1995) and Landeo (2009) for experimental evidence of self-serving bias, and Landeo et al. (2013) for theoretical work on self-serving bias in incomplete-information environments. See also Landeo (2018) for further discussion of theoretical and experimental studies on self-serving bias.

[60]The expected fine for the biased injurer is 300 irrespective of the reporting decision of the other injurer. Importantly, under these biased beliefs, report is the dominant strategy for the biased injurer. This is because his expected fines under no-report are 360 or 810, depending on the choice of no-report or report by the other injurer, respectively. Then, as in the case of the environment with unbiased injurers, the biased injurer will choose to report.

[61]The expected fine for the biased injurer is 420 irrespective of the reporting decision of the other injurer. Importantly, under these biased beliefs, report will be the best response when the other injurer chooses to report ($420 < 810$) and no-report will be the best response when the other injurer chooses no-report ($360 < 420$). Then, as in the case of the environment with unbiased injurers, the biased injurer might choose to report or no-report, according to his beliefs about the strategy that the other injurer will choose and his assessment of the riskiness of each strategy.

[62]Following prospect theory (Kahneman and Tversky, 1979), lower than predicted deterrence thresholds might also suggest risk-seeking behavior triggered by the subjects' focus on the loss (fine) component of the payoff function.

Table 3 – Effect of Ordered- Leniency Policies on Self-Reporting[a]

	Number of Pairs	Report Rates (R, R)	Report Rates (R, NR)/(NR, R)	No-Report Rate (NR, NR)
Strong Leniency v. No Leniency	74	.79 v. .00	.21 v. .11	.00 v. .89
		$p < .01$	$p = .23$	$p < .01$
Mild Leniency v. No Leniency	75	.58 v. .00	.38 v. .11	.04 v. .89
		$p < .01$	$p < .01$	$p < .01$
Mild Leniency v. Strong Leniency	79	.58 v. .79	.38 v. .21	.04 v. .00
		$p < .05$	$p < .10$	$p = .25$

Notes: [a]p-values correspond to the one-sided Fisher-exact test.

$p < .01$). Our findings also indicate that the (NR, NR) rate is significantly higher under No Leniency (.89 v. .00, for No Leniency and Strong Leniency, respectively; $p < .01$). These results are aligned with our theoretical qualitative predictions, and provide strong support to Hypothesis 1.

Result 1. *Strong Leniency increases the rate of self-reporting by both injurers, with respect to No Leniency. No Leniency increases the rate of no-reporting by both injurers, with respect to Strong Leniency.*

Second, we analyze the effect of implementing a Mild Leniency policy on self-reporting (lines 3 and 4 of Table 3). In theory, when the risk-dominance refinement is applied, Mild Leniency increases the likelihood of self-reporting by both injurers, with respect to No Leniency; and, No Leniency increases the likelihood of no-reporting by both injurers with respect to Mild Leniency. If the Pareto-dominance refinement is applied instead, Mild Leniency and No Leniency exhibit the same 100% likelihood of no-reporting by both injurers. Our results suggest that Mild Leniency significantly increases the (R, R) rate (.58 v. zero, for Mild and No Leniency, respectively; $p < .01$). Our findings also indicate that No Leniency significantly increases the (NR, NR) rate (.89 v. .04, for No Leniency and Mild Leniency, respectively; $p < .01$). These findings are aligned with our theoretical predictions under the risk-dominance refinement. Importantly, our results under Mild Leniency demonstrate that the majority of pairs chose the risk-dominant N.E (i.e., chose the (R, R) outcome in 58% of the cases), and only a minority of pairs chose the Pareto-dominant N.E. (i.e., chose the (NR, NR) outcome in 4% of the cases). Our findings also indicate the presence of

a coordination problem under Mild Leniency: 38% of the pairs ended up at the (R, NR)/(NR, R) outcomes. Our results provide strong support to Hypothesis 2.

Result 2. *Mild Leniency increases the rate of self-reporting by both injurers, with respect to No Leniency. No Leniency increases the rate of no-reporting by both injurers, with respect to Mild Leniency.*

Third, we evaluate the effect of implementing Mild and Strong Leniency on self-reporting (lines 5 and 6 of Table 3). In theory, when the risk-dominance refinement is applied, Strong and Mild Leniency exhibit the same rates of self-reporting by both injurers and the same rates of no-reporting by both injurers. If the Pareto-dominance refinement is applied instead, Strong Leniency increases the rate of self-reporting by both injurers, and Mild Leniency increases the rate of no-reporting by both injurers. Our findings suggest that Strong Leniency marginally increases the (R, R) rate (.79 v. .58, for Strong and Mild Leniency, respectively; $p < .05$). This result might be explained by the coordination problem exhibited under Mild Leniency: Although the majority of pairs chose the risk-dominant N.E. (i.e., chose the (R, R) outcome in 58% of the cases), 38% of the pairs ended up at the (R, NR)/(NR, R) outcomes. Importantly, the rates of self-reporting by one or both injurers for Strong and Mild Leniency are not significantly different (.96 v. 1, for Mild and Strong Leniency, respectively; $p = .25$). In other words, Mild Leniency does not increase the rate of no-reporting by both injurers (.04 and zero, for Mild and Strong Leniency, respectively; $p = .25$). Hence, our findings are aligned with our theoretical predictions under the risk-dominance refinement and provide support to Hypothesis 3.

Result 3. *Mild Leniency and Strong Leniency exhibit the same rates of self-reporting by one or both injurers. Mild Leniency and Strong Leniency exhibit the same zero rate of no-reporting by both injurers.*

Next, we present an analysis of the report time under ordered-leniency policies. Table 4 summarizes our findings. In theory, self-reporting occurs immediately under Strong and Mild Leniency. Recall that subjects had between zero and 90 seconds to decide to report the act and submit their reports. Then, it is expected that "immediate report" will be represented by a time slightly higher than zero seconds. Our results suggest that the modal report times under Strong and Mild Leniency are equal to 1 second;[63] and the average report times for Strong and Mild

[63]In fact, the majority of subjects exhibited a 1-second report time: 57 and 51%, for Strong and Mild Leniency, respectively.

Table 4 – Report Time under Ordered-Leniency Policies[a]
(Within-Condition Analysis)

	Number of Individuals	Report Time[b]	
		Average Value	Modal Value
Strong Leniency (S)	70	1.64	1.00
		(1.30)	
Mild Leniency (M)	61	2.21	1.00
		(2.70)	

Notes: [a]Standard deviations in parentheses; [b]report time in seconds.

Leniency are not greater than 2.21 seconds.[64] Importantly, although the median report times for Strong and Mild Leniency are significantly different from zero seconds (as expected), we cannot reject the null hypotheses that the median report time under Strong Leniency is equal to 1 second and the median report time under Mild Leniency is equal to 1.5 seconds (Wilcoxon signed-rank test, $p = .15$ and $p = .61$, respectively).[65] Our results suggest that ordered-leniency policies incentivized the participants to minimize their report times in order to get the first position in the self-reporting queue. In other words, ordered-leniency policies generated a "race to the courthouse" between the two members of the group of injurers. These findings are aligned with our theory and provide strong support to Hypothesis 4.

Result 4. *Strong and Mild Leniency exhibit a "race to the courthouse" effect – self-reporting occurs immediately.*

We now study, the effect of the benefits derived from the commission of the harmful act (b-value) on deterrence, a within-condition analysis. Table 5 outlines our findings. For each condition, we consider the theoretical deterrence threshold, and compare the deterrence rates for b-values below and above this threshold. Given that our previous analysis of the effect of ordered-leniency policies on self-reporting suggests that the risk-dominant N.E. (the (R, R) outcome) is chosen by the majority of subjects under Mild Leniency, we consider the deterrence threshold

[64]When the subjects decided to report under No Leniency (an off-equilibrium strategy), the average report time was equal to 16.2 seconds. Our results suggest that the median report time under No Leniency is significantly higher than the median report times under Strong and Mild Leniency; Wilcoxon rank-sum (Mann-Whitney) test, $p = .02$ and $p = .04$, respectively.

[65]Our findings also indicate that the median report times for Strong and Mild Leniency are not significantly different, $p = .41$, Wilcoxon rank-sum (Mann-Whitney) test.

Table 5 – Effect of b-Value of Deterrence[a]
(Within-Condition Analysis)

	Number of Pairs	Deterrence Rate
Strong Leniency (S)		
$b \in [200, 600]$	15	.47
$b \in (600, 1600]$	33	.06
		$p < .01$
Mild Leniency (M)		
$b \in [200, 660]$	26	.31
$b \in (660, 1600]$	22	.00
		$p < .01$
No Leniency (N)		
$b \in [200, 360]$	4	.75
$b \in (360, 1600]$	44	.23
		$p = .06$

Notes: [a]p-value corresponds to the one-sided Fisher-exact test.

that corresponds to this refinement. Two important insights deserve attention. First, our findings suggest that the benefits from the harmful act incentivize the potential injurers to commit the act. Consider, for instance, the effect of the b-value on deterrence for the case of Strong Leniency (first three lines of Table 5). The data indicate that when the benefits are above the theoretical threshold (600), the deterrence rate is equal to 6%; and, when the benefits are below this threshold, the deterrence rate rises to 47% (a statistically significant effect, $p < .01$). More generally, for each leniency environment, the likelihood of deterrence is significantly lower when the benefits are greater than the theoretical deterrence threshold. These results are aligned with our theory and provide strong support to Hypothesis 5.

Result 5. *Within each leniency environment, the deterrence rate is lower when the benefits from the harmful act are greater than the deterrence threshold.*

Second, our results indicate that the deterrence rates for b-values below the theoretical thresholds are lower than the rates predicted by the theory for the Strong and Mild Leniency policies (47 and 31% instead of 100%). The low deterrence rate under Strong Leniency might suggest that some subjects considered an alternative deterrence threshold. In theory, both injurers report immediately, and hence, they are equally likely to get the first position in the self-reporting queue

Table 6 – Effect of Ordered-Leniency Policies on Deterrence

	Number of Pairs	Deterrence Rate	p-value[a]
Strong v. No Leniency			
$b \in [200, 660]$	52	.31 v. .38	$p = .39$
$b \in (660, 1600]$	44	.05 v. .14	$p = .30$
Mild v. No Leniency			
$b \in [200, 660]$	52	.31 v. .38	$p = .39$
$b \in (660, 1600]$	44	.00 v. .14	$p = .12$
Mild v. Strong Leniency			
$b \in [200, 660]$	52	.31 v. .31	$p = .62$
$b \in (660, 1600]$	44	.00 v. .05	$p = .50$

Notes: [a]p-values correspond to the one-sided Fisher-exact test.

(i.e., each has a 50% chance to get the first position). In the lab, some subjects might exhibit cognitive biases, such as self-serving bias (Babcock et al., 1995), and believe that their chances to be the first to report are greater than the chances of the other injurers (i.e., that their chances are greater than 50%). In the limiting case, some subjects might believe that they will always be the first to report. As a result, they will consider a fine equal to 300 instead of an expected fine equal to 600 when making their decisions about committing the act. Hence, they will also choose to commit an act when the benefits are lower than 600.

Regarding Mild Leniency, the low deterrence rate might be explained, in part, by the decisions in Stage 2. Remember that, although the majority of subjects coordinated on the risk-dominant N.E. (R, R), some subjects attempted to coordinate on the Pareto-dominant N.E. (NR, NR). Subjects who expect to coordinate on (NR, NR) will consider a lower deterrence threshold (360 instead of 660). Hence, they will also choose to commit the act when the benefits are lower than 660. The low deterrence rate might be also explained by the presence of self-serving bias on the subjects' beliefs about their chances to be the first to report. Some subjects might believe that they will always be the first to report. As a result, they will consider a fine equal to 420 instead of an expected fine equal to 660. Hence, they will also choose to commit the act when the benefits are lower than 660.[66]

Finally, we assess the effect of ordered-leniency policies on deterrence. Table 6 summarizes our

[66]Failure to apply backward induction due to limited computational abilities might explain some of the deviations from the theoretical predictions on deterrence, under Strong and Mild Leniency. See Johnson et al. (2002) for seminal experimental work on backward induction failure in sequential bargaining models.

findings. In theory, when the benefits from committing the harmful act are not greater than 660 and the risk-dominance refinement is applied, the deterrence rate under Mild Leniency is higher than the deterrence rate under Strong Leniency; and, the deterrence rates under Strong and Mild Leniency are higher than the deterrence rate under No Leniency. Our finding suggest that the deterrence rates across conditions are not significantly different. These results are explained by the low deterrence rates under Strong and Mild Leniency. As discussed, the low deterrence rates might indicate that some subjects considered alternative deterrence thresholds when making their decisions about committing the act, due, for instance, to self-serving bias.

In sum, our experimental findings demonstrate the effectiveness of ordered-leniency policies as detection mechanisms. In particular, our results indicate that the implementation of either Strong or Mild Leniency policies significantly increased the likelihood of reporting by one or both injurers. Importantly, our findings suggest that the majority of subjects chose the risk-dominant N.E. under the Mild Leniency policy. We provide empirical evidence of a "race to the courthouse" effect of ordered-leniency policies: Immediate self-reporting is observed when Strong or Mild Leniency policies are implemented.

6 Extensions

This section discusses several relevant extensions to our benchmark model. We first consider a setting where groups involve multiple potential injurers. Second, we study an environment with high-harm activities. Then, we consider two additional extensions. The first setting relaxes the assumption of deterministic probabilities of detection by considering an environment where the detection rates depend on the enforcement effort in a stochastic way. The second environment allows for asymmetric benefits across injurers.

6.1 Groups with Multiple Members

Our key insights extend to settings involving groups with multiple injurers. Suppose $n > 2$ is the number of potential injurers, and let p_i for $i = 0, 1, ..., n-1$ be the probability that any given injurer will be detected and fined if exactly i injurers have self-reported. We assume that $0 \leq p_0 < p_1 < .... < p_{n-1} < 1$, so self-reporting by an injurer raises the probability that the silent injurers will be apprehended. The injurers decide simultaneously whether to participate in the act and, if the act is committed, decide whether to report and the time of reporting, $t \in [0, 1]$.

As before, multiple equilibria may arise in the Stage 2 self-reporting subgame. We will restrict attention to coalition-proof Nash equilibria – CPNE (Bernheim et al., 1987).[67]

As in our benchmark model, an ordered-leniency policy $\boldsymbol{r} = (r_1, r_2, ..., r_n)$ that grants a reduced fine for the first position in the queue and possibly leniency for latter positions as well, $r_1 < r_2 < ... \leq r_n \leq 1$, can lead to both faster detection and stronger deterrence. Consider first the time-to-report decision. Since higher positions in the queue receive greater leniency, waiting to report the act is a weakly-dominated strategy. So, in equilibrium, any injurer who chooses to report the act will do so immediately (i.e., choose $t = 0$). In other words, an ordered-leniency policy will create a race to the courthouse. Note also that if exactly m injurers report the act at time $t = 0$, then the first m positions in the queue are randomly allocated and each of the m injurers receives a payoff of $b - \frac{1}{m}\sum_{i=1}^{m} r_i f$.

Next, consider the injurers' decisions about whether to report the act in Stage 2. In an equilibrium where all injurers self-report, it must be the case that

$$\frac{1}{m}\sum_{i=1}^{m} r_i \leq p_{m-1}, \tag{4}$$

for all $m = 1, 2, ..., n$. If this condition holds, no *individual* would want to deviate since the average fine from self-reporting, $\frac{1}{n}\sum_{i=1}^{n} r_i$, is smaller than average fine from remaining silent, p_{n-1}. A *coalition of two injurers* would not deviate either. If one of the coalition members expected the other coalition member to remain silent, that coalition member would prefer to join the $n-2$ self-reporters since $\frac{1}{n-1}\sum_{i=1}^{n-1} r_i \leq p_{n-2}$ according to condition (4). Following this basic logic, no coalition of any size can deviate in a way that is mutually self-enforcing. As shown in the appendix, condition (4) is necessary as well as sufficient for self-reporting to be a CPNE.

When designing an ordered-leniency policy, $\boldsymbol{r}$ the enforcement agency will seek to maximize the average fine paid by the injurers, $\frac{1}{n}\sum_{i=1}^{n} r_i$, subject to the constraints that $\frac{1}{m}\sum_{i=1}^{m} r_i \leq p_{m-1}$ and $r_m \leq 1$ for all $m = 1, 2, ..., n$. The solution to this program will create the strongest possible deterrence of the illegal activity. We denote this policy as "optimal ordered-leniency policy."

Proposition 5. *With an optimal ordered-leniency policy, all injurers self-report immediately in*

[67] An outcome is self-enforcing if and only if no proper subset (coalition) of players can deviate in a way that makes all of its members better off. The CPNE refinement captures the concept of efficient self-enforcing outcomes for environments with more than two players: An outcome is a CPNE if and only if it is Pareto efficient within the class of self-enforcing outcomes. Finally note that the application of the Pareto- or risk-dominance refinements in two-player games with no communication implicitly assumes that the players agree on the refinement. The application of the CPNE refinement here follows a similar approach, and hence, communication is not required.

equilibrium. An ordered-leniency policy with $r_1 = p_0$ *and* $r_m = min\{mp_{m-1} - \sum_{i=1}^{m-1} r_i, 1\} > 0$ *for all* $m > 1$ *is a solution to the enforcement agency's problem and represents an optimal ordered-leniency policy.*

Our analysis demonstrates that the highest level of deterrence is achieved when all injurers who commit the act later self-report immediately, and receive successive discounts for self-reporting based on their positions in the self-reporting queue. In general, leniency for the first to report will not be full and the fine for the last to report may not be maximal. Next, we provide a numerical example to illustrate our main insights.

Example. Suppose that the group includes three injurers, $n = 3$, $\bar{f} = 1$, and $(p_0, p_1, p_2) = (.2, .4, .5)$. Suppose that leniency is granted only to the first injurer to report the act, $\boldsymbol{r} = (.2, 1, 1)$. In equilibrium, only one injurer self-reports and the average fine is .33.[68] The enforcement agency can increase deterrence by also giving leniency to the second and third injurers to report the act. The optimal ordered-leniency policy is $\boldsymbol{r} = (.2, .6, .7)$. In equilibrium, all injurers self-report immediately. The average fine is .5.[69]

6.2 Additional Extensions

Stochastic Detection Rate

In our benchmark model, we assume that the social planner could perfectly control the probabilities of detection, p_0 and p_1, via its enforcement effort e. Injurers, when deciding whether to commit the harmful act, know exactly what these probabilities are, and therefore can accurately forecast their future self-reporting decisions. In the second-best enforcement mechanism, injurers who decide to commit the act in Stage 1 later decide to self-report in Stage 2 (see Proposition 3). Thus, in our baseline model, self-reporting of harmful acts is ubiquitous. Our framework can be extended to allow for a stochastic detection rate.

Consider first our benchmark environment. Suppose that the inculpatory evidence is strong enough to convict a silent injurer with almost certainty. Then, $p_1 = 1 - \varepsilon$, where $\varepsilon > 0$ is an an arbitrarily small number.[70] Suppose also that all the other assumptions of our benchmark model hold. Recall that the probability of detection in the absence of self-reporting is $p_0 = e$. Following

[68] The likelihood of detection is $p_1 = .4$. Then, the average fine is $(.2 + .4 + .4)/3 = 1/3$.

[69] The average fine is $(.2 + .6 + .7)/3 = 1/2$.

[70] For simplicity, and without loss of generality, we abstract from ε for the rest of the analysis.

our main analysis, the maximal deterrence will be obtained with multipliers $(r_1^S, r_2^S) = (e, 1)$ and $(r_1^M, r_2^M) = \left(\frac{1+2e}{3}, 1\right)$, for the Pareto- and risk-dominance refinements, respectively. Then, the act will be deterred if $b \leq \hat{b}^S(e) = \left(\frac{1+e}{2}\right)\bar{f}$ and $b \leq \hat{b}^M(e) = \left(\frac{2+e}{3}\right)\bar{f}$, for the Pareto- and risk-dominance refinements, respectively (see Proposition 3).

Now suppose that the detection rate p_0 is stochastic. Specifically, after the injurers commit the act, p_0 is drawn from common-knowledge density $q(p_0; e)$ on the unit interval where the median value is e (the enforcement effort of the agency). Holding the leniency multipliers, $(r_1^i, r_2^i), i = S, M$, fixed as described above, if $p_0 < e$ (i.e., if detection is relatively unlikely), then the injurers will both remain silent in Stage 2 and not report the act, and will pay a sanction $p_0\bar{f}$. If instead $p_0 > e$ (i.e., if detection is relatively likely), then the injurers will choose to self-report in Stage 2 and will pay an expected sanction $\hat{b}^i(e), i = S, M$.

In Stage 1, before learning the realization of the random variable p_0, the injurers must decide whether to commit the act. They are deterred from committing the act when

$$b < \int_0^e p_0\bar{f}q(p_0; e)dp_0 + \int_e^1 \hat{b}^i(e)q(p_0; e)dp_0. \tag{5}$$

Note that the deterrence threshold in this stochastic environment (right-hand side of the inequality) is smaller than $\hat{b}^i(e)$, the deterrence threshold with a certain detection rate. Then, having an uncertain detection rate compromises deterrence in Stage 1. Intuitively, when p_0 is stochastic with a median value of e rather than a deterministic value of e, the potential injures benefit from the option of not reporting the act when the probability of detection is small ($p_0 < e$) but do not experience any loss when the probability of detection is large ($p_0 > e$). As a result, the deterrence threshold is lower, and hence, harmful acts are committed more frequently in stochastic environments. As demonstrated earlier, deterrence is at its socially optimal level when the harm is not too high (see Proposition 3). Hence, social welfare will be unambiguously lower in environments with stochastic detection rates.[71]

Asymmetric Benefits to Group Members

In our benchmark framework, we assume that the two injurers derive the same private benefit from committing the harmful act. Injurers might not be always symmetric.[72] Our model can be extended to allow for asymmetric benefits to group members.

[71] As in our benchmark model, the optimal enforcement effort and the leniency multipliers that maximize deterrence will depend on a variety of factors including the characteristics of the densities $q(p_0; e)$ and $g(b)$.

[72] For instance, asymmetries might arise in environments where one injurer is the mastermind who conceives and plans the harmful act, and recruits others to help him commit the act. Then, the mastermind is the residual

Suppose that b_1 and b_2 are drawn from a joint density $\phi(b_1, b_2)$ and so the benefit to the first injurer, b_1, could be larger than or smaller than the benefit to the second injurer, b_2. Suppose also that all the other assumptions of our benchmark model hold. The act is socially desirable if the sum of the private benefits of the act exceed the total harm, $b_1 + b_2 > 2h$ or equivalently $(b_1 + b_2)/2 > h$, and socially undesirable otherwise. When transfer payments between the two injurers are impossible, the act will be committed when both potential injurers are willing to participate: $\min(b_1, b_2) \geq \hat{b}$, where $\hat{b}$ is the expected sanction. Interestingly, this new environment may feature overdeterrence of certain socially beneficial acts. To see why, consider an act where the private benefit to the first injurer is very high, $b_1 > 2h$, and the benefit to the second injurer is zero, $b_2 = 0$. This act is socially desirable, since $b_1 + b_2 \geq 2h$, but the act will not be committed for any positive expected sanction $\hat{b} > 0$. The second potential injurer will simply refuse to participate.

With side payments, the potential injurers will commit the act when their joint benefit exceeds the joint expected sanction, $b_1 + b_2 \geq 2\hat{b}$.[73] Hence, our earlier results will carry over to this enriched setting with bargaining at Stage 1. To illustrate this point, consider the environment presented in the previous paragraph. The first potential injurer who anticipates receiving $b_1 > 2h$ can pay the second potential injurer with $b_2 = 0$ to participate in the act as well. Finally note that, since $(b_1 + b_2)/2 \geq \min(b_1, b_2)$, the potential injurers will commit the act for a broader range of values when bargaining is possible.

Although environments involving groups with multiple injurers, high-harm activities, stochastic detection rates, or asymmetric benefits to group members obviously raise some new issues, the main insights derived from our benchmark model and the implications for the design of optimal enforcement policies remain relevant.

7 Summary and Conclusions

This paper studies the optimal design of enforcement schemes with ordered-leniency policies for detecting and preventing harmful short-term activities conducted by groups of injurers. We show that ordered-leniency policies that generate maximal deterrence give successively larger discounts to injurers who secure higher positions in the reporting queue, creating a so-called "race to the courthouse" among the members of the group of injurers. In equilibrium, all injurers report the

claimant of the benefits of the act, while the accomplices are just hired hands.

[73] Note that this environment does not allow side payments to depend on future self-reporting.

act immediately.

We provide a social welfare rationale for the current use of ordered-leniency policies in the real world. Our analysis demonstrates that the socially-optimal level of deterrence can be obtained at zero cost with an enforcement policy with ordered leniency when the externalities associated with the harmful activities are not too high. In contrast, enforcement policies that do not grant leniency for self-reporting cannot implement the first-best outcome when the set of fines is bounded. As a consequence, underdeterrence and costly enforcement are observed. Hence, enforcement policies with ordered leniency are superior to no-leniency policies.

Our experimental findings suggest that ordered-leniency policies are effective detection mechanisms. In particular, our results indicate that the implementation of either Strong or Mild Leniency policies significantly increased the likelihood of self-reporting by one or both injurers. Importantly, our findings suggest that the majority of subjects chose the risk-dominant N.E. under the Mild Leniency policy. We provide empirical evidence of a "race to the courthouse" effect of ordered-leniency policies: Immediate self-reporting is observed when Strong or Mild Leniency policies are implemented. Interestingly, our results on deterrence indicate that some subjects considered an alternative deterrence threshold when making their decisions about committing the harmful act. These findings might suggest the presence of self-serving bias on subjects' beliefs about getting the first position in the self-reporting queue. As a result, the deterrence power of ordered-leniency policies was weakened, and harmful acts were committed more frequently.

Several relevant extensions are discussed. First, we consider an environment with multiple potential injurers and show that our main insights also hold in this setting. Second, we explore an environment where where the detection rate depends on the enforcement effort in a stochastic way. In this setting, injurers who commit the act may refrain from self-reporting if the probabilities of detection are sufficiently low. As a result, deterrence might be compromised. Third, we discuss a setting that allows for asymmetric benefits from committing a harmful act across injurers. In this setting, the equilibrium outcomes heavily depend on the ability of group members to write side contracts with each other and negotiate transfer payments. Our earlier results carry over when monetary transfers are possible. These and other extensions are fruitful topics for future research.

References

Andreoni, James. 1991. "The Desirability of a Permanent Tax Amnesty." *Journal of Public Economics*, 45: 143–159.

Arlen, Jennifer and Reinier Kraakman. 1997. "Controlling Corporate Misconduct: An Analysis of Corporate Liability Regimes." *New York University Law Review,* 72: 687–779.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased Judgments of Fairness in Bargaining." *American Economic Review*, 11:109–26.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76: 169–217.

Bernheim, Douglas B., Bezalel Peleg, and Michael D. Whinston. 1987. "Coalition Proof Nash Equilibria I: Concepts." *Journal of Economic Theory*, 42: 1–12.

Bigoni, Maria, Sven-Olof Fridolfsson, Chloe Le Coq, and Giancarlo Spagnolo. 2012. "Fines, Leniency, and Rewards in Antitrust." *RAND Journal of Economics*, 43: 368–90.

Buccirossi, Paolo and Giancarlo Spagnolo. 2006. "Leniency Policies and Illegal Transactions." *Journal of Public Economics*, 90: 1281–1297.

Burton, Anthony and Martin Sefton. 2004. "Risk, Pre-Play Communication and Equilibrium." *Games and Economic Behavior*, 46: 23–40.

Ceresney, Andrew. 2015. "The SEC's Cooperation Program: Reflections on Five Years of Experience." http://www.sec.gov./news/speech/sec-cooperation-program.html.

Che, Yeon-Koo, and and Seung-Weon Yoo. 2001. "Optimal Incentives for Teams." *American Economic Review,* 91: 525–541.

Chen, Zhijun, and Patrick Rey. 2013. "On the Design of Leniency Programs." *Journal of Law and Economics*, 56: 917–957.

Cooper, Russell W., Douglas V. DeJong, D.V., Robert, Forsythe, and Thomas W. Ross. 1992. "Communication in Coordination Games." *Quarterly Journal of Economics*, 107: 739–771.

FBI. 2012. "Financial Crimes Report 2010–2011." https://www.fbi.gov/stats-services/publications/financial-crimes-report-2010-2011.

Feess, Eberhardt, and Markus Walzl. 2010. "Evidence Dependence of Fine Reductions in Corporate Leniency Programs," *Journal of Institutional and Theoretical Economics*, 166: 573-590.

Feess, Eberhardt, and Markus Walzl. 2004. "Self-Reporting in Optimal Law Enforcement When There Are Criminal Teams," *Economica*, 71: 333-348.

Grossman, Gene M. and Michael L. Katz. 1983. "Plea Bargaining and Social Welfare." *American Economic Review*, 73: 749–757.

Harrington, Joseph E. 2013. "Corporate Leniency Programs When Firms Have Private Information: The Push of Prosecution and the Pull of Pre-Emption." *Journal of Industrial Economics*, 51: 1–27.

Harsanyi, John C. and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge: MIT Press.

Innes, Robert. 1999. "Remediation and Self-reporting in Optimal Law Enforcement." *Journal of Public Economics*, 72, 379-393.

Johnson, Eric J., Colin Camerer, Sankar Sen, and Talya Rymon. 2002. "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining." *Journal of Economic Theory*, 104: 16–47.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47: 263–291.

Kaplow, Louis and Steven Shavell. 1994. "Optimal Law Enforcement with Self-Reporting of Behavior." *Journal of Political Economy*, 102: 583–606.

Kobayashi, Bruce. 1992. "Deterrence with Multiple Defendants: An Explanation for "Unfair" Plea Bargains." *RAND Journal of Economics*, 23, 507–517.

Kornhauser, Lewis A. and Richard L. Revesz. 1994. "Multidefendant Settlements under Joint and Several Liability: The Problem of Insolvency." *Journal of Legal Studies*, 23: 517–542.

Kraakman, Reinier H. 1986. "Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy." *Journal of Law, Economics & Organization*, 2, 53–104.

Kunda, Ziva. 1990. "The Case of Motivated Reasoning." *Psychological Bulletin,* 108: 480–98.

Kunda, Ziva. 1987. "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories." *Journal of Personality and Social Psychology,* 53: 636–47.

Landeo, Claudia M. 2018. "Law and Economics and Tort Litigation Institutions: Theory and Experiments." In K. Zeiler and J. Teitelbaum, eds., *Research Handbook on Behavioral Law and Economics*. Cheltenham, UK: Edward Elgar Publishing.

Landeo, Claudia M. 2009. "Cognitive Coherence and Tort Reform." *Journal of Economic Psychology*, 6: 898–912.

Landeo, Claudia M. and Kathryn E. Spier. 2015. "Incentive Contracts for Teams: Experimental Evidence." *Journal of Economic Behavior and Organization*, 119: 496–511.

Landeo, Claudia M. and Kathryn E. Spier. 2012. "Exclusive Dealing and Market Foreclosure: Further Experimental Results." *Journal of Institutional and Theoretical Economics*, 168: 150–170.

Landeo, Claudia M. and Kathryn E. Spier. 2009. "Naked Exclusion: An Experimental Study of Contracts with Externalities." *American Economic Review*, 99: 1850–1877.

Landeo, Claudia M., Maxim Nikitin, and Sergei Izmalkov. 2013. "Incentives for Care, Litigation, and Tort Reform under Self-Serving Bias." In T. Miceli and M. Baker, eds., *Research Handbook on Economic Models of Law.*

Landes, William M. 1971. "An Economic Analysis of the Courts." *Journal of Law and Economics*, 14: 61–108.

Livernois, John and C.J. McKenna. 1999. "Truth or Consequences: Enforcing Pollution Standards with Self-Reporting." *Journal of Public Economics,* 71: 415–440.

Malik, Arun S. 1993. "Self-Reporting and the Design of Policies for Regulating Stochastic Pollution." *Journal of Environmental Economics and Management,* 24: 241–257.

Malik, Arun S. and Robert M. Schwab. 1991. "The Economics of Tax Amnesties." *Journal of Public Economics*, 46: 29–49.

Motta, Massimo and Michele Polo. 2003. "Leniency Programs and Cartel Prosecution." *International Journal of Industrial Organization,* 21: 347–379.

Ochs, Jack. 1995. "Coordination Problems." In *Handbook of Experimental Economics*, ed. John H. Kagel and Alvin E. Roth. New Jersey: Princeton University Press Inc.

Polinsky, A. Mitchell and Steven Shavell. 1984. "The Optimal Use of Fines and Imprisonment" *Journal of Public Economics*, 24: 89–99.

Reinganum, Jennifer F. 1988. "Plea Bargaining and Prosecutorial Discretion." *American Economic Review,* 78: 713–728.

Spagnolo, Giancarlo and Catarina Marvão. 2016. "Cartels and Leniency: Taking Stock of What We Learnt." In *Handbook of Game Theory and Industrial Organization.* Edward Elgar Publishing.

Spagnolo, Giancarlo. 2005. "Divide et Impera: Optimal Leniency Programs." Mimeo, Stockholm School of Economics.

Spier, Kathryn E. 1994. "A Note on Joint and Several Liability: Insolvency, Settlement, and Incentives." *Journal of Legal Studies,* 23: 559–568.

Appendix

This Appendix presents formal proofs of the lemmas and propositions.

Proof of Lemma 1. Denote the strategy of player j as $\sigma_j = (\rho_j, t_j)$ where $\rho_j \in \{R, NR\}$ is whether to report the act and $t_j \in [0,1]$ is when to report the act. Suppose $r_1 < r_2$. If $\sigma_{-j} = (NR, t_{-j})$, then player j is indifferent about their reporting time, $(R,0) \sim (R,t_j)\ \forall t_j \in (0,1]$. If $\sigma_{-j} = (R, t_{-j})$, then for player j we have $(R,0) \sim (R,t_j)\ \forall t_j < t_{-j}$ and $(R,0) \succ (R,t_j)\ \forall t_j \geq t_{-j}$. Therefore $(R,0)$ weakly dominates $(R,t_j)\ \forall t_j \in (0,1]$ when $r_1 < r_2$. Suppose instead that $r_1 > r_2$. f $\sigma_{-j} = (NR, t_{-j})$, then player j is indifferent, $(R,1) \sim (R,t_j)\ \forall t_j \in [0,1)$. If $\sigma_{-j} = (R, t_{-j})$, then $(R,1) \sim (R,t_j)\ \forall t_j > t_{-j}$ and $(R,1) \succ (R,t_j)\ \forall t_j \leq t_{-j}$. Therefore $(R,1)$ weakly dominates $(R,t_j)\ \forall t_j \in [0,1)$ when $r_1 > r_2$. If $r_1 = r_2$ then there is no advantage to being first or second and so the players are indifferent as to the reporting time. ■

Proof of Lemma 2. In Case 1, $b - r_1 f \geq b - p_0 f$ and $b - \left(\frac{r_1+r_2}{2}\right) f \geq b - p_1 f$. With the tie-breaking assumption, self-reporting is a dominant strategy and (R,R) is the unique Nash equilibrium (NE). In Case 4, $b - r_1 f < b - p_0 f$ and $b - \left(\frac{r_1+r_2}{2}\right) f < b - p_1 f$ so not reporting is a dominant strategy and (NR,NR) is the unique NE. In Case 2, $b - r_1 f < b - p_0 f$ and $b - \left(\frac{r_1+r_2}{2}\right) f \geq b - p_1 f$ so (R, NR) and (NR, R) are both pure-strategy NE. In Case 3 there are two pure-strategy NE, (R, R) and (NR, NR). (R, R) Pareto-dominates (NR, NR) if $b - \left(\frac{r_1+r_2}{2}\right) f \geq b - p_0 f$ or $\frac{r_1+r_2}{2} \leq p_0$. (R, R) risk-dominates (NR, NR) if the former is preferred by player j if player $-j$ is randomizing 50/50 between R and NR, or $\frac{1}{2}(b - r_1 f) + \frac{1}{2}\left(b - \left(\frac{r_1+r_2}{2}\right) f\right) \geq \frac{1}{2}(b - p_0 f) + \frac{1}{2}(b - p_1 f)$, or $\frac{3r_1+r_2}{4} \leq \frac{p_1+p_1}{2}$. ■

Proof Lemma 3. Consider the four cases included in Lemma 2. In Case 1, (R, R) is the unique NE and each injurer receives a payoff of $b - \left(\frac{r_1+r_2}{2}\right) f$. It is therefore a weakly dominant strategy for an injurer to participate in the act if $b > \left(\frac{r_1+r_2}{2}\right) f$. In Case 2, (R, NR) and (NR, R) are both pure-strategy NE with an average payoff of $b - \left(\frac{r_1+p_1}{2}\right) f$. The act is committed when $b > \left(\frac{r_1+p_1}{2}\right) f$. In Case 3 three are two NE, (R, R) and (NR, NR). The act is committed if $b > p_0 f$ or $b > \left(\frac{r_1+r_2}{2}\right) f$, depending on which of the two equilibria is expected to prevail. Finally, in Case 4, (R, NR) is the unique pure-strategy NE and the act is committed if $b > p_0 f$. ■

Proof of Proposition 3. First, by Lemma 1, since $r_1^j < r_2^j$ for $j = S, M$, all reporting takes place at $t = 0$.

Second, we characterize the expected fine for each of the four cases included in Lemma 2, and identify the maximal expected fines.

Case 1. Both injurers self-report in this case. We now characterize the values (r_1, r_2) that maximize the expected fine$\left(\frac{r_1+r_2}{2}\right) f$ subject to the constraints that (i) $\frac{r_1+r_2}{2} \leq p_1$, (ii) $r_1 \in [0, p_0]$, and (iii) $r_2 \in [0,1]$. Two sub-cases are considered.

Case 1.1 The first case refers to $p_1 \leq \frac{1+p_0}{2}$. If $p_1 \leq \frac{1+p_0}{2}$, then constraint (i) must hold with equality, $\frac{r_1+r_2}{2} = p_1$. Suppose not: $\frac{r_1+r_2}{2} < p_1$. This would imply that both $r_1 = p_0$ and $r_2 = 1$, for

otherwise the expected fine $\left(\frac{r_1+r_2}{2}\right) f$ could be increased. Then, $\frac{r_1+r_2}{2} = \frac{1+p_0}{2} < p_1$, a contradiction. Therefore $\frac{r_1+r_2}{2} = p_1$. We can write $(r_1, r_2) = (p_1 - \Delta, p_1 + \Delta)$, where Δ is a constant. Since $r_1 \in [0, p_0]$, it must be that $p_1 - p_0 \leq \Delta \leq p_1$. Since $r_2 \in [0, 1]$, it must be that $-p_1 \leq \Delta \leq 1 - p_1$. Taken together, $\Delta \in [p_1 - p_0, \min\{p_1, 1 - p_1\}]$. $p_1 \leq \frac{1+p_0}{2}$ implies that $p_1 - p_0 \leq \min\{p_1, 1 - p_1\}$, so this range exists. *The expected fine is* $p_1 f$.

Case 1.2. The second case refers to $p_1 > \frac{1+p_0}{2}$. If $p_1 > \frac{1+p_0}{2}$, then constraint (i) does not bind at the optimum: $\frac{r_1+r_2}{2} < p_1$. Suppose not: $\frac{r_1+r_2}{2} = p_1$, Then, as above we would have $(r_1, r_2) = (p_1 - \Delta, p_1 + \Delta)$, where $\Delta \in [p_1 - p_0, \min\{p_1, 1 - p_1\}]$. But $p_1 > \frac{1+p_0}{2}$ implies $2p_1 > 1 + p_0$, which implies further that $p_1 - p_0 > \min\{p_1, 1 - p_1\}$. So no such value for Δ exists. Therefore $\frac{r_1+r_2}{2} < p_1$. It must also be true that $(r_1, r_2) = (p_0, 1)$. If $r_1 < p_0$ and/or $r_2 < 1$, then the expected fine would be higher (and no constraints violated) if r_1 and/or r_2 were raised. *The expected fine is* $\left(\frac{1+p_0}{2}\right) f < p_1 f$.

Case 2. Since only one injurer self-reports, the expected fine is $\left(\frac{r_1+p_1}{2}\right) f$. Since r_1 is constrained to be less than or equal to p_0 in this case, the strongest possible deterrence is obtained when $r_1 = p_0$. So *the expected fine is less than or equal to* $\left(\frac{p_0+p_1}{2}\right) f$. This expected fine is strictly lower than the expected fine in Case 1.

Case 3. There are multiple equilibria in this case.

With *Pareto dominance*, the injurers self-report if and only if $\frac{r_1+r_2}{2} \leq p_0$. *The expected fine is less than or equal to* $p_0 f$. This expected fine is always strictly lower than the expected fine in Case 1.

With *risk dominance*, the enforcer maximizes $\frac{r_1+r_2}{2}$ subject to the constraints that (i) $\frac{3r_1+r_2}{4} \leq \frac{p_0+p_1}{2}$, (ii) $r_1 \in [p_0, 1]$, and (iii) $r_2 \in [0, 1]$. Holding r_1 fixed, deterrence is increased by raising r_2 to the point where constraint (i) or constraint (iii) binds. Given r_1, we must have $r_2 = \min\{2(p_0 + p_1) - 3r_1, 1\}$. The enforcer's problem can be represented as choosing $r_1 \in [p_0, 1]$ to maximize $\frac{r_1+\min\{2(p_0+p_1)-3r_1,1\}}{2}$. Two sub-cases are considered.

Case 3.1 The first case refers to *risk dominance* and $p_1 \leq \frac{1+p_0}{2}$. If $p_1 \leq \frac{1+p_0}{2}$, then $2p_1 \leq 1 + p_0$. This implies that $2(p_0+p_1)-3r_1 \leq 1-3(r_1-p_0) \leq 1$, for all $r_1 \in [p_0, 1]$. So $\min\{2(p_0+p_1)-3r_1, 1\} = 2(p_0+p_1)-3r_1$, and the expected fine is $(p_0+p_1-r_1)f$ for all $r_1 \in [p_0, 1]$. Deterrence is maximized by making r_1 as small as possible, so $r_1 = p_0$ and $r_2 = 2(p_0+p_1)-3r_1 = 2p_1-p_0$, and *the expected fine is* $p_1 f$. This expected fine is the same as the expected fine in Case 1.

Case 3.2 The second case refers to *risk dominance* and $p_1 > \frac{1+p_0}{2}$. If $p_1 > \frac{1+p_0}{2}$, then r_1 will be strictly greater than p_0, and the expected fine strictly higher than $p_1 \bar{f}$. To see why this is true, suppose $r_1 = p_0 + \varepsilon$ where $\varepsilon > 0$. Since $p_1 > \frac{1+p_0}{2}$ implies $2p_1 > 1 + p_0$, we have $2(p_0+p_1)-3r_1 = 2p_1-p_0-3\varepsilon > 1$ when ε is not too large. Therefore $\min\{2(p_0+p_1)-3r_1, 1\} = 1$ when $r_1 = p_0+\varepsilon$ for $\varepsilon > 0$ sufficiently small. The expected fine in this case is $\left(\frac{r_1+1}{2}\right) f$. Deterrence would be higher if r_1 were raised above p_0. r_1 will be raised to the point where $2(p_0+p_1)-3r_1 = 1$ and so $r_1 = \frac{2(p_0+p_1)-1}{3}$ and $r_2 = 1$. *The expected fine is* $\left(\frac{1+p_0+p_1}{3}\right) f$. This expected fine is strictly higher than the expected fine in Case 1.

Case 4. Neither injurer self-reports. *The expected fine is* $p_0 f$. This expected fine is strictly lower than the expected fine in Case 1.

Hence, when *Pareto dominance* is applied in Case 3, the maximal expected fine always corresponds to Case 1. When *risk dominance* is applied in Case 3 and $p_1 \leq \frac{1+p_0}{2}$, the maximal expected fine corresponds to Case 1 or Case 3; when *risk dominance* is applied in Case 3 and $p_1 > \frac{1+p_0}{2}$, the maximal expected fine corresponds to Case 3.

Third, since the equilibria of the self-reporting subgame described in Lemmas 1 and 2 do not depend on the level of the fine, f, the highest deterrence is obtained with the maximal fine, $f = \bar{f}$. ■

Proof of Lemma 4. Proposition 3 implies (1) if $p_1 \leq \frac{1+p_0}{2}$, then $\hat{b}^S = \hat{b}^M = p_1\bar{f}$; and, (2) if $p_1 > \frac{1+p_0}{2}$, then $\hat{b}^S = \left(\frac{1+p_0}{2}\right)\bar{f}$, $\hat{b}^M = \left(\frac{1+p_0+p_1}{3}\right)\bar{f}$, and $\hat{b}^S < \hat{b}^M$. Substituting $p_0 = e$ and $p_1 = e + (1-e)\pi$ gives parts (1) and (2) of the lemma. ■

Proof of Proposition 4. First, the characterization of the first-best outcome follows immediately from the proofs of Proposition 3 and Lemma 4.

Second, the characterization of the fine and leniency multipliers implemented in the second-best outcome follow the proofs of Proposition 3 and Lemma 4.

Third, we demonstrate that the second-best outcome involves positive enforcement efforts. The social welfare function is given by:

$$W = \int_{\hat{b}^i(e,\pi)}^{\infty} (b-h)g(b)db - c(e),$$

where $\hat{b}^i(e,\pi)$, $i = S, M$, correspond to the deterrence thresholds under the Pareto-dominance and risk-dominance refinements, respectively. The enforcement agency chooses e to maximize social welfare. The first-order condition is:

$$(h - \hat{b}^i(e,\pi))\frac{\partial \hat{b}^i(e,\pi)}{\partial e} g(\hat{b}^i(e,\pi)) - c'(e) = 0.$$

As before, the first term represents the incremental benefit from increasing the probability e: $h - \hat{b}^i(e,\pi)$ is the social gain associated with deterring an additional harmful act, and $\frac{\partial \hat{b}(e,\pi)}{\partial e} g(\hat{b}^i(e,\pi))$ is the incremental volume of harmful acts that are deterred when the detection rate e increases. The second term, $c'(e)$, represents the marginal cost of effort. Rearranging terms, we find that the second-best optimal deterrence threshold (optimal expected fine) satisfies:

$$\hat{b}^i(e,\pi) = h - \frac{c'(e)}{\frac{\partial \hat{b}^i(e,\pi)}{\partial e} g(\hat{b}^i(e,\pi))}.$$

We need to show that the second-best outcome involves $e^i > 0$. Suppose not: $e^i = 0$. In this case, $h > \hat{b}^i(0,\pi)$ since by assumption the first-best enforcement policy cannot be obtained; $\frac{\partial \hat{b}^i(e,\pi)}{\partial e} > 0$ by Lemma 4; and $g(\hat{b}^i(0,\pi)) > 0$ since the density function has full support. Since $c'(0) = 0$, we have that the slope of the social welfare function is strictly positive when $e^i = 0$ and so we conclude that $e^i > 0$. Next, we show that $\hat{b}^i(e^i,\pi) < h$. Suppose instead that $\hat{b}^i(e^i,\pi) \geq h$. Since

$\frac{\partial \hat{b}^i(e,\pi)}{\partial e} g(\hat{b}^i(e,\pi)) > 0$, the slope of the welfare function would be strictly negative. Social welfare would be higher if e were reduced. ■

Proof of Proposition 5. Without loss of generality, we normalize the fine to unity, $f = 1$.

First, by Lemma 1, since $r_1 < r_2 < ... \leq r_n \leq 1$, all reporting takes place at $t = 0$.

Second, we prove that in an optimal ordered-leniency policy, $\boldsymbol{r}$, all injurers report. Suppose that in the optimal policy $\boldsymbol{r}$ that no injurer self-reports. The average expected fine is p_0. Consider an alternative policy with $r_1 = p_0 - \varepsilon$ where $\varepsilon > 0$ (small) and $r_i = 1$ for all $i \geq 2$. It is a CPNE for exactly one injurer to self-report. If one injurer self-reports, no remaining injurer (or coalition of injurers) would want to self-report since $r_i < p_{i-1}$ $\forall i \geq 2$. The expected fine is therefore $\frac{1}{n}[p_0 - \varepsilon + (n-1)p_1] > p_0$ and so deterrence is stronger. Now consider leniency policy $\boldsymbol{r}$ where exactly $m - 1 < n$ injurers self-report. The expected fine with leniency policy $\boldsymbol{r}$ is:

$$\frac{1}{n}\left(\sum_{i=1}^{m-1} r_i + \sum_{i=m}^{n} p_{m-1}\right).$$

Since injurer $m-1$ is willing to self-report, the expected fine from self-reporting must be weakly lower than the fine from remaining silent. So it must be the case that $\frac{1}{m-1}\sum_{i=1}^{m-1} r_i \leq p_{m-2}$. We will now show that the social player can increase deterrence by inducing injurer m to self-report as well. Consider a new leniency policy $\boldsymbol{r'}$ where $r'_i = r_i$ for $i \leq m-1$, $r'_m = p_{m-1} - \varepsilon$, and $r'_i = 1$ for $i \geq m+1$. Under $\boldsymbol{r'}$, it is a CPNE for injurers $i = 1, ..., m$ to self-report since

$$\frac{1}{m}\sum_{i=1}^{m} r'_i = \frac{1}{m}\left(\sum_{i=1}^{m-1} r'_i + r'_m\right) \leq \left(\tfrac{1}{m}\right)\left[(m-1)p_{m-2} + p_{m-1} - \varepsilon\right] < p_{m-1}.$$

The last step follows from the assumption that $p_{m-1} > p_{m-2}$ and $\varepsilon > 0$ is small. The expected fine under $\boldsymbol{r'}$ is

$$\frac{1}{n}\left(\sum_{i=1}^{m-1} r_i + (p_{m-1} - \varepsilon) + \sum_{i=m+1}^{n} p_m\right),$$

which is higher than the expected fine under $\boldsymbol{r}$. This concludes the proof that in an optimal ordered-leniency policy, all injurers self-report.

Third, in the main text, we showed that (4) is sufficient for self-reporting by all injurers to be a CPNE. We now demonstrate that (4) is also necessary. Suppose self-reporting by all n injurers is a CPNE. In the CPNE, no *individual* injurer is better off deviating: $\frac{1}{n}\sum_{i=1}^{n} r_i \leq p_{n-1}$. Now consider a deviation by a coalition of size $m' = n - m + 1$. Note that since $m' + (m-1) = n$, $m-1$ injurers are not part of the deviating coalition and continue to self-report. So, the m' injurers in the coalition would pay an expected fine of p_{m-1}. There are two cases to consider. (i) Suppose $\frac{1}{n}\sum_{i=1}^{n} r_i > p_{m-1}$, so a coalition of size m' is jointly better off not reporting. Since self-reporting by all n injurers is a CPNE, it must be the case that a deviation by this coalition is not self-enforcing. Thus, we require that an individual would prefer to abandon the coalition and join the group of $m-1$ injurers who self-report: $\frac{1}{m}\sum_{i=1}^{m} r_i \leq p_{m-1}$. This is condition (4). (ii) Suppose $\frac{1}{n}\sum_{i=1}^{n} r_i \leq p_{m-1}$, so a coalition of size m' is collectively worse off reporting. Since

$r_i \leq r_j$ for all $i < j$, this condition implies that $\frac{1}{m}\sum_{i=1}^{m} r_i \leq p_{m-1}$. This is condition (4). So in any CPNE where all injurers self-report, condition (4) is necessary and sufficient.

Fourth, we demonstrate that the characterization in Proposition 4 corresponds to the optimal ordered-leniency policy. For notational ease, define $\rho_n(\boldsymbol{r}) = min\{mp_{m-1} - \sum_{i=1}^{m-1} r_i, 1\}\ \forall m$. We show first that for any optimum, $\widehat{\boldsymbol{r}}$, there is an alternative leniency policy $\tilde{\boldsymbol{r}}$ that achieves the same level of deterrence and satisfies $\tilde{r}_m = \rho_m(\tilde{\boldsymbol{r}})$. If $\widehat{\boldsymbol{r}}$ is a solution to the social planner's program, then it must be the case that the injurer in the very last position, n, pays $\widehat{r}_n = \rho_n(\widehat{\boldsymbol{r}})$. Suppose not: $\widehat{r}_n < \rho_n(\widehat{\boldsymbol{r}})$. Can increase $\widehat{r}_n$ to $\widehat{r}_n + \Delta$ for Δ small. No constraints are violated, and the optimand is larger. Now suppose that $\rho_m(\widehat{\boldsymbol{r}}) - \widehat{r}_m = \Delta > 0$ for some $m < n$. So constraint m is slack. Consider an alternative policy $\tilde{\boldsymbol{r}} = (\widehat{r}_1, ..., \widehat{r}_{m-1}, \widehat{r}_m + \Delta, \widehat{r}_{m+1} - \Delta, \widehat{r}_{m+2}, ..., \widehat{r}_n)$. Note that this new policy $\tilde{\boldsymbol{r}}$ is identical to $\widehat{\boldsymbol{r}}$ for all elements except m and $m+1$. The value taken by the optimand under $\tilde{\boldsymbol{r}}$ is unchanged. Since $\rho_i(\tilde{\boldsymbol{r}}) = \rho_i(\widehat{\boldsymbol{r}})\ \forall i \neq n, m+1$, we need only check the constraints for m and $m+1$ are satisfied. $\tilde{r}_m = \widehat{r}_m + \Delta = \rho_m(\widehat{\boldsymbol{r}}) = \rho_m(\tilde{\boldsymbol{r}})$ by construction. So constraint m is satisfied by $\tilde{\boldsymbol{r}}$. $\rho_{m+1}(\tilde{\boldsymbol{r}}) = min\{(m+1)p_m - (\widehat{r}_1 + ... + \widehat{r}_m + \Delta), 1\} \geq \rho_{m+1}(\widehat{\boldsymbol{r}}) - \Delta$. Finally, we show that $\tilde{r}_{m+1} \leq \rho_{m+1}(\tilde{\boldsymbol{r}})$. $\tilde{r}_{m+1} = \widehat{r}_{m+1} - \Delta \leq \rho_{m+1}(\widehat{\boldsymbol{r}}) - \Delta \leq \rho_{m+1}(\tilde{\boldsymbol{r}})$. So constraint $m+1$ is satisfied by $\tilde{\boldsymbol{r}}$. ■

PLEASE GIVE THIS MATERIAL TO THE EXPERIMENTER AT THE END OF THE EXPERIMENT

INSTRUCTIONS

This is an experiment in the economics of decision-making. The National Science Foundation has provided the funds for this research.

In this experiment you will be asked to play an economic decision-making computer game. The experiment currency is the "token." The instructions are simple. If you follow them closely and make appropriate decisions, you may make a large amount of money. At the end of the session you will be paid your game earnings in CASH. If you have any questions at any time, please raise your hand and the experimenter will go to your desk.

PROBABILITY OR CHANCE

The concept of probability or chance will be used in this experiment. **PROBABILITY OR CHANCE (EXPRESSED IN PERCENTAGES)** indicates the likelihood of occurrence of uncertain events. The concept of probability or chance can be illustrated as follows. Suppose that an urn contains 100 balls: 20 out of 100 balls are white, and 80 out of 100 balls are black. Suppose that you randomly extract one ball from the urn. The chance that the ball will be white is equal to 20% because 20 out of 100 balls in the urn are white. Similarly, the chance that the ball will be black is equal to 80%, because 80 out of 100 balls in the urn are black.

SESSION AND PLAYERS

The session is made up of 6 matches. The first 5 matches are practice matches. After the last practice match, **ONE ACTUAL MATCH** will be played.

1) At the beginning of the session, every participant will be randomly assigned a role. The equally likely roles are: **Player B1** and **Player B2**.

   The **ROLES WILL REMAIN THE SAME until the end of the session**.

2) Before the beginning of **EACH PRACTICE MATCH**, the computer will randomly form pairs of **TWO PEOPLE**: **Player B1** and **Player B2**.

   YOU WILL NOT KNOW THE IDENTITY OF YOUR PARTNER.
   YOU WILL PLAY WITH A DIFFERENT PARTNER IN EVERY PRACTICE MATCH.

3) Before the beginning of the **ACTUAL MATCH**, the computer will randomly form pairs of **TWO PEOPLE**: **Player B1** and **Player B2**.

   YOU WILL NOT KNOW THE IDENTITY OF YOUR PARTNER.
   YOU WILL NOT PLAY WITH ANY OF YOUR PREVIOUS PARTNERS.

MATCH STAGES

STAGE 1: DECISION WHETHER TO JOINTLY COMMIT THE ACT

1) Each player has an **initial endowment** equal to **700 tokens**.

2) **THE DECISION TO JOINTLY COMMIT THE ACT** refers to an economic decision involving potential economic benefits and potential economic losses.

- **ECONOMIC BENEFITS** might occur in **STAGE 1**.
- **ECONOMIC LOSSES** might occur in **STAGE 2**.

3) **THE COMPUTER** randomly determines the **NUMBER OF TOKENS X** that each player will get **IF BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT.** Both players will receive the same number of tokens X.

- The number of tokens X can be equal to **200, …, 1598, 1599, 1600 tokens**.
- The number of tokens X will be revealed to both players.

4) **Player B1** and **Player B2** decide whether to **JOINTLY COMMIT THE ACT**

- Each player will have **60 SECONDS TO SEND MESSAGES TO THE OTHER PLAYER REGARDING WHETHER TO AGREE TO JOINTLY COMMIT THE ACT.**
    - Messages containing personal or identifying information are **NOT** allowed.
    - Messages including physical threats are **NOT** allowed.
- Then, each player will have **30 SECONDS TO DECIDE WHETHER TO AGREE TO JOINTLY COMMIT THE ACT OR NOT AGREE TO JOINTLY COMMIT THE ACT AND PRESS THE NEXT BUTTON.**
    - If a player **FAILS TO MAKE A CHOICE AND TO PRESS THE NEXT BUTTON WITHIN THE 30-SECOND PERIOD,** it will be implied that he/she decided **NOT TO AGREE TO JOINTLY COMMIT** the act.

5) The possible outcomes are as follows.

- **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT:** Each player gets **X TOKENS** (in addition to the initial endowment of 700 tokens) and **STAGE 2 BEGINS**.
- **ONLY ONE PLAYER AGREES TO JOINTLY COMMIT THE ACT:** Each player gets **ZERO TOKENS** and the **MATCH ENDS**. The match payoff for each player will be 700 tokens (initial endowment).
- **NEITHER PLAYER AGREES TO JOINTLY COMMIT THE ACT:** Each player gets **ZERO TOKENS** and the **MATCH ENDS**. The match payoff for each player will be 700 tokens (initial endowment).

STAGE 2: DECISION WHETHER TO REPORT THE ACT

1) If **both payers agreed to jointly commit the act**, then Stage 2 begins.

2) **A FINE EQUAL TO 900 TOKENS MIGHT BE DEDUCTED FROM A PLAYER'S TOKEN BALANCE AS A CONSEQUENCE OF JOINTLY COMMITTING THE ACT.**

   - A player's decision to report the act **MIGHT DECREASE THE FINE HE/SHE WILL PAY** from 900 tokens to 420 tokens.

   - A player's decision to report the act **MIGHT INCREASE HIS/HER PARTNER'S CHANCE TO PAY A FINE** from 40% to 90% or from 40% to 100%.

   THE SPECIFIC FINE AND THE CHANCE OF PAYING THAT FINE DEPEND ON THE DECISIONS OF BOTH PLAYERS ABOUT REPORTING THE ACT.

3) Each player will have **90 SECONDS TO DECIDE WHETHER TO REPORT OR NOT TO REPORT THE ACT AND PRESS THE NEXT BUTTON.**

   - If a player **FAILS TO MAKE A CHOICE AND TO PRESS THE NEXT BUTTON WITHIN THE 90-SECOND PERIOD,** it will be implied that he/she decided **NOT REPORT** the act.

4) The possible outcomes and match payoffs are presented below.

POSSIBLE OUTCOME 1: BOTH PLAYERS DECIDE NOT TO REPORT THE ACT

- **NEITHER PLAYER GETS A FINE REDUCTION**
- **EACH PLAYER'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 40%**: Each player pays a fine equal to **900 tokens** with **40% chance** and **does not pay any fine** with **60% chance**.

Hence, the match payoffs are as follows.

With a **40% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 900 tokens
Player B2's match payoff = 700 tokens + X tokens – 900 tokens

With a **60% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 0 tokens
Player B2's match payoff = 700 tokens + X tokens – 0 tokens

POSSIBLE OUTCOME 2: ONLY PLAYER B1 DECIDES TO REPORT THE ACT

- **ONLY PLAYER B1 GETS A FINE REDUCTION**: Instead of paying a fine equal to 900 tokens, Player B1 **always** pays only **420 tokens**.
- **PLAYER B2'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 90%**: Player B2 pays a fine equal to **900 tokens** with **90% chance** and **does not pay any fine** with **10% chance**.

Hence, the match payoffs are as follows.

With a **90% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 420 tokens
Player B2's match payoff = 700 tokens + X tokens – 900 tokens

With a **10% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 420 tokens
Player B2's match payoff = 700 tokens + X tokens – 0 tokens

POSSIBLE OUTCOME 3: ONLY PLAYER B2 DECIDES TO REPORT THE ACT

- **ONLY PLAYER B2 WILL GET A FINE REDUCTION**: Instead of paying a fine equal to 900 tokens, Player B2 **always** pays only **420 tokens**.
- **PLAYER B1'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 90%**: Player B1 pays a fine equal to **900 tokens** with **90% chance** and **does not pay any fine** with **10% chance**.

Hence, the match payoffs are as follows.

With a **90% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 900 tokens
Player B2's match payoff = 700 tokens + X tokens – 420 tokens

With a **10% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 0 tokens
Player B2's match payoff = 700 tokens + X tokens – 420 tokens

POSSIBLE OUTCOME 4: BOTH PLAYERS DECIDE TO REPORT THE ACT

- **IF PLAYER B1 REPORTS FIRST**
  - **ONLY PLAYER B1 GETS A FINE REDUCTION**: Instead of paying a fine equal to 900 tokens, Player B1 **always** pays only **420 tokens**.
  - **PLAYER B2'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 100%**: Player B2 **always** pays a fine equal to **900 tokens.**

Hence, the match payoffs are as follows.

With a **100% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 420 tokens
Player B2's match payoff = 700 tokens + X tokens – 900 tokens

- **IF PLAYER B2 REPORTS FIRST**
  - **ONLY PLAYER B2 GETS A FINE REDUCTION**: Instead of paying a fine equal to 900 tokens, Player B2 **always** pays only **420 tokens**.
  - **PLAYER B1'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 100%**: Player B1 **always** pays a fine equal to **900 tokens**.

Hence, the match payoffs are as follows.

With a **100% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 900 tokens
Player B2's match payoff = 700 tokens + X tokens – 420 tokens

- **IF BOTH PLAYERS REPORT AT THE SAME TIME**
  - **EACH PLAYER GETS A FINE REDUCTION WITH 50% CHANCE**: Instead of paying a fine equal to 900 tokens, each player pays only **420 tokens** with **50% chance**.
  - **EACH PLAYER'S CHANCE OF PAYING A FINE EQUAL TO 900 TOKENS IS 50%:** Each player pays a fine equal to **900 tokens** with **50% chance**.

Hence, the match payoffs are as follows.

With a **50% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 420 tokens
Player B2's match payoff = 700 tokens + X tokens – 900 tokens

With a **50% CHANCE,** the match payoffs will be:

Player B1's match payoff = 700 tokens + X tokens – 900 tokens
Player B2's match payoff = 700 tokens + X tokens – 420 tokens

EXERCISES

Suppose that the number of tokens that each player gets **IF BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT** is equal **to X tokens.**

Nine exercises, based on the possible outcomes, are presented below. Please fill the blanks.

Exercise 1

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **NOT TO REPORT** the act and **Player B2** decides **NOT TO REPORT** the act.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 2

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **TO REPORT** the act and **Player B2** decides **NOT TO REPORT** the act.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 3

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **NOT TO REPORT** the act and **Player B2** decides **TO REPORT** the act.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 4

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **TO REPORT** the act and **Player B2** decides **TO REPORT** the act, and that **Player B1 IS THE FIRST TO REPORT**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 5

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **TO REPORT** the act and **Player B2** decides **TO REPORT** the act, and that **Player B2 IS THE FIRST TO REPORT**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 6

Suppose that **BOTH PLAYERS AGREE TO JOINTLY COMMIT THE ACT**. Then, each player gets _____ tokens. Suppose also that **Player B1** decides **TO REPORT** the act and **Player B2** decides **TO REPORT** the act, and that **BOTH PLAYERS REPORT AT THE SAME TIME**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 7

Suppose that **Player B1 AGREES TO JOINTLY COMMIT THE ACT** and **Player B2 DOES NOT AGREE TO JOINTLY COMMIT THE ACT**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 8

Suppose that **Player B1 DOES NOT AGREE TO JOINTLY COMMIT THE ACT** and **Player B2 AGREES TO JOINTLY COMMIT THE ACT**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

Exercise 9

Suppose that **NEITHER PLAYER AGREES TO JOINTLY COMMIT THE ACT**.

The **MATCH PAYOFFS (IN TOKENS)** are as follows.

Player B1:

Chance	**Payoff**	**Chance**	**Payoff**

Player B2:

Chance	**Payoff**	**Chance**	**Payoff**

SESSION PAYOFF

The game earnings in tokens will be equal to the **PAYOFF FOR THE ACTUAL MATCH**. The game earnings in dollars will be equal to: (game earnings in tokens)/29 (29 tokens = 1 dollar). The session payoff will be equal to the game earnings in dollars plus the $10 participation fee.

GAME SOFTWARE

The game will be played using a computer terminal. You will need to enter your decisions by using the mouse. In some instances, you will need to wait until the other players make their decisions before moving to the next screen. Please **BE PATIENT**. There will be a box, displayed in the upper right-hand side of your screen, which indicates the "Match Number," "Your Role," and "Your Balance."

Please press the **NEXT >>** button to move to the next screen. **DO NOT TRY TO GO BACK TO THE PREVIOUS SCREEN AND DO NOT CLOSE THE BROWSER**: The software will stop working.

Next, the **5 PRACTICE MATCHES** will begin. After that, the **ACTUAL MATCH** will be played. **YOU CAN CONSULT THESE INSTRUCTIONS AT ANY TIME DURING THE EXPERIMENT**.

THANKS FOR YOUR PARTICIPATION IN THIS STUDY!!

PLEASE GIVE THIS MATERIAL TO THE EXPERIMENTER AT THE END OF THE EXPERIMENT