If we Confess our Sins*

Francisco Silva†

July 24, 2018

Abstract

I consider a scenario where a social planner suspects that a crime has been committed. There are many suspects and at most one of them is guilty. I characterize the optimal mechanism for the social planner under two different assumptions with respect to her commitment power: full commitment power and partial commitment power. I find that, in either case, the optimal mechanism is what I call a "confession inducing mechanism", where, before an investigation takes place, each agent has the opportunity to confess to being guilty in exchange for a reduced punishment. I find that these mechanisms do better than the traditional trial mechanism because of information externalities: when an agent credibly confesses his guilt, he reveals everyone else's innocence.

JEL classification: D82, K14

Keywords: leniency, criminal justice system, mechanism design, commitment power

*I would like to thank David Abrams, Mustafa Dogan, Selman Erol, Nicolas Figueroa, Nicholas Janetos, Sangmok Lee, George Mailath, Steven Matthews, Timofiy Mylovanov, Daniel Neuhann, Qiusha Peng, Andrew Postlewaite, Bernardo S. da Silveira, Rakesh Vohra and Yanhao Wei as well as the UPenn's Micro Lunch seminar participants for their useful comments.

†Department of Economics, Pontificia Universidad Catolica de Chile, Vicuña Mackenna 4860, Piso 3. Macul, Santiago, Chile. (email: franciscosilva@uc.cl).

1 Introduction

Imagine that there is the suspicion that a crime has been committed. There are N suspects (the agents) and at most one of them is guilty. Each agent knows only whether they are guilty or innocent. I consider the following question: "what is the best mechanism that a principal can design if her goal is to punish only the guilty agent (if any)?".

The traditional solution is a "trial" mechanism. In a trial mechanism, if the principal suspects a crime has been committed, she initiates an investigation aimed at obtaining evidence - some exogenous signal θ, correlated with the agents' guilt. Based on that evidence θ, the principal forms beliefs about the guilt of each agent and chooses punishments accordingly: each agent n receives a punishment of some $x_n = d_n(\theta)$. However, I argue in this paper that these systems are suboptimal for the principal: she will be able to do better by communicating with the agents prior to launching an investigation.

The communication takes place as follows: the principal gives each agent n, who is privately informed of his innocence/guilt, the opportunity to choose some message m_n from some "message set" M_n before evidence θ has been collected; then the principal starts an investigation which produces θ; and, finally, based on message vector $m = (m_1, ..., m_N)$ and on evidence θ, she chooses a punishment $x_n = d_n(m, \theta)$ for each agent n. Message set $M = M_1 \times ... \times M_N$ and function $d = (d_1, ..., d_N)$ form a "mechanism" (the trial mechanism is a "special" mechanism where punishments do not depend on the message m of the agents). The goal of the paper is to study which mechanisms (M, d) are optimal for the principal under two different assumptions with respect to the principal's ability to commit to a mechanism: full commitment power and partial commitment power.

The assumption of full commitment power is the standard one in mechanism design: the principal can commit to any mechanism (M, d) she chooses. In this setting, I find that the optimal mechanism is what I call a "confession inducing" mechanism (CIM). A CIM has two properties. The first property, which is implied by the revelation principle (Myerson (1979)), is that each agent's message set is only made of two messages, labeled c and $\overline{c}$, i.e., $M_n = \{c, \overline{c}\}$ for all n. Message c is interpreted as a confession, while message $\overline{c}$ is interpreted as a refusal to confess: in a CIM, each agent is able to confess to being guilty before an investigation takes place. The second

property is that, whenever an agent confesses, his punishment is independent of what other agents may have reported and of the evidence, i.e., $d_n(c, m_{-n}, \theta)$ is independent of m_{-n} and θ (where m_{-n} has the standard interpretation).

This description is reminiscent of two mechanisms that already exist in American law: "self-reporting", which exists in environmental law and "plea bargaining" in criminal law. The idea behind self-reporting in environmental law is that firms which infringe environmental regulations are able to contact the corresponding law enforcement authority and self-report this infringement in exchange for a smaller punishment than the one they would have received if they were later found guilty. In plea bargaining, defendants are given the chance to confess to having committed the crime in exchange for a reduced sentence.

In the literature, there have been papers that have highlighted two advantages of CIMs when compared to trial systems: Kaplow and Shavell (1994) show that CIMs can do better because they can reduce costs (in a setting with only a single agent, a credible confession eliminates the need for a costly investigation); and Grossman and Katz (1983) and Siegel and Strulovici (2018) show that CIMs can do better because they can reduce risk (for example, the risk of acquitting guilty agents). In the text, I characterize the optimal CIM under the assumption of full commitment power and show that it is preferred to the trial mechanism but for a different reason: information externalities. More precisely, there are information externalities that are generated by each credible confession: when an agent who is guilty confesses, that confession informs the principal that all other agents are innocent. Therefore, when the principal has to decide the punishments of the agents who did not confess, she will not make the mistake of punishing them. In other words, the fact that it is commonly known that there is at most one guilty agent implies that the agents' guilt is correlated, which is what generates these information externalities.[1]

The commitment power of the principal can be used to i) impose small punishments on agents who are believed to be guilty, and ii) to punish agents who are believed to be innocent. In order to implement the optimal CIM, the principal uses both of these: by the revelation principle, following the optimal CIM, each agent confesses if guilty but refuses if innocent. Therefore, the principal knows that in equilibrium every agent who confesses is guilty and every agent who refuses is innocent.

[1] A similar argument has been made to argue in favor of leniency policies as a way to reduce collusion. I review this literature on the literature review section.

So, the principal uses i) to be able to punish agents who have confessed less than what she would have preferred; and uses ii) to be able to punish agents who have not confessed more than what she would have wanted. Assumption ii) seems particularly problematic: it is hard to see how a mechanism that punishes agents who are known to be innocent could be implemented. In light of that, I consider the case of partial commitment power, which is defined to allow the principal to commit to i) but not to ii).[2]

When the principal has partial commitment power, it is as if she is unable to commit not to renegotiate: the principal cannot commit to punish those she believes are innocent because both her and the agent would be better off by renegotiating and selecting a smaller punishment. In this context, the problem of finding the optimal mechanism is more challenging seeing as, in general, the revelation principle need not follow (see Bester and Strausz (2000). The main technical contribution of this paper is to show that, even in this context, CIMs are optimal. The optimal renegotiation proof CIM is, however, different than the optimal CIM when the principal has full commitment power in two ways. First, it is such that, in equilibrium, agents do not completely separate, i.e., unlike the case with full commitment power, if the agent is guilty, he randomizes between confessing and not confessing (while the innocent never confesses). This means that, upon observing a refusal to confess, the principal is no longer certain that the agent is innocent, which allows her to impose some punishments on non-confessing agents, which is precisely what makes them willing to confess if guilty. The second difference is that punishments that follow refusals to confess are sequentially optimal, i.e., if the principal observes an agent refusing to confess, she will update her beliefs about the guilt of that agent using all information available at that time (every agents' message and the evidence) and impose what would be the optimal punishment given those beliefs.[3] In other words, unlike the previous mechanism, this CIM does not induce any regret by the principal towards agents who chose not to confess.

The paper is organized as follows. In the following section, I present a simple

[2]In appendix B, I also briefly discuss the case where the principal cannot commit to neither i) nor ii).

[3]The optimal renegotiation proof CIM induces an equilibrium where guilty agents randomize between confessing and refusing to confess, which is similar to what happens in some of the literature (Baker and Mezzetti (2001), Kim (2010), etc.). The key difference, however, is that, in this paper, I show that CIMs are actually optimal among all possible renegotiation proof mechanisms.

example to illustrate the main results of the paper. In section 3, I describe the model. In section 4, I study the case where the principal has full commitment power. In section 5, I consider the partial commitment power case. In section 6, I discuss the related literature and in section 7 I conclude.

2 Example

Consider a small town where, for simplicity, only $N = 2$ agents live. Imagine that there has been a fire which damaged the local forest. The principal suspects that it might not have been an accident, so that one of the agents might have set the fire: each agent n is either guilty ($t_n = g$) or innocent ($t_n = i$). Her prior belief is that there is a 40% chance that agent 1 is guilty, a 40% chance that agent 2 is guilty, and a 20% chance that none of the agents is guilty (so that the fire was an accident). Each agent knows only whether they are innocent or guilty.

The principal is able to conduct a (costless) investigation, which produces evidence θ: a random variable correlated with the agents' guilt. In particular, let us assume that $\theta \in \{0, 1\}$ and that

$$\Pr\{\theta = 1 | t = (i, i)\} = \frac{1}{2},\ \Pr\{\theta = 1 | t = (g, i)\} = \frac{2}{3},\ \Pr\{\theta = 1 | t = (i, g)\} = \frac{1}{3} \quad (1)$$

where $t = (t_1, t_2)$. For example, an investigation could be to go to the local forest and look for forensic evidence that links any of the agents to the fire.

If the principal decides on vector of punishments $x = (x_1, x_2) \in [0, 1]^2$, then agent n's payoff is simply $u(x_n) = -x_n$ for $n = 1, 2$, while the principal's payoff is $v(t, x) = v_1(t_1, x_1) + v_2(t_2, x_2)$, where

$$v_n(t_n, x_n) = \begin{cases} x_n \text{ if } t_n = g \\ -x_n \text{ if } t_n = i \end{cases}$$

for $n = 1, 2$. In words, each agent simply wants to minimize his expected punishment, while the principal wants to maximize the expected punishment of the agent if he is guilty, but minimize it if he is innocent.

Imagine that the principal implements a trial mechanism. A trial mechanism works in the following manner: the principal initiates a (costless) investigation that produces evidence θ, and then chooses punishments based on θ, i.e., for each $\theta \in \{0,1\}$, the principal chooses $d(\theta) \in [0,1]^2$. The optimal trial mechanism, denoted by d^{Tr}, is such that for $n = 1, 2$,

$$d_n^{Tr}(\theta) \in \arg\max_{x_n \in [0,1]} \left\{ \left(\Pr\{t_n = g|\theta\} - \Pr\{t_n = i|\theta\} \right) x_n \right\}$$

for all $\theta \in \{0,1\}$, which can be written as

$$d_n^{Tr}(\theta) = \begin{cases} 1 \text{ if } \Pr\{t_n = g|\theta\} \geq \frac{1}{2} \\ 0 \text{ if } \Pr\{t_n = g|\theta\} < \frac{1}{2} \end{cases}$$

for all $\theta \in \{0,1\}$.[4] In words, whenever there is a suspicion that a crime has been committed, the principal conducts an investigation which produces θ; then, based on θ, the principal updates her beliefs about each agent's guilt; if the principal finds that the agent is more likely to be guilty than not, she will punish him in 1; otherwise, the agent will be acquitted. Under (1), we have that

$$\Pr\{t_1 = g|\theta = 1\} = \Pr\{t_2 = g|\theta = 0\} = \frac{\frac{4}{10}\frac{2}{3}}{\frac{4}{10}\frac{2}{3} + \frac{4}{10}\frac{1}{3} + \frac{2}{10}\frac{1}{2}} = \frac{8}{15} \geq \frac{1}{2}$$

and

$$\Pr\{t_1 = g|\theta = 0\} = \Pr\{t_2 = g|\theta = 1\} = \frac{\frac{4}{10}\frac{1}{3}}{\frac{4}{10}\frac{1}{3} + \frac{4}{10}\frac{2}{3} + \frac{2}{10}\frac{1}{2}} = \frac{4}{15} < \frac{1}{2}$$

which implies that

$$d_1^{Tr}(\theta) = \begin{cases} 1 \text{ if } \theta = 1 \\ 0 \text{ if } \theta = 0 \end{cases} \text{ and } d_2^{Tr}(\theta) = \begin{cases} 0 \text{ if } \theta = 1 \\ 1 \text{ if } \theta = 0 \end{cases}$$

Given d^{Tr}, one can calculate the expected punishment of each agent, depending on his type: if agent n is guilty, his expected punishment is $\frac{2}{3}$, while if he is innocent, his expected punishment is

$$\frac{\frac{4}{10}\frac{1}{3} + \frac{2}{10}\frac{1}{2}}{\frac{6}{10}} = \frac{7}{18}$$

[4] Without loss of generality, in the example and throughout the paper, I assume that ties are broken in favor of a conviction.

for $n = 1, 2$. The principal's expected payoff is given by

$$2 * \left(\frac{4}{10} \frac{2}{3} - \frac{6}{10} \frac{7}{18} \right) = \frac{1}{15}$$

Trial mechanisms are restrictive because they do not allow for any type of communication between the principal and the agents. In general, however, the principal may benefit from that communication as I now illustrate.

Imagine that the principal allows each agent to choose between confessing to being guilty (option c) and not confessing (option $\overline{c}$). And imagine also that the principal makes the following promise to each agent n: "tell me whether you are innocent or guilty and I promise that your punishment will not depend on what you report but only on the evidence and on the other agent's report". If the principal can commit to such a promise, each agent becomes indifferent between his reports. Therefore, the principal is able to induce truthful reporting by the agents: guilty agents confess to being guilty and choose c, while innocent agents refuse and choose $\overline{c}$. While the information that is being provided by each agent cannot be used to determine his own punishment, it can be used to determine the *other* agent's punishment. In particular, if agent $n = 1$ confesses, the principal learns that agent $n = 2$ is innocent, so it is in her best interest to acquit him. If agent $n = 1$ refuses to confess, because the principal infers that agent $n = 1$ is innocent, she chooses to punish agent $n = 2$ if and only if

$$\Pr \{t_2 = g | t_1 = i, \theta\} \geq \frac{1}{2}$$

The difference to the trial mechanism is that, when deciding each agent's punishment, the principal uses not only the evidence θ but also the knowledge of the other agent's type - his guilt - which is obtained from his report. Seeing as the agents' types are correlated, that knowledge ends up being valuable, which is why the trial mechanism ends up being suboptimal.[5]

[5] If the agents' types were independent, the two mechanisms would be equivalent because

$$\Pr \{t_n = g | \theta, t_{-n}\} = \Pr \{t_n = g | \theta\}$$

for all t_{-n} and for all n.

If we use the distribution of (1), we have that

$$\Pr\{t_1 = g|\theta, t_2 = g\} = \Pr\{t_2 = g|\theta, t_1 = g\} = 0 < \frac{1}{2}$$

for all θ, while

$$\Pr\{t_1 = g|\theta = 1, t_2 = i\} = \Pr\{t_2 = g|\theta = 0, t_1 = i\} = \frac{\frac{4}{10}\frac{2}{3}}{\frac{4}{10}\frac{2}{3} + \frac{2}{10}\frac{1}{2}} = \frac{8}{11} \geq \frac{1}{2}$$

and

$$\Pr\{t_1 = g|\theta = 0, t_2 = i\} = \Pr\{t_2 = g|\theta = 1, t_1 = i\} = \frac{\frac{4}{10}\frac{1}{3}}{\frac{4}{10}\frac{1}{3} + \frac{2}{10}\frac{1}{2}} = \frac{4}{7} \geq \frac{1}{2}$$

so that each agent is punished if and only if the other agent refuses to confess, which, in equilibrium, only happens when the other agent is innocent.

To make matters a little more formal, recall that, as discussed in the introduction, a mechanism is a pair (M, d), where $M = M_1 \times M_2$ and where M_n represents the message set of each agent n, while mapping d maps message vectors m and evidence θ into punishment vectors: $d(m, \theta) \in [0, 1]^2$. The previous description corresponds to a mechanism labeled (M', d'), where $M'_n \in \{c, \overline{c}\}$ for $n = 1, 2$ and where

$$d'_n(m, \theta) = \begin{cases} 0 \text{ if } m_{-n} = c \\ 1 \text{ if } m_{-n} = \overline{c} \end{cases} \quad \text{for all } (m_{-n}, \theta)$$

Given mechanism (M', d'), each agent is indifferent as to what to report, so it is an equilibrium for guilty agents to confess and for innocent agents to refuse. The expected punishment of a guilty agent is 1: if agent n is guilty, agent $-n$ is certainly innocent and sends message $\overline{c}$, which results in agent n being punished in 1 regardless of θ. The expected punishment of an innocent agent is $\frac{\frac{2}{10}}{\frac{6}{10}} = \frac{1}{3}$, which is smaller than in the trial mechanism. Therefore, it is clear that the principal is better off than with the trial mechanism: her expected payoff is now

$$2 * \left(\frac{4}{10}1 - \frac{6}{10}\frac{1}{3}\right) = \frac{2}{5}$$

In the text, I show that mechanism (M', d') is in fact one of the optimal mech-

anisms for the principal if she can commit. Using mechanism (M', d'), it is easy to build a CIM $\left(M^{FC}, d^{FC}\right)$ that is also optimal (FC stands for "full commitment") - recall that any mechanism (M, d) is a CIM if and only if, for $n = 1, 2$, i) $M_n = \{c, \overline{c}\}$ and ii) $d_n\left(c, m_{-n}, \theta\right)$ is independent of (m_{-n}, θ). For $n = 1, 2$, let $M_n^{FC} = M_n' = \{c, \overline{c}\}$ and

$$d_n^{FC}\left(\overline{c}, m_{-n}, \theta\right) = d_n'\left(\overline{c}, m_{-n}, \theta\right)$$

for all (m_{-n}, θ), while

$$d_n^{FC}\left(c, m_{-n}, \theta\right) = 1$$

In words, provided agents report truthfully (guilty agents confess, while innocent agents refuse), if agent n refuses to confess, he gets the same lottery of punishments as in mechanism d_n', while, if he confesses, he gets 1 - the certainty equivalent punishment of the guilty type. So, innocent agents refuse to confess because $1 > \frac{1}{3}$, while guilty agents are indifferent between the two options (so they have just enough incentives to confess). And this mechanism is just as good as mechanism (M', d') because, regardless of their type, the expected punishment of each agent is the same under both mechanisms.[6]

The "problem" with mechanisms $\left(M^{FC}, d^{FC}\right)$ and (M', d') is that they induce agents to confess if and only if they are guilty. This means that, in equilibrium, when the principal observes an agent refusing to confess, she knows that the agent is innocent but is "forced" by the mechanism to punish him sometimes: in this example, if the principal observes that both agents have refused to confess, she correctly infers that both are innocent but is still required by the mechanism to punish them. Naturally, whenever that happens, the principal and the agent have an incentive to renegotiate and agree that it would be best for there not to be any punishment.

In the second part of the paper, I define what are renegotiation proof mechanisms: loosely speaking, a mechanism is renegotiation proof if, after observing any message vector m and evidence θ, the principal does not want to reduce the ensuing punishment. The challenge of finding the optimal renegotiation proof mechanism is that the revelation principle need not hold. In fact, Bester and Strausz (2000) suggest that it may fail in a context with multiple agents. Nevertheless, I show in the text that CIMs

[6] This multiplicity is also true in general: there are many optimal mechanisms and one of them is a CIM. If, for example, I was to assume that the players were risk averse (like in Siegel and Strulovici (2018)), the CIM would be uniquely optimal.

are still optimal even when one restricts attention to renegotiation proof mechanisms, i.e., there is an optimal renegotiation proof mechanism $\left(M^{RP}, d^{RP}\right)$ that is a CIM. There are however a couple of differences between $\left(M^{RP}, d^{RP}\right)$ and $\left(M^{FC}, d^{FC}\right)$.

The first difference is that, while both mechanisms induce the agents to refuse to confess if innocent, if the mechanism is $\left(M^{RP}, d^{RP}\right)$ each agent $n = 1, 2$ confesses only with a probability $\tau_n \in [0, 1]$ when guilty. The reason why truthful revelation is no longer optimal is that, if each agent was to confess if guilty and refuse if innocent, it would have to be that, following the observation that agent n has chosen to refuse to confess, the mechanism would impose a punishments of 0 on agent n, because any positive punishment would be renegotiated. But, of course, this also implies that punishments that follow confessions must be 0, for otherwise the agent would never choose to confess even if guilty. So, if a mechanism induces truthful reporting by the agents, it essentially acquits them no matter what, which, in general, is not optimal. Under (1), one can show that the optimal renegotiation proof mechanism induces $\tau_1 = \tau_2 = \frac{1}{4}$.

The second difference has to do with the punishments themselves. Mapping d^{RP} is such that, for $n = 1, 2$,

$$d_n^{RP}\left(\overline{c}, m_{-n}, \theta\right) = \begin{cases} 1 \text{ if } \Pr\left\{t_n = g | m_n = \overline{c}, m_{-n}, \theta, \tau_1 = \tau_2 = \frac{1}{4}\right\} \geq \frac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

for all m_{-n} and θ. In words, the optimal renegotiation proof mechanism is sequentially optimal after a refusal to confess, i.e., if agent n refuses to confess, the principal decides his punishment based on all the information available: evidence θ, the other agent's message and the agent's own message, whereas before, in mechanism $\left(M^{FC}, d^{FC}\right)$, she would not consider the agent's own message. Notice that

$$\begin{aligned} & \Pr\left\{t_1 = g | \theta = 1, m = \left(\overline{c}, \overline{c}\right), \tau_1 = \tau_2 = \frac{1}{4}\right\} \\ = & \Pr\left\{t_2 = g | \theta = 0, m = \left(\overline{c}, \overline{c}\right), \tau_1 = \tau_2 = \frac{1}{4}\right\} \\ = & \frac{\frac{4}{10}\frac{2}{3}\left(1 - \frac{1}{4}\right)}{\frac{4}{10}\frac{2}{3}\left(1 - \frac{1}{4}\right) + \frac{4}{10}\frac{1}{3}\left(1 - \frac{1}{4}\right) + \frac{2}{10}\frac{1}{2}} \\ = & \frac{1}{2} \geq \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned}
& \Pr\left\{t_1 = g|\theta = 0, m = (\overline{c}, \overline{c}), \tau_1 = \tau_2 = \frac{1}{4}\right\} \\
= & \Pr\left\{t_2 = g|\theta = 1, m = (\overline{c}, \overline{c}), \tau_1 = \tau_2 = \frac{1}{4}\right\} \\
= & \frac{\frac{4}{10}\frac{1}{3}\left(1 - \frac{1}{4}\right)}{\frac{4}{10}\frac{1}{3}\left(1 - \frac{1}{4}\right) + \frac{4}{10}\frac{2}{3}\left(1 - \frac{1}{4}\right) + \frac{2}{10}\frac{1}{2}} \\
= & \frac{1}{4} < \frac{1}{2}
\end{aligned}$$

so that

$$d_1^{RP}(\overline{c}, m_2, \theta) = \begin{cases} 1 \text{ if } m_2 = \overline{c} \text{ and } \theta = 1 \\ 0 \text{ otherwise} \end{cases}$$

and

$$d_2^{RP}(\overline{c}, m_1, \theta) = \begin{cases} 1 \text{ if } m_1 = \overline{c} \text{ and } \theta = 0 \\ 0 \text{ otherwise} \end{cases}$$

while

$$d_n^{RP}(c, m_{-n}, \theta) = \frac{2}{3} \text{ for all } (m_{-n}, \theta)$$

The expected punishment of a guilty agent is $\frac{2}{3}$ (because a guilty agent is indifferent between sending the two messages), while the expected punishment of an innocent agent is

$$\frac{\frac{4}{10}\left(\frac{1}{4}0 + \frac{3}{4}\left(\frac{1}{3}1 + \frac{2}{3}0\right)\right) + \frac{2}{10}\left(\frac{1}{2}1 + \frac{1}{2}0\right)}{\frac{6}{10}} = \frac{1}{3}$$

The principal's expected payoff is

$$2 * \left(\frac{4}{10}\frac{2}{3} - \frac{6}{10}\frac{1}{3}\right) = \frac{2}{15}$$

It is then clear that this mechanism is worse than the full commitment one but better than the trial mechanism.

3 Model

Types: There are N agents and a principal. Each agent n randomly draws a type $t_n \in \{i, g\} \equiv T_n$, which is his private information: each agent n is either innocent (i) or guilty (g) of committing the crime. Let $T \equiv T_1 \times ... \times T_N$ be the set of all possible vectors of agents' types and T_{-n} be the set of all possible vectors of types of agents other than n. The ex-ante probability that vector t is realized is denoted by $\pi(t)$ for all $t \in T$ and is assumed to be common knowledge. I assume that there is, at most, one guilty agent: $\pi(t) = 0$ for all t for which there is (n, n') such that $t_n = t_{n'} = g$. This implies that, when an agent is guilty, he knows that everyone else is innocent. I make this assumption because i) it is likely to be true for most crimes, and ii) in circumstances where there is a large probability that there are multiple criminals, one would think that the criminals would know each other's identities, which would considerably complicate the analysis.

Evidence: After $t \in T$ has been drawn, evidence is produced as a result of an investigation. Evidence is modelled as an exogenous random vector $\theta \in \Theta$ that is correlated with the agents' guilt. I define $p(\theta|t)$ as the conditional distribution of θ given vector $t \in T$. The fact that θ and t are correlated allows the principal to use θ as a way to improve his knowledge about the guilt of each agent. I assume that it is costless to produce evidence in order to distinguish my work from Kaplow and Shavell (1994), who point out that CIMs might improve upon trial systems because they are cheaper: there would be less of a need to have costly investigations if agents were to credibly confess their guilt. By making these investigations costless, I am negating this advantage of CIMs, and making it clear that the advantages that CIMs have over trial mechanisms in this paper have a different origin.

Preferences: I assume that each agent n's payoff is given by $u(x_n) = -x_n$, where $x_n \in [0, 1]$ represents the punishment agent n receives - it could be time in prison, community service time, physical punishment or a monetary fine. As for the principal, she is thought of as representing society in a way. Her goal is to maximize the punishments of the guilty agents and minimize those of the innocent. In particular, I assume that the principal's payoff, as a function of the type vector $t \in T$ and of the

punishment vector $x = (x_1, .., x_N) \in [0,1]^N$, is given by

$$v(t,x) = \sum_{n=1}^{N} v_n(t_n, x_n)$$

where

$$v_n(t_n, x_n) = \begin{cases} x_n \text{ if } t_n = g \\ -\alpha_n x_n \text{ if } t_n = i \end{cases}$$

with $\alpha_n > 0$ for all n. Parameters α_n measure the importance of protecting innocent agents relative to the importance of punishing guilty agents.[7] In a first best world, where the principal observes vector t, she would like to punish agent n in $x_n = 1$ if he is guilty and in $x_n = 0$ if he is innocent.

Notice that the agents and the principal are assumed to be risk neutral. This assumption is made in order to separate my argument from Grossman and Katz (1983), who show that risk aversion leads to CIMs doing better than trial mechanisms (see related literature). In this paper, CIMs do better than trial systems for different reasons (if I was to assume risk aversion on either side - the agents or the principal - CIMs would do even better).

The way the payoff function of the principal is specified is supposed to be general enough to embed a variety of theories about the goal of the criminal justice system: for example, it might be that the State wishes to punish guilty agents to prevent them from committing further crimes; it might also be that the State wants to punish guilty agents to satisfy a public need of revenge; and, as I make explicit in Appendix A, it is also possible to interpret the principal's preferences as representing society's desire to deter crime. In particular, in appendix A, I show that there is no loss of generality in not modelling the agents' decision of whether to commit a crime. The basic logic is that, if the principal wants to stop agent n from committing a crime, it should maximize the difference between the expected punishment of agent n should he choose to become guilty, and the expected punishment of agent n should he choose to become innocent. This can be achieved by properly choosing α_n.

Timing: The timing is the following. Before any evidence is generated (and, so, before an investigation has been initiated), the principal selects a mechanism.

[7] A reasonable assumption (yet unnecessary for any of the results to follow) is that α_n is the same for all n.

A mechanism is made of a message set $M = M_1 \times ... \times M_N$ and a mapping $d : M \times \Theta \rightarrow [0,1]^N$. Given the mechanism (M,d), each agent n chooses to send a message $m_n \in M_n$ to the principal. After message vector $m \in M$ has been sent, $\theta \in \Theta$ is realized and vector $x = d(m,\theta)$ is implemented, i.e., each agent does not observe the actual realization of θ before choosing his message.

In this context, an agent's strategy is a probability distribution over his message set M_n for each type: $\sigma_n : \{i,g\} \rightarrow \Delta M_n$ for $t_n \in \{i,g\}$. Vector $\sigma = (\sigma_1, ..., \sigma_N)$ represents the strategy profile of the N agents. Each profile $((M,d),\sigma)$ is called a system.

Definition 1: *System $((M,d),\sigma)$ is incentive compatible if and only if σ is a Bayes-Nash equilibrium of the game induced by mechanism (M,d).*

Formally, σ is a Bayes-Nash equilibrium of the game induced by mechanism (M,d) if and only if, for all n, whenever $\sigma_n(t_n)(m_n) > 0$ then

$$E^{\sigma}\left(d_n(m_n, m_{-n}, \theta)\,|t_n\right) \leq E^{\sigma}\left(d_n(m'_n, m_{-n}, \theta)\,|t_n\right)$$

for all $m'_n \in M_n$, where E^{σ} represents the expectation formed under the belief that strategy profile σ will be played by all other players.

Definition 2: *A trial mechanism (M,d) is such that d is independent of m for all $m \in M$.*

A trial mechanism is defined to be such that the punishment that each agent receives does not depend on his report but only on the evidence produced: it is as if there is no report by the agents; the principal simply observes the evidence θ and chooses punishments.

Definition 3: *A confession inducing mechanism (CIM) (M,d) is such that, for all n, $M_n = \{c, \overline{c}\}$ and $d_n(c, m_{-n}, \theta)$ is independent of $m_{-n} \in M_{-n}$ and $\theta \in \Theta$.*

One can interpret a CIM as having two stages. In the first stage, before any investigation has been initiated, every agent is given the opportunity to confess to being guilty. If the agent confesses (sends message c), he receives a constant punishment. If he refuses to confess (sends message $\overline{c}$), his punishment will be determined in a second stage, after an investigation has occurred and after all other agents have chosen

whether to confess or not. If a system $((M,d),\sigma)$ is such that mechanism (M,d) is a CIM, then I say that the system is a confession inducing system (CIS).

4 Full commitment power

When the principal chooses a particular mechanism (M,d), she anticipates that some strategy σ will be played by the agents. Typically, it is the system $((M,d),\sigma)$, and not only the mechanism (M,d), that determines the payoff of both the agents and the principal. If the principal has "full" commitment power, she is able to pick any incentive compatible system.[8] The optimal mechanism is the mechanism associated with the optimal incentive compatible system. I start by considering the case where the principal is restricted to trial mechanisms as a benchmark.

4.1 The optimal trial mechanism

Recall that, in any trial mechanism, the reports of the agents are irrelevant as the punishments are only a function of the evidence. Let $\left(M^{Tr},d^{Tr}\right)$ denote the optimal trial mechanism: the mechanism that maximizes the principal's expected payoff among all trial mechanisms. Seeing as d^{Tr} is independent of $m\in M^{Tr}$ by definition, one can, without loss of generality, let $M_n^{Tr}=\{\overline{c}\}$, where $\overline{m}=(\overline{c},...,\overline{c})$. As for d^{Tr}, it follows that

$$d_n^{Tr}(\overline{m},\theta)\in\arg\max_{x_n\in[0,1]}E\left(v_n\left(t_n,x_n\right)|\theta\right)$$

for all $\theta\in\Theta$ and n. As illustrated in the example, this means that

$$d_n^{Tr}(\overline{m},\theta)=\begin{cases}1 \text{ if } \Pr\left(t_n=g|\theta\right)\geq\alpha_n\Pr\left(t_n=i|\theta\right)\\ 0 \text{ otherwise}\end{cases}$$

for all $\theta\in\Theta$ and n.

[8] Here, I implicitly follow the standard assumption in mechanism design of assuming that, in the event that there are many strategy profiles σ for which $((M,d),\sigma)$ is an incentive compatible system, the principal can choose her preferred one.

As one would have expected, in the optimal trial mechanism, parameters α_n determine the standard of proof: if α_n is large, there must be more conclusive evidence against agent n for him to be punished, i.e., $\Pr(t_n = g|\theta)$ must be larger.

4.2 The optimal mechanism

As described above, in order to find the optimal mechanism, one must find the optimal system $\left(\left(M^{FC}, d^{FC}\right), \sigma^{FC}\right)$. The revelation principle (see, for example, Myerson (1979)) is extremely useful in this context seeing as it states that, without loss of generality, one can restrict attention to revelation mechanisms that induce thruthful reporting by the agents, i.e., without loss of generality, for all n, $M_n^{FC} = \{c, \overline{c}\}$ for all n, while

$$\sigma_n^{FC}(g)(m_n) = \begin{cases} 1 \text{ if } m_n = c \\ 0 \text{ if } m_n = \overline{c} \end{cases} \text{ and } \sigma_n^{FC}(i)(m_n) = \begin{cases} 0 \text{ if } m_n = c \\ 1 \text{ if } m_n = \overline{c} \end{cases}$$

The goal then is to find the optimal d^{FC} subject to the so called incentive constraints: guilty agents must prefer to confess (send message c) and innocent agents must prefer not to (send message $\overline{c}$).

Given that $v(t,x) = \sum_{n=1}^{N} v_n(t_n, x_n)$, this problem can be transformed into N independent problems where, in each n^{th} problem, one chooses $d_n^{FC} : M^{FC} \times \theta \to [0,1]$ subject to the two incentive constraints with respect to agent n. Formally, in the n^{th} problem, the principal chooses over mappings $d_n : M^{FC} \times \theta \to [0,1]$ in order to maximize her expected utility that refers to agent n, which is equal to

$$\Pr\{t_n = g\} E^{\sigma^{FC}}(d_n(c, m_{-n}, \theta)|t_n = g) - \alpha_n \Pr\{t_n = i\} E^{\sigma^{FC}}(d_n(\overline{c}, m_{-n}, \theta)|t_n = i) \tag{2}$$

where $\Pr\{t_n\}$ simply represents the ex-ante probability that t_n is realized; subject to the "guilty's incentive constraint", which can be written as

$$E^{\sigma^{FC}}(d_n(c, m_{-n}, \theta)|t_n = g) \leq E^{\sigma^{FC}}(d_n(\overline{c}, m_{-n}, \theta)|t_n = g) \tag{3}$$

and to the "innocent's incentive constraint", which can be written as

$$E^{\sigma^{FC}}\left(d_n\left(\overline{c},m_{-n},\theta\right)|t_n=i\right)\leq E^{\sigma^{FC}}\left(d_n\left(c,m_{-n},\theta\right)|t_n=i\right) \tag{4}$$

Proposition 1 *d^{FC} is a CIM such that, for all n,*

$$d_n^{FC}\left(\overline{c},m_{-n},\theta\right)=\begin{cases} 1 \text{ if } \Pr^{\sigma^{FC}}\left\{t_n=g|m_{-n},\theta\right\}\geq\alpha_n\Pr^{\sigma^{FC}}\left\{t_n=i|m_{-n},\theta\right\} \\ 0 \text{ otherwise}\end{cases}$$

for all $m_{-n}\in M_{-n}^{FC}$ and for all $\theta\in\Theta$, and

$$d_n^{FC}\left(c,m_{-n},\theta\right)=E^{\sigma^{FC}}\left(d_n^{FC}\left(\overline{c},m_{-n},\theta\right)|t_n=g\right)$$

for all $m_{-n}\in M_{-n}^{FC}$ and for all $\theta\in\Theta$.

Proof. I guess (and later verify) that constraint (4) does not bind. This immediately implies that constraint (3) binds: if none of the constraints was to bind, the principal would select the "first best" mechanism, consisting of a punishment of 1 should the agent confess (for any (m_{-n},θ)) and a punishment of 0 otherwise, which would violate constraint (3). If constraint (3) binds, it holds with equality. By plugging that equality into (2), we get

$$\Pr\left\{t_n=g\right\}E^{\sigma^{FC}}\left(d_n\left(\overline{c},m_{-n},\theta\right)|t_n=g\right)-\alpha_n\Pr\left\{t_n=i\right\}E^{\sigma^{FC}}\left(d_n\left(\overline{c},m_{-n},\theta\right)|t_n=i\right) \tag{5}$$

which only depends on the punishments that follow message $\overline{c}$. Using the fact that, under strategy profile σ^{FC} agents report truthfully, and simple algebra allows me to write (5) as

$$E^{\sigma^{FC}}\left[\left(\Pr^{\sigma^{FC}}\left\{t_n=g|m_{-n},\theta\right\}-\alpha_n\Pr^{\sigma^{FC}}\left\{t_n=i|m_{-n},\theta\right\}\right)d_n\left(\overline{c},m_{-n},\theta\right)\right] \tag{6}$$

which is maximized by choosing $d_n\left(\overline{c},m_{-n},\theta\right)=d_n^{FC}\left(\overline{c},m_{-n},\theta\right)$ for all (m_{-n},θ).

As for the choice of the punishments that follow a confession (the $d_n\left(c,m_{-n},\theta\right)$), they do not enter condition (6) directly, so the only requirement is that condition (3) holds with equality: it is only necessary that, when guilty, the expected punishment of

confessing is the same as the expected punishment of not confessing, which explains why

$$d_n^{FC}\left(c, m_{-n}, \theta\right) = E^{\sigma^{FC}}\left(d_n\left(\overline{c}, m_{-n}, \theta\right) | t_n = g\right)$$

for all (m_{-n}, θ).

Finally, I show in appendix C that

$$E^{\sigma^{FC}}\left(d_n\left(\overline{c}, m_{-n}, \theta\right) | t_n = g\right) \geq E^{\sigma^{FC}}\left(d_n\left(\overline{c}, m_{-n}, \theta\right) | t_n = i\right)$$

which implies that, under system $\left(\left(M^{FC}, d^{FC}\right), \sigma^{FC}\right)$, the expected punishment agent n receives when he is innocent and chooses not to confess is smaller than the punishment of confessing. Therefore, one confirms that constraint (4) is satisfied under system $\left(\left(M^{FC}, d^{FC}\right), \sigma^{FC}\right)$, so that the guess that (4) did not bind was correct.[9] ■

In the CIS $\left(\left(M^{FC}, d^{FC}\right), \sigma^{FC}\right)$, if agent n refuses to confess, he receives a punishment of 1 if

$$\Pr\left\{t_n = g | t_{-n}, \theta\right\} \geq \alpha_n \Pr\left\{t_n = i | t_{-n}, \theta\right\} \tag{7}$$

and 0 otherwise (because, under strategy profile σ^{FC}, each agent confesses if guilty and refuses if innocent). As explained in the proof of proposition 2, the punishments that follow a refusal to confess directly determine the punishment that follows a confession (the guilty type must be indifferent between confessing and refusing to confess). So, in the CIS $\left(\left(M^{FC}, d^{FC}\right), \sigma^{FC}\right)$, it is "as if" each agent n is punished in 1 if condition (7) holds, and in 0 otherwise.[10] In words, each agent's punishment is chosen by the principal after updating her beliefs about the guilt of the agent using the evidence (θ) and the information that every *other* agent has provided (t_{-n}).

[9] As in standard principal agent models with only two types, the only incentive constraint that binds is the one from the type the principal would like to punish the most. In this model, it is the guilty type. If, for example, the principal was a seller and the agent a buyer, that type would be the one who values the good that is being sold the most, for, in the absence of incentive constraints, it would be the one the principal would like to charge a larger price.

[10] Put differently, just like in the example, there is another system $\left(\left(M', d'\right), \sigma'\right)$ that is also optimal, where

$$M' = M^{FC},\ \sigma' = \sigma^{FC}$$

and, for all n,

$$d'_n\left(m, \theta\right) = d_n^{FC}\left(\overline{c}, m_{-n}, \theta\right)$$

for all (m, θ).

By contrast, in the optimal trial mechanism, each agent n is punished in 1 if

$$\Pr\{t_n = g|\theta\} \geq \alpha_n \Pr\{t_n = i|\theta\}$$

and in 0 otherwise. Therefore, one can see why it is that the CIM $\left(M^{FC}, d^{FC}\right)$ does better than the optimal trial mechanism: it allows the principal to use more information when punishing each agent. Naturally, this advantage would no longer exist if the extra information (the t_{-n}) was not informative about agent n's type, which would happen if the agents' types were independent, which is not the case.

5 Renegotiation Proof Systems

In this section, I restrict attention to mechanisms that the players do not want to renegotiate, i.e., in equilibrium, the principal must not be better off by reducing some agent's punishment. The principal's willingness to renegotiate depends on the strategy profile of the agents: the "problem" of mechanism $\left(M^{FC}, d^{FC}\right)$ is that it induces agents to report truthfully, which makes the punishments that follow a refusal to confess renegotiable. Therefore, the notion of "renegotiation proofness" is one that must be applied to systems.

Definition 4: *A system* $((M,d),\sigma)$ *is renegotiation proof if and only if, for all* $m \in M$, $\theta \in \Theta$ *and* n,

$$d_n(m,\theta) \leq \gamma_n^{\sigma}(m,\theta)$$

where

$$\gamma_n^{\sigma}(m,\theta) \equiv \max\left\{\arg\max_{x_n \in [0,1]} E^{\sigma}\left(v_n(t_n,x_n)|m,\theta\right)\right\}$$

Set

$$\arg\max_{x_n \in [0,1]} E^{\sigma}\left(v_n(t_n,x_n)|m,\theta\right)$$

represents the set of punishments that the principal would prefer to choose after observing (m,θ) and given strategy profile σ. If a system is renegotiation proof, it is never the case that, after observing some (m,θ), the principal prefers to choose a smaller punishment, because such a reduction would be promptly accepted by the

agent in question. If system $((M,d),\sigma)$ is renegotiation proof, then I say that mechanism (M,d) is renegotiation proof.

Notice that, because the principal is risk neutral,

$$\gamma_n^{\sigma}(m,\theta)=\begin{cases} 1 \text{ if } \Pr^{\sigma}(t_n=g|m,\theta)\geq\alpha_n\Pr^{\sigma}(t_n=i|m,\theta) \\ 0 \text{ otherwise} \end{cases}$$

The goal is to find the optimal renegotiation proof incentive compatible (RPIC) system.

Proposition 2 *The optimal RPIC system* $\left(\left(M^{RP},d^{RP}\right),\sigma^{RP}\right)$ *is a CIS and has the following properties:*

i) for all n,

$$\sigma_n^{RP}(g,m_n)=\begin{cases} \tau_n^{RP} \text{ if } m_n=c \\ 1-\tau_n^{RP} \text{ if } m_n=\overline{c} \end{cases} \quad \text{and } \sigma_n^{RP}(i,m_n)=\begin{cases} 0 \text{ if } m_n=c \\ 1 \text{ if } m_n=\overline{c} \end{cases}$$

for some $\tau^{RP}\in[0,1]^N$;

ii) for all n,

$$d_n^{RP}(\overline{c},m_{-n},\theta)=\gamma_n^{\sigma^{RP}}(\overline{c},m_{-n},\theta)$$

and

$$d_n^{RP}(c,m_{-n},\theta)=E^{\sigma^{RP}}\left(d_n^{RP}(\overline{c},m_{-n},\theta)\,|t_n=g\right)$$

for all $m_{-n}\in M_{-n}^{RP}$ *and for all* $\theta\in\Theta$.

Proof. See Appendix C. ■

Just like when the principal had commitment power, there is still a CIM that is optimal among all renegotiation proof mechanisms. There are, however, two main differences. The first difference is that, unlike the mechanism of the previous section, CIM $\left(M^{RP},d^{RP}\right)$ does not induce complete separation between each agent's type. In particular, while innocent agents always refuse to confess, it is no longer the case that guilty agents always confess. Instead, in equilibrium, they randomize between confessing and not confessing.

This is a consequence of limiting the principal's ability to commit, very much like in several other related papers in the literature of law enforcement. For example, Baker and Mezzetti (2001) assume that prosecutors are able to choose how much effort to put into gathering evidence about the crime, after having given the opportunity for the defendant to confess. Given that the prosecutors have no commitment power, in equilibrium, only some guilty agents will choose to confess, while the remaining ones (alongside the innocents) will not. Other examples include Kim (2010), Franzoni (1999) or Bjerk (2007). What is different about this result is that I show that a CIM that induces partial separation is optimal, while these other papers directly solve for the equilibria of an exogenously specified game.

The second major difference has to do with what happens after an agent has refused to confess. In the previous section, agents who refused to confess were sometimes punished depending on the evidence and on the other agents' reports even though it was commonly known they were innocent. But with system $\left(\left(M^{RP}, d^{RP}\right), \sigma^{RP}\right)$, the punishments that follow a refusal to confess are sequentially optimal, i.e., the punishment agent n receives after choosing $m_n = \overline{c}$ and given the other agents' reports (m_{-n}) and evidence θ is exactly the punishment the principal finds optimal given his properly updated beliefs about agent n's innocence. Therefore, the mechanism requires no commitment power after an agent has refused to confess. The only commitment power required is the power to commit not to increase the punishment of the agents who do confess.

To see why that is, notice that, in order for system $\left(\left(M^{RP}, d^{RP}\right), \sigma^{RP}\right)$ to be renegotiation proof, it must be that $d^{RP}(\overline{c}, m_{-n}, \theta) \leq \gamma_n^{\sigma}(\overline{c}, m_{-n}, \theta)$. Imagine that $d^{RP}(\overline{c}, m_{-n}, \theta) < \gamma_n^{\sigma}(\overline{c}, m_{-n}, \theta)$. Then, it would be possible for the principal to do better: she could simultaneously increase $d_n^{RP}(\overline{c}, m_{-n}, \theta)$ until it reached $\gamma_n^{\sigma}(\overline{c}, m_{-n}, \theta)$ (which would make her better off, conditional on agent n sending message $\overline{c}$, by definition of $\gamma_n^{\sigma}(\overline{c}, m_{-n}, \theta)$) and, in order to keep the guilty type of agent n indifferent, the punishment after a confession (to raise $d_n^{RP}(c, m_{-n}, \theta)$ would also make the principal better off, conditional on agent n sending message c, because only guilty agents confess).

6 Related Literature

There is a considerable amount of literature in economics that argues for the use of variants of CIMs in law enforcement. Kaplow and Shavell (1994) add a stage, where agents can confess to being guilty, to a standard model of negative externalities and argue that this improves social welfare because it saves monitoring costs. By making the punishment after a confession equal to the expected punishment of not confessing, the law enforcer is able to deter crime to the same extent as he was without the confession stage, but without having to monitor the confessing agents.

Grossman and Katz (1983) discuss the role of plea bargaining in reducing the amount of risk in the criminal justice system. The argument is that, by letting guilty agents confess and punishing them with the corresponding certainty equivalent punishment of going to trial, the principal reduces the risk of acquitting guilty agents.

Siegel and Strulovici (2018) consider a setting with a risk averse principal and a single risk averse agent and analyze alternatives to the traditional criminal trial procedure, where agents are either convicted or acquitted. The authors demonstrate that there is a welfare gain in increasing the number of verdicts an agent can receive: so, for example, a verdict of "not proven" in addition to the traditional verdicts of "guilty" and "not guilty". The paper also considers plea bargaining, interpreting a guilty plea as a special type of a third verdict that agents can choose, and show it is uniquely optimal in such a setup.

The main difference between these papers and mine is that the argument I make about the optimality of CIS's does not depend on the agents or the principal being risk averse (as these are assumed to be risk neutral) nor on them being cheaper (as there are no costs in my paper), but, rather on the fact that CIMs explore the correlation between the agents' innocence.

A key aspect of my argument has to do with the fact that the principal deals with different agents. There are a few articles on law enforcement which have also considered multiple defendants - for example Kim (2009), Bar-Gill and Ben Shahar (2009) and Kobayashi (1992). However, in those papers, it is assumed that all defendants are guilty and the emphasis is on finding the best strategy to make sure that they are punished, which is in contrast with this paper. Berg and Kim (2016) are an exception in that they allow for two states of nature in a context with two defendants:

either both defendants are guilty or both are innocent. In my paper, this information structure (perfect correlation) would allow for the implementation of the first best, provided there is commitment power by the principal.

The literature on industrial organization that considers the design of leniency programs in Antitrust law also considers multiple agents: see, among others, Motta and Polo (2003), Spagnolo (2008) or Harrington (2008). The typical approach of these papers is to analyze the impact that different exogenously imposed leniency policies have on cartel formation. The overall message is that leniency policies can be beneficial if properly designed. This message is shared by my work. The logic of the argument is quite similar in that, in either setting, it is the correlation between the agents' guilt that leads to this result. However, this correlation has a different nature: in my work, the correlation comes from the fact that the guilty agent knows that all others are innocent, while in this other literature, agents who are a part of a cartel know the identity of the other cartel members. Furthermore, unlike this literature, the exercise in this paper is one of mechanism design, where I find that the optimal policy is a leniency policy, rather than directly assuming it is.

In terms of the methodology, the environment studied in this paper is characterized by the fact that there is a single type of good denominated "punishment". The allocation of that good has implications not only to the agents but also to the principal's expected utility. There is some literature on mechanism design which considers similar environments by assuming that the principal cannot rely on transfer payments. In these environments, because the principal is deprived of an important instrument in satisfying incentive compatibility, it is necessary to find other ways of screening the different types of agents. One such way is to create hurdles in the mechanism that only some types are willing to go through. For example, Banerjee (1997), in solving the government's problem of assigning a number of goods to a larger number of candidates with private valuations, argues that, if these candidates are wealth constrained, it is efficient to make them go trough "red tape", in order to guarantee that those who value the good the most end up getting it. In Lewis and Sappington (2000), the seller of a productive resource uses the share of the project it keeps in its possession as a tool to screen between high and low skilled operators, which are wealth-constrained. Another approach is to assume that the principal is able to verify the report provided by the agents. This is the case, for example, of Ben-Porath, Dekel and Lipman (2014) and Mylovanov and Zapechelnyuk (2014), where it

is assumed that this verification, while maybe costly, is always accurate. This paper's approach is the latter: the principal is able to imperfectly and costlessly verify the agents' claims through evidence and by combining the reports from multiple agents.[11]

7 Conclusion

In this paper, I show how best to use the fact that agents' guilt is correlated when determining their punishment. In settings where there is at most one guilty person, guilt is necessarily correlated, because the fact that a certain person is guilty implies that no one else is. Given this, when someone credibly confesses their guilt, they produce information externalities: they reveal that no one else is guilty.

Under the assumption that it is possible/acceptable to show leniency towards agents who are believed to be guilty, I find that the optimal mechanism is a "confession inducing mechanism", where agents have the opportunity to confess in exchange for a constant punishment, rather than risking being subject to a full investigation, which may end up with them being acquitted or heavily punished. This description is clearly remisniscent of plea bargaining, a common criminal law practice in the United States. Plea bargaining gives the prosecutor the ability to commit to certain promises made to the defendant(s): the prosecutor is able to promise a reduced punishment (compared to the punishment the defendant would obtain if convicted in court) in exchange for a guilty plea. Furthermore, the fact that plea bargaining exists and is commonly used validates the assumption that it is acceptable to have mechanisms where agents who are believed to be guilty receive reduced punishments. This ability to commit to showing leniency comes from the Law: in particular, Rule 11 of the federal rules of criminal procedures protects the rights of defendants who choose to confess to being guilty.

Despite the similarities, plea bargaining does not fit that well into the point I am making because, in general, the negotiation initiated by the prosecutor occurs (too) late in the criminal process, at a time where there is only a single defendant who is being investigated. Naturally, for there to be information externalities, there must be

[11] Midjord (2013) also considers a setup without transfers, where the principal is able to imperfectly and costlessly verify the agents' reports through evidence. The main theoretical difference to this paper is that the author does not investigate the optimal mechanism.

more than one agent who is being considered. A possible takeaway from this paper is precisely that the opportunity to confess should come earlier in the criminal process, so as to be able to explore these information externalities. Such a system would be similar to the system of "self-reporting", which exists in environmental law, except that it would be applied to criminal law: in the context of the environmental law, self reporting allows firms to take the initiative to confess to having broken environmental regulations in exchange for smaller punishments, even when there is no investigation that is taking place.

There are, however, a few problems with expanding the policy of self-reporting to criminal cases, which are not directly studied in the text. One such problem is that innocent agents might be given enough incentives by guilty agents to confess in their stead, either through bribery or coercion. A related problem is the possibility of agents confessing to lesser crimes, rather than the ones they have committed. For example, someone who has committed first degree murder might be tempted to confess to manslaughter, as presumably the latter crime would render a smaller punishment. The implementation of a CIM in criminal law would then depend on whether it is possible to resolve these type of problems in a satisfying manner. A way to, at least, mitigate them would be to "validate" the confession of any given agent only if the evidence supports the claim.

A second problem with implementing such a system is that it is not clear how large punishments that follow confessions should be. In the model, punishments are a function of preferences, which are assumed to be observable. In reality though, preferences are not observable. Hence, the implementation of a CIM would necessarily have to rely on the existing and future research on defendants' preferences (see, for example, Tor, Gazal-Ayal and Garcia (2010) or Dervan and Edkins (2013)). I believe the careful analysis of these and other problems is essential to be able to convincingly argue for the introduction of this type of system in criminal law.

8 Appendix

The appendix is divided into three parts. In appendix A, I show why there is no loss of generality in not modelling the agents' decisions of committing the crime; in appendix B, I consider the case where the principal cannot commit; in appendix C, I present all the proofs of the main text.

8.1 Appendix A

In the main text, unlike some of the literature on law enforcement, I do not explicitly model the agents' decision of committing the crime.[12] I simply assume that the crime has been committed already and that the randomness of the agents' innocence (vector t) reflects the fact that the principal does not know the identity of the criminals. The concern that the reader might have about my approach is that the design of the mechanism itself might influence the agents' decisions of whether to become a criminal. In particular, in theory, it would be possible that the optimal mechanism that I find causes an increase in the number of people who choose to become criminals. In this extension, I address this concern.

In a typical law enforcement model, where agents choose whether to commit a crime, the problem each agent faces is the following.[13] There is some exogenous benefit for the agent of committing the crime, some negative externality caused by the crime and some cost depending on what he chooses to do. The benefit of committing the crime is normally thought of as something exogenous, independent of the criminal justice system. Therefore, the design of the criminal justice system only impacts the decision of each agent of whether or not to commit the crime insomuch as it affects his cost. In particular, what will determine whether each agent commits a crime is the difference between the expected punishment that the agent will receive if he commits it (and becomes guilty) and if he does not (and becomes innocent). Let v_n^g and v_n^i denote the expected punishment of agent n if he chooses to commit the crime and

[12] There is a branch of the literature on law enforcement, initiated by Becker (1968), that explicitly models the decisions of the agents of whether or not to commit a crime.

[13] See Garoupa (1997).

if he chooses not to commit the crime respectively, and let b_n denote the benefit of committing the crime. It follows that each agent n commits the crime if and only if

$$b_n \geq \upsilon_n^g - \upsilon_n^i$$

Hence, if the goal of the criminal justice is **only** to deter crime, the preferences of the principal should be

$$\sum_{n=1}^{N} \left(\upsilon_n^g - \upsilon_n^i \right)$$

But these are exactly the preferences defined in the text. Recall that

$$v_n \left(t_n, x_n \right) = \begin{cases} x_n \text{ if } t_n = g \\ -\alpha_n x_n \text{ if } t_n = i \end{cases}$$

so that

$$E \left(v_n \right) = \Pr \left\{ t_n = g \right\} \upsilon_n^g - \alpha_n \Pr \left\{ t_n = i \right\} \upsilon_n^i$$

which is proportional to $\upsilon_n^g - \upsilon_n^i$ provided that

$$\alpha_n = \frac{\Pr \left\{ t_n = g \right\}}{\Pr \left\{ t_n = i \right\}} \tag{8}$$

In other words, if (8) holds, the principal's goal is exclusively to deter crime.

8.2 Appendix B - Sequentially optimal systems

In this section, I briefly discuss the case where the principal has no commitment power.

Definition 5: *A system* $\left(\left(M, d \right), \sigma \right)$ *is sequentially optimal if and only if, for all* $m \in M$, $\theta \in \Theta$ *and* n,

$$d_n \left(m, \theta \right) \in \arg \max_{x_n \in [0,1]} E^{\sigma} \left(v_n \left(t_n, x_n \right) | m, \theta \right)$$

If a system is sequentially optimal, the principal never has remorse. It is never the case that after observing any (m, θ), the principal has a strict preference to choose a

punishment different than $d_n(m, \theta)$.

Let

$$q(\theta|n) \equiv p(\theta|t_n = g, t_{-n} = (i, ..., i))$$

and let

$$q(\theta|0) = p(\theta|t = (i, ..., i))$$

Proposition 3 *If i) θ is a continuous random variable and, ii) for all n, there is $\underline{\theta}_n \in \Theta$ and $\overline{\theta}_n \in \Theta$ such that*

$$\lim_{\theta \to \underline{\theta}_n} \frac{q(\theta|n)}{q(\theta|n')} = 0 \text{ for some } n' \geq 0$$

and

$$\lim_{\theta \to \overline{\theta}_n} \frac{q(\theta|n)}{q(\theta|n')} = \infty \text{ for all } n' \geq 0$$

then the optimal trial system is an optimal incentive compatible sequentially optimal system.

Proof. If there is a system that does better than the optimal trial system, it must be that, in equilibrium, some agent n sends two distinct messages a and b with positive probability that induce a different posterior belief regarding agent n's guilt. Say that, after message a, the belief that agent n is innocent is larger than after message b. This implies that, for the system to be sequentially optimal, punishments after message a must be lower than after message b. If conditions i) and ii) hold, these are strictly lower. But, in that case, agent n would never choose to send message b, a contradiction to a and b being sent in equilibrium. ■

Conditions i) and ii) are meant to capture the idea that there are all sorts of evidence that can potentially be found: that the evidence set is sufficiently large. They imply that any posterior belief can be formed after observing the evidence for any prior belief about each agent's guilt, i.e., for any $z \in (0, 1)$, there is always some $\theta \in \Theta$ that leads the principal to believe that agent n is guilty with probability z, for any system. In particular, if θ is close to $\underline{\theta}_n$, then agent n is very likely to be innocent, while if θ is close to $\overline{\theta}_n$ he is very likely to be guilty. I would argue that these conditions are very likely to be true for most scenarios being considered. Thus,

one can conclude that, by eliminating the principal's commitment power, in general, one also eliminates her ability to collect any information from the agents, so that the best she can do is the trial mechanism.

8.3 Appendix C - Proofs

8.3.1 Proof of Proposition 1

Proof. The proof of proposition 1 is completed by showing that the "innocent's incentive constraint" (condition (4)) does not bind. In order to show it, it is enough to show that condition (4) holds true under mapping d_n^{FC} for any n. Suppose not. Then, if agent n is innocent, he would be strictly better off by reporting c:

$$E^{\sigma^{FC}}\left(d_n^{FC}\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right) > E^{\sigma^{FC}}\left(d_n^{FC}\left(c, m_{-n}, \theta\right)|t_n = i\right)$$

Given that the interests of the principal and of the innocent are perfectly aligned (recall that the principal wants to minimize the expected punishment of agent n whenever he is guilty), it would actually be in the interest of the principal that the innocent agent did report c. Formally, let $d_n' : M^{FC} \times \Theta \to [0,1]$, where

$$d_n'\left(c, m_{-n}, \theta\right) = d_n^{FC}\left(c, m_{-n}, \theta\right) \text{ and } d_n'\left(\overline{c}, m_{-n}, \theta\right) = d_n^{FC}\left(c, m_{-n}, \theta\right)$$

for all (m_{-n}, θ) and for all n. Then, when we evaluate the principal's objective function, given by expression (2), at $d_n = d_n'$, we get that it is strictly larger than when evaluated at $d_n = d_n^{FC}$, because

$$E^{\sigma^{FC}}\left(d_n^{FC}\left(c, m_{-n}, \theta\right)|t_n = g\right) = E^{\sigma^{FC}}\left(d_n'\left(c, m_{-n}, \theta\right)|t_n = g\right)$$

but

$$\begin{aligned} E^{\sigma^{FC}}\left(d_n^{FC}\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right) &> E^{\sigma^{FC}}\left(d_n^{FC}\left(c, m_{-n}, \theta\right)|t_n = i\right) \\ &= E^{\sigma^{FC}}\left(d_n'\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right) \end{aligned}$$

But this is a contradiction to d_n^{FC} being the mapping that maximizes (2) subject to (3). ■

8.3.2 Proof of proposition 2

Let me start by introducing some additional notation: let ξ_n denote the probability that agent n is guilty and let

$$r_n^\sigma(m_n) \equiv \frac{\sigma_n(g)(m_n)}{\sigma_n(i)(m_n)}$$

denote the "likelihood ratio" of agent n under strategy profile σ and message m_n. Notice that

$$\Pr^\sigma(t_n = g|m,\theta) = \frac{\xi_n r_n^\sigma(m_n) p(\theta|t_n=g)}{\sum_{\widehat{n}=1}^{N} \xi_n r_{\widehat{n}}^\sigma(m_{\widehat{n}}) p(\theta|t_{\widehat{n}}=g) + \pi(i,...,i) p(\theta|t=(i,...,i))}$$

which means that the dependence of $\gamma_n^\sigma(m,\theta)$ over σ and m only comes from the likelihood ratio of each m_n, so that there is some function $h : \mathbb{R}_+^N \cup \Theta \to \{0,1\}^N$ such that, for all (m,θ),

$$\prod_{\widehat{n}} \sigma_{\widehat{n}}(i,m_n) > 0 \Rightarrow h_n\left(r^\sigma(m),\theta\right) = \gamma_n^\sigma(m,\theta) \tag{9}$$

for all n, where $r^\sigma(m) = (r_1^\sigma(m_1),...,r_n^\sigma(m_n))$.

The main challenge of doing mechanism design with limited commitment power is that the revelation principle need not hold. In fact, Bester and Strausz (2000) suggest it does not hold, so that, it is unclear what should set M be: in principle, it is possible that the optimal RPIC system is something quite intricate where agents send multiple messages. The next lemma allows me to restrict attention to a particular message set:

Lemma 2.1. *There is an optimal RPIC system* $\left(\left(M^{RP}, d^{RP}\right), \sigma^{RP}\right)$ *where* $M_n^{RP} = \mathbb{R}_+ \cup \{c\}$ *for all* n.

Proof. Whether a system is or is not renegotiation proof depends on the posterior beliefs of the principal after observing each vector (m,θ); in particular it depends on $\gamma_n^\sigma(m,\theta)$ for each n. By (9), those beliefs only depend on each likelihood ratio $r_n^\sigma(m_n)$. Therefore, M_n^{RP} only has to be large enough to accomodate all these ratios,

plus a message that is not sent by the innocent type: message c, which is such that $r_n^\sigma(c) = \infty$.

Formally, take any system $((M,d),\sigma)$ where, for some n, there are m_n' and m_n'' such that $r_n^\sigma(m_n') = r_n^\sigma(m_n'')$. The goal of the proof is to show that it is possible to eliminate one such message. In this way, the set of messages only needs to be large enough as $\mathbb{R}_+ \cup \{c\}$ because the range of $r_n^\sigma(\cdot)$ is $\mathbb{R}_+$ to which one adds the confessing message c.

Consider the alternative system $\left(\left(M,\overline{d}\right),\overline{\sigma}\right)$ that is equal to $((M,d),\sigma)$ except that:

$$\begin{cases} i)\ \overline{\sigma}_n(t_n)(m_n') = \sigma_n(t_n)(m_n') + \sigma_n(t_n)(m_n'') \text{ for } t_n = i,g \\ ii)\ \overline{\sigma}_n(t_n)(m_n'') = 0 \text{ for } t_n = i,g \\ ii)\ \overline{d}(m_n', m_{-n}, \theta) = \left(\begin{array}{c} \frac{\sigma_n(t_n)(m_n')}{\sigma_n(t_n)(m_n')+\sigma_n(t_n)(m_n'')} d(m_n', m_{-n}, \theta) \\ + \frac{\sigma_n(t_n)(m_n'')}{\sigma_n(t_n)(m_n')+\sigma_n(t_n)(m_n'')} d(m_n'', m_{-n}, \theta) \end{array} \right) \text{ for } t_n = i,g \\ iii)\ \overline{d}(m_n'', m_{-n}, \theta) = \overline{d}(m_n', m_{-n}, \theta) \end{cases}$$

The new system $\left(\left(M,\overline{d}\right),\overline{\sigma}\right)$ merges the two messages and effectively eliminates message m_n'': each agent is indifferent between the two messages but only sends message m_n'. If we denote the expected payoff of agent $\widehat{n}$, type $t_{\widehat{n}}$, under system $((M,d),\sigma)$ by $U_{\widehat{n}}^{t_{\widehat{n}}}(((M,d),\sigma))$, I claim that, for any $\widehat{n}$ and for any $t_{\widehat{n}}$, $U_{\widehat{n}}^{t_{\widehat{n}}}(((M,d),\sigma)) = U_{\widehat{n}}^{t_{\widehat{n}}}\left(\left(\left(M,\overline{d}\right),\overline{\sigma}\right)\right)$.

If $\widehat{n} = n$, this follows because, under system $((M,d),\sigma)$, the expected payoff of sending messages m_n' and m_n'' is the same as under system $\left(\left(M,\overline{d}\right),\overline{\sigma}\right)$.

If $\widehat{n} \neq n$, this follows because, for $t_{\widehat{n}} = i,g$,

$$\left\{ \begin{array}{c} \Pr^{\overline{\sigma}}\{m_n', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\} \overline{d}_n(m_{\widehat{n}}, m_n', m_{-\widehat{n},n}, \theta) + \\ \Pr^{\overline{\sigma}}\{m_n'', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\} \overline{d}_n(m_{\widehat{n}}, m_n'', m_{-\widehat{n},n}, \theta) \end{array} \right\} =$$
$$\left\{ \begin{array}{c} \Pr^{\sigma}\{m_n', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\} d_n(m_{\widehat{n}}, m_n', m_{-\widehat{n},n}, \theta) + \\ \Pr^{\sigma}\{m_n'', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\} d_n(m_{\widehat{n}}, m_n'', m_{-\widehat{n},n}, \theta) \end{array} \right\}$$

for any $m_{\widehat{n}}$ such that $\sigma_{\widehat{n}}(t_{\widehat{n}})(m_{\widehat{n}}) > 0$, where $\Pr^\sigma\{m_{-\widehat{n}}, \theta | t_{\widehat{n}}\}$ represents the density of event $(m_{-\widehat{n}}, \theta)$, given agent $\widehat{n}$'s type. In words, the distribution of punishments of agent $\widehat{n}$ is unchanged. The fact that $U_{\widehat{n}}^{t_{\widehat{n}}}(((M,d),\sigma)) = U_{\widehat{n}}^{t_{\widehat{n}}}\left(\left(\left(M,\overline{d}\right),\overline{\sigma}\right)\right)$ for any $\widehat{n}$

and for any $t_{\widehat{n}}$ implies that system $\left(\left(M,\overline{d}\right),\overline{\sigma}\right)$ is incentive compatible and is such that the expected utility of the principal is equal to when the system is $\left(\left(M,d\right),\sigma\right)$.

The proof is complete by showing that system $\left(\left(M,\overline{d}\right),\overline{\sigma}\right)$ is renegotiation proof. Notice that

$$r_n^{\sigma}\left(m_n'\right)=r_n^{\sigma}\left(m_n''\right)=r_n^{\overline{\sigma}}\left(m_n'\right)$$

which implies that, for all $\widehat{n}$ and for all $\left(m_{-n},\theta\right)$,

$$\gamma_{\widehat{n}}^{\sigma}\left(m_n',m_{-n},\theta\right)=\gamma_{\widehat{n}}^{\sigma}\left(m_n'',m_{-n},\theta\right)=\gamma_{\widehat{n}}^{\overline{\sigma}}\left(m_n',m_{-n},\theta\right)$$

and, consequently,

$$\overline{d}_{\widehat{n}}\left(m_n',m_{-n},\theta\right)\leq\gamma_{\widehat{n}}^{\overline{\sigma}}\left(m_n',m_{-n},\theta\right)$$

■

In what follows, without loss of generality, I assume that $M_n=\mathbb{R}_+\cup\{c\}$ for all n. Let me introduce some additional notation. Let

$$\upsilon_n^{\sigma}\equiv\inf_{m_n\in\mathbb{R}_+}\left\{r_n^{\sigma}\left(m_n\right):\sigma_n\left(i\right)\left(m_n\right)>0\right\}$$

for each σ and n. The interpretation of υ_n^{σ} is that it is the likelihood ratio of the message that has the lowest likelihood ratio of those that are sent with positive probability.

Lemma 2.2. *For any $\sigma:\{i,g\}\to\Delta\left\{\mathbb{R}_+\cup\{c\}\right\}$ and for every agent n, there is some message m_n^{σ} such that $\sigma_n\left(i\right)\left(m_n^{\sigma}\right)>0$ and for which*

$$h_n\left(\upsilon_n^{\sigma},r_{-n}^{\sigma}\left(m_{-n}\right),\theta\right)=0\Rightarrow h_n\left(r_n^{\sigma}\left(m_n^{\sigma}\right),r_{-n}^{\sigma}\left(m_{-n}\right),\theta\right)=0$$

for all $\left(m_{-n},\theta\right)$.

Proof. Suppose not. Then, for some $\left(m_{-n},\theta\right)$, it must be that

$$h_n\left(\upsilon_n^{\sigma},r_{-n}^{\sigma}\left(m_{-n}\right),\theta\right)=0\text{ but }h_n\left(r_n^{\sigma}\left(m_n\right),r_{-n}^{\sigma}\left(m_{-n}\right),\theta\right)=1$$

for all m_n for which $\sigma_n\left(i\right)\left(m_n\right)>0$. This implies that

$$\widehat{h}_n\left(r_n,r_{-n}^{\sigma}\left(m_{-n}\right),\theta\right)\geq\frac{\alpha_n}{1+\alpha_n}$$

for all $r_n > \upsilon_n^\sigma$ where

$$\widehat{h}_n\left(r_n, r_{-n}^\sigma\left(m_{-n}\right), \theta\right) = \frac{\xi_n r_n p\left(\theta | t_n = g\right)}{\sum_{\widehat{n}=1}^{N} \xi_n r_{\widehat{n}}^\sigma\left(m_{\widehat{n}}\right) p\left(\theta | t_{\widehat{n}} = g\right) + \pi\left(i, ..., i\right) p\left(\theta | t = (i, ..., i)\right)}$$

Therefore,

$$\lim_{r_n \to \upsilon_n^\sigma} \widehat{h}_n\left(r_n, r_{-n}^\sigma\left(m_{-n}\right), \theta\right) = \widehat{h}_n\left(\upsilon_n^\sigma, r_{-n}^\sigma\left(m_{-n}\right), \theta\right) \geq \frac{\alpha_n}{1 + \alpha_n}$$

which is a contradiction, because $h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma\left(m_{-n}\right), \theta\right) = 0$. ■

Lemma 2.2. allows me to treat m_n^σ as if $r_n^\sigma\left(m_n^\sigma\right) = \upsilon_n^\sigma$: it is as if m_n^σ is the message after which the principal believes agent n is the most likely to be innocent. For each σ, define mapping $d^\sigma : M \times \Theta \to [0,1]^N$ to be such that

$$d_n^\sigma\left(m, \theta\right) = h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma\left(m_{-n}\right), \theta\right)$$

for all n and for all $(m, \theta) \in M \times \Theta$. In words, in system $\left(\left(M, d^\sigma\right), \sigma\right)$, the principal asks each agent for a report and then decides their punishment using her most optimistic beliefs about the agent's guilt, i.e., when deciding agent n's punishment after receiving message (m_n, m_{-n}, θ), it is as if the principal pretends she has observed vector $(m_n^\sigma, m_{-n}, \theta)$ instead and then uses all the information available to select the corresponding punishment. Put differently, what this means is that, for any message m_n for which $r_n^\sigma\left(m_n\right) \neq \upsilon_n^\sigma$, the principal chooses a punishment that is smaller than what she would preferred ($d_n^\sigma\left(m, \theta\right) \leq h_n\left(r^\sigma\left(m\right), \theta\right)$ if $r_n^\sigma\left(m_n\right) \neq \upsilon_n^\sigma$), so that it is as if the principal has a bias towards each agent's innocence.

Notice that system $\left(\left(M, d^\sigma\right), \sigma\right)$ is trivially incentive compatible (as each agent's punishment does not depend on his own report) and is renegotiation proof because

$$h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma\left(m_{-n}\right), \theta\right) \leq h_n\left(r_n^\sigma\left(m\right), \theta\right)$$

for all (m, θ) and for all n by definition of υ_n^σ.

Lemma 2.3. *There is an optimal RPIC system* $\left(\left(M^{RP}, d^{RP}\right), \sigma^{RP}\right)$ *such that* $d^{RP} = d^{\sigma^{RP}}$.

Proof. Take any RPIC system $((M,d),\sigma)$. Without loss of generality, assume that, for all n, $M_n = \mathbb{R}_+ \cup \{c\}$ and $\sigma_n(i)(c) = 0$. Take some agent n. There are two cases to consider.

Case 1: m_n^σ is such that $\sigma_n(g)(m_n^\sigma) = 0$.

In this case, it follows that $\upsilon_n^\sigma = 0$. Given that $h_n\left(0, r_{-n}^\sigma(m_{-n})\right) = 0$ for all m_{-n}, it follows that $d_n(m_n^\sigma, m_{-n}, \theta) = 0$ for all (m_{-n}, θ) for the system to be renegotiation proof. But then, for the system to be incentive compatible, it must be that $d_n(m,\theta) = 0$ for all θ and for all m that are sent with positive probability, which proves the result.

Case 2: m_n^σ is such that $\sigma_n(g)(m_n^\sigma) > 0$.

Seeing as message m_n^σ is sent by both types, it follows that the expected payoff of agent n, type t_n, is exactly equal to $E^\sigma\left(d_n(m_n^\sigma, m_{-n}, \theta)\,|t_n\right)$. Therefore, an alternative system $((M,d'),\sigma)$, where $d' = d$ except that

$$d'_n(m,\theta) = d_n(m_n^\sigma, m_{-n}, \theta)$$

for all (m,θ), is just as good for the principal (recall that the principal's expected payoff is a function of the expected payoff of each agent, when guilty and when innocent). System $((M,d'),\sigma)$ is incentive compatible by definition and it is renegotiation proof because, for any (m_{-n},θ),

$$h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma(m_{-n}),\theta\right) \leq h_n\left(m_n, r_{-n}^\sigma(m_{-n}),\theta\right)$$

for all m_n sent with positive probability.

So, then the issue is how to choose punishments that follow message m_n^σ. I claim that, no matter which vector (m,θ) is observed by the principal, she would prefer to impose punishment $h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma(m_{-n}),\theta\right)$ over punishment $d_n(m_n^\sigma, m_{-n},\theta)$, which proves the statement.

If $m_n = m_n^\sigma$, this follow directly by definition:

$$h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma(m_{-n}),\theta\right) = h_n\left(r_n^\sigma(m_n^\sigma), r_{-n}^\sigma(m_{-n}),\theta\right)$$

which is defined to be the principal's preferred punishment conditional on observing vector $(m_n^\sigma, m_{-n},\theta)$. If $m_n \neq m_n^\sigma$, the principal would prefer to select the punishment

that is the closest to $h_n\left(r_n^\sigma(m_n), r_{-n}^\sigma(m_{-n}), \theta\right)$. Seeing as

$$h_n\left(r_n^\sigma(m_n), r_{-n}^\sigma(m_{-n}), \theta\right) \geq h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma(m_{-n}), \theta\right) \geq d_n(m_n^\sigma, m_{-n}, \theta)$$

the closest is $h_n\left(\upsilon_n^\sigma, r_{-n}^\sigma(m_{-n}), \theta\right)$. The argument is complete by applying the same logic to each agent n, which results in system $((M, d^\sigma), \sigma)$. ■

Following Lemma 2.3., the problem of finding the optimal system becomes solely the problem of finding the optimal strategy profile $\sigma : \{i, g\} \to \Delta\{\mathbb{R}_+ \cup \{c\}\}$. In other words, of all the possible σ, one must find the one that maximizes the principal's expected utility under the assumption that the corresponding system is $((M, d^\sigma), \sigma)$, which is incentive compatible and renegotiation proof by definition. In the next lemma, I show that there is an optimal strategy profile σ where only two messages are sent with positive probability, and one of them is only sent by the guilty agent. Given that messages that are not sent with positive probability can be deleted, the message set of each agent only needs to contain those two messages, labeled c and $\overline{c}$.

Lemma 2.4. *There is an optimal RPIC system* $\left(\left(M^{RP}, d^{RP}\right), \sigma^{RP}\right)$ *such that, for all* n*, i)* $M_n^{RP} = \{c, \overline{c}\}$*, ii)* $\sigma_n(i)(c) = 0$ *and iii)* $d_n^{RP}(m, \theta) = \gamma_n^{\sigma^{RP}}(\overline{c}, m_{-n}, \theta)$ *for all* $(m, \theta) \in M^{RP} \times \Theta$*.*

Proof. Take any system $((M, d^\sigma), \sigma)$ where $M_n = \mathbb{R}_+ \cup \{c\}$. Take any agent n. I show how to construct an alternative RPIC system $\left(\left(M, d^{\sigma'}\right), \sigma'\right)$ where i) $\sigma'_{\widehat{n}} = \sigma_{\widehat{n}}$ for all $\widehat{n} \neq n$, ii) $\sigma'_n(i)(c) = \sigma_n(i)(c) = 0$ and iii) $\sigma'_n(t_n)(\upsilon_n^\sigma) + \sigma'_n(t_n)(c) = 1$ for $t_n = i, g$. By doing the same to all agents, the result follows.

For every m_n sent with positive probability under σ, let $a_n^\sigma(m_n) \geq 0$ be such that

$$\sigma_n(g)(m_n) - a_n^\sigma(m_n) = \upsilon_n^\sigma \sigma_n(i)(m_n)$$

and define σ''_n as follows: $\sigma''_{\widehat{n}} = \sigma_{\widehat{n}}$ for all $\widehat{n} \neq n$, $\sigma''_n(i) = \sigma_n(i)$ but

$$\sigma''_n(g)(m_n) = \begin{cases} \sigma_n(g)(m_n) - a_n^\sigma(m_n) & \text{if } m_n \neq c \\ \sigma_n(g)(m_n) + \sum_{m_n} a_n^\sigma(m_n) & \text{if } m_n \neq c \end{cases}$$

In words, in σ''_n, agent n confesses more than under system σ_n. In particular, each message m_n that was not the one that induced the largest belief of innocence is less

likely to be sent by the guilty type, to the extent that all non-confessing messages (all messages different than c) lead to the same posterior belief under σ_n''. Therefore, by Lemma 2.1., it follows that such a system is equivalent to the system $\left(\left(M, d^{\sigma'}\right), \sigma'\right)$ described above (because all non-confessing messages can be merged). Therefore, the proof is complete by showing that system $\left(\left(M, d^{\sigma''}\right), \sigma''\right)$ is weakly preferred by the principal to system $\left(\left(M, d^{\sigma}\right), \sigma\right)$.

First, let us think about the expected utility that the principal gets from agent n. That depends exclusively on agent n's expected punishment for each type, which, in turn, only depends on the smallest likelihood ratio of the messages that are sent with positive probability. But the smallest likelihood ratio under σ_n and under σ_n'' is the same: by definition, $\upsilon_n^{\sigma} = \upsilon_n^{\sigma''}$. So, the expected utility that the principal gets from agent n is the same under σ_n and σ_n''.

Now, let us think about the expected utility that the principal gets from some agent $\widehat{n} \neq n$. The difference for agent $\widehat{n}$ will be that each message $m_n \neq m_n^{\sigma}$ sent by agent n is split into two. So, for the result to follow it is enough to show that this split is beneficial for the principal.

Fix any $(m_{-n,\widehat{n}}, \theta)$ and take any $m_n \neq m_n^{\sigma}$ sent with positive probability. The expected utility that the principal gets from agent $\widehat{n}$ under σ, conditional on receiving $(m_n, m_{-n,\widehat{n}}, \theta)$, is given by $\chi\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma}, r_n^{\sigma}\left(m_n\right), -\right)\right)$, where

$$\chi\left(x_{\widehat{n}}\right) = \left\{ \begin{array}{c} \left(\begin{array}{c} \xi_n \sigma_n\left(g\right)\left(m_n\right) \Pr\left\{m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = g\right\} + \\ \left(1-\xi_n\right) \sigma_n\left(i\right)\left(m_n\right) \Pr\left\{t_n = i, m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = i\right\} \end{array} \right) \upsilon_{\widehat{n}}\left(i, x_{\widehat{n}}\right) + \\ \left(1-\xi_n\right) \sigma_n\left(i\right)\left(m_n\right) \Pr\left\{t_n = g, m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = i\right\} \upsilon_{\widehat{n}}\left(g, x_{\widehat{n}}\right) \end{array} \right.$$

Notice that

$$\chi\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma}, r_n^{\sigma}\left(m_n\right), -\right)\right) = \widehat{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma}, r_n^{\sigma}\left(m_n\right), -\right)\right) + \widetilde{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma}, r_n^{\sigma}\left(m_n\right), -\right)\right)$$

where

$$\widehat{\chi}\left(x_{\widehat{n}}\right) = \left\{ \begin{array}{c} \left(\begin{array}{c} \xi_n \left(\sigma_n\left(g\right)\left(m_n\right) - a_n^{\sigma}\left(m_n\right)\right) \Pr\left\{m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = g\right\} + \\ \left(1-\xi_n\right) \sigma_n\left(i\right)\left(m_n\right) \Pr\left\{t_n = i, m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = i\right\} \end{array} \right) \upsilon_{\widehat{n}}\left(i, x_{\widehat{n}}\right) + \\ \left(1-\xi_n\right) \sigma_n\left(i\right)\left(m_n\right) \Pr\left\{t_n = g, m_{-n,\widehat{n}}, \theta | t_{\widehat{n}} = i\right\} \upsilon_{\widehat{n}}\left(g, x_{\widehat{n}}\right) \end{array} \right.$$

and

$$\widetilde{\chi}\left(x_{\widehat{n}}\right)=a_{n}^{\sigma}\left(m_{n}\right)\Pr\left\{m_{-n,\widehat{n}},\theta|t_{\widehat{n}}=g\right\}v_{\widehat{n}}\left(i,x_{\widehat{n}}\right)$$

Under $\widehat{\sigma}$, there are two differences: first, when agent $\widehat{n}$ is guilty, there is a shift in $a_{n}^{\sigma}\left(m_{n}\right)$ towards confessing; and second, this causes the payoffs that follow message m_{n} to change (because $r_{n}^{\sigma''}\left(m_{n}\right)=r_{n}^{\sigma}\left(m_{n}^{\sigma}\right)\leq r_{n}^{\sigma}\left(m_{n}\right)$). As a result, it is enough to show that

$$\widehat{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},r_{n}^{\sigma}\left(m_{n}\right),-\right)\right)+\widetilde{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},r_{n}^{\sigma}\left(m_{n}\right),-\right)\right)\leq\widehat{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},\upsilon_{n}^{\sigma},-\right)\right)+\widetilde{\chi}\left(0\right)$$

which shows the statement.

Notice that

$$h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},r_{n}^{\sigma}\left(m_{n}\right),-\right)\leq h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},\upsilon_{n}^{\sigma},-\right)\leq h_{\widehat{n}}\left(1,\upsilon_{n}^{\sigma},-\right)$$

where the first inequality follows because there is negative correlation between the agents' types: if agent n is more likely to be innocent, agent $\widehat{n}$ is more likely to be guilty. Seeing as

$$h_{\widehat{n}}\left(1,\upsilon_{n}^{\sigma},-\right)\in\arg\max_{x_{\widehat{n}}\in[0,1]}\ \widehat{\chi}\left(x_{\widehat{n}}\right)$$

and $\widehat{\chi}\left(\cdot\right)$ is (weakly) concave, it follows that

$$\widehat{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},r_{n}^{\sigma}\left(m_{n}\right),-\right)\right)\leq\widehat{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},\upsilon_{n}^{\sigma},-\right)\right)$$

It is also the case that

$$0=\arg\max_{x_{n}\in[0,1]}\ \widetilde{\chi}\left(x_{n}\right)$$

which implies that

$$\widetilde{\chi}\left(h_{\widehat{n}}\left(\upsilon_{\widehat{n}}^{\sigma},r_{n}^{\sigma}\left(m_{n}\right),-\right)\right)\leq\widetilde{\chi}\left(0\right)$$

which completes the proof. ■

Finally, the last step is to transform the optimal system of Lemma 2.4. into a CIS.

Lemma 2.5. *There is an optimal RPIC system* $\left(\left(M^{RP},d^{RP}\right),\sigma^{RP}\right)$ *such that, for all* n*, i)* $M_{n}^{RP}=\{c,\overline{c}\}$*, ii)* $\sigma_{n}\left(i\right)\left(c\right)=0$*, iii)* $d_{n}^{RP}\left(\overline{c},m_{-n},\theta\right)=\gamma_{n}^{\sigma^{RP}}\left(\overline{c},m_{-n},\theta\right)$

for all $(m_{-n},\theta)\in M_{-n}^{RP}\times\Theta$ *and iv)*

$$d_n^{RP}(c,m_{-n},\theta)=E^{\sigma^{RP}}\left(\gamma_n^{\sigma^{RP}}(\overline{c},m_{-n},\theta)\,|t_n=g\right)$$

for all $(m_{-n},\theta)\in M_{-n}^{RP}\times\Theta$.

Proof. Take the optimal RPIC system of Lemma 2.4. denoted by $((M,d^{\sigma}),\sigma)$. Consider system $((M,d'),\sigma)$, where $d'_n(\overline{c},m_{-n},\theta)=d_n^{\sigma}(\overline{c},m_{-n},\theta)$ for all (m_{-n},θ) and n, but,

$$d'_n(c,m_{-n},\theta)=E^{\sigma}(\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)\,|t_n=g)$$

for all (m_{-n},θ) and n. This new system is just as good for the principal, because the expected punishment of each agent of each type is the same as with $((M,d^{\sigma}),\sigma)$. It is also trivially renegotiation proof because

$$E^{\sigma}(\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)\,|t_n=g)\leq\gamma_n^{\sigma}(c,m_{-n},\theta)=1$$

for all (m_{-n},θ) and n. So, the proof is complete by showing that system $((M,d'),\sigma)$ is incentive compatible. By definition, the guilty type of each agent is indifferent between the two reports, so what is left is to show that the innocent type does not strictly prefer to report c, i.e., I show that

$$E^{\sigma}(\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)\,|t_n=g)\geq E^{\sigma}(\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)\,|t_n=i)$$

for all n.

Take any n and recall that, for all (m_{-n},θ) and n, one can write

$$\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)=\begin{cases}1\text{ if }\frac{\Pr^{\sigma}\{t_n=g|\overline{c},m_{-n},\theta\}}{\Pr^{\sigma}\{t_n=i|\overline{c},m_{-n},\theta\}}\geq\alpha_n\\ 0\text{ otherwise}\end{cases}$$

which can be rewritten as

$$\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)=\begin{cases}1\text{ if }\frac{\xi_n\sigma_n(g)(\overline{c})}{(1-\xi_n)\sigma_n(i)(\overline{c})}\frac{\Pr^{\sigma}\{m_{-n},\theta|t_n=g\}}{\Pr^{\sigma}\{m_{-n},\theta|t_n=i\}}\geq\alpha_n\\ 0\text{ otherwise}\end{cases}$$

Let

$$E_n\equiv\{(m_{-n},\theta):\gamma_n^{\sigma}(\overline{c},m_{-n},\theta)=1\}$$

If $E_n = \varnothing$, then

$$E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = g\right) = E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right) = 0$$

If $\rceil E_n = \varnothing$, then

$$E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = g\right) = E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right) = 1$$

If $\frac{\Pr^\sigma\{m_{-n},\theta|t_n=g\}}{\Pr^\sigma\{m_{-n},\theta|t_n=i\}} > 1$ for all $(m_{-n}, \theta) \in E_n$ then

$$E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = g\right) > E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right)$$

Finally, if there is $(m_{-n}, \theta) \in E_n$ such that $\frac{\Pr^\sigma\{m_{-n},\theta|t_n=g\}}{\Pr^\sigma\{m_{-n},\theta|t_n=i\}} \leq 1$ and given that $\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)$ is increasing with $\frac{\Pr^\sigma\{m_{-n},\theta|t_n=g\}}{\Pr^\sigma\{m_{-n},\theta|t_n=i\}}$, then it must be that $\frac{\Pr^\sigma\{m_{-n},\theta|t_n=g\}}{\Pr^\sigma\{m_{-n},\theta|t_n=i\}} < 1$ for all $(m_{-n}, \theta) \notin E_n$. Hence,

$$\Pr^\sigma\left\{(m_{-n}, \theta) \notin E_n | t_n = g\right\} < \Pr^\sigma\left\{(m_{-n}, \theta) \notin E_n | t_n = i\right\}$$

which implies that

$$\Pr^\sigma\left\{(m_{-n}, \theta) \in E_n | t_n = g\right\} > \Pr^\sigma\left\{(m_{-n}, \theta) \in E_n | t_n = i\right\}$$

and so

$$E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = g\right) > E^\sigma\left(\gamma_n^\sigma\left(\overline{c}, m_{-n}, \theta\right)|t_n = i\right)$$

■

References

[1] Baker, Scott, and Claudio Mezzetti. "Prosecutorial resources, plea bargaining, and the decision to go to trial." Journal of Law, Economics, & Organization (2001): 149-167.

[2] Banerjee, Abhijit V. "A theory of misgovernance." The Quarterly Journal of Economics (1997): 1289-1332.

[3] Bar-Gill, Oren, and Omri Ben-Shahar. "The Prisoners'(Plea Bargain) Dilemma." Journal of Legal Analysis 1.2 (2009): 737-773.

[4] Becker, Gary S. "Crime and Punishment: An Economic Approach." The Journal of Political Economy (1968): 169-217.

[5] Ben-Porath, Elchanan, Eddie Dekel, and Barton L. Lipman. "Optimal Allocation with Costly Verification." The American Economic Review 104.12 (2014).

[6] Bester, Helmut, and Roland Strausz. "Imperfect commitment and the revelation principle: the multi-agent case." Economics Letters 69.2 (2000): 165-171.

[7] Bjerk, David. "Guilt shall not escape or innocence suffer? The limits of plea bargaining when defendant guilt is uncertain." American Law and Economics Review 9.2 (2007): 305-329.

[8] Dervan, Lucian E., and Vanessa A. Edkins. "The Innocent Defendant's Dilemma: An Innovative Empirical Study of Plea Bargaining's Innocence Problem." J. Crim. L. & Criminology 103 (2013): 1.

[9] Franzoni, Luigi Alberto. "Negotiated enforcement and credible deterrence." The Economic Journal 109.458 (1999): 509-535.

[10] Garoupa, Nuno. "The theory of optimal law enforcement." Journal of economic surveys 11.3 (1997): 267-295.

[11] Grossman, Gene M., and Michael L. Katz. "Plea bargaining and social welfare." The American Economic Review (1983): 749-757.

[12] Harrington Jr, Joseph E. "Optimal corporate leniency programs." The Journal of Industrial Economics 56.2 (2008): 215-246.

[13] Kaplow, Louis, and Steven Shavell. "Optimal Law Enforcement with Self-Reporting of Behavior." Journal of Political Economy 102.3 (1994).

[14] Kim, Jeong-Yoo. "Secrecy and fairness in plea bargaining with multiple defendants." Journal of Economics 96.3 (2009): 263-276.

[15] Kim, Jeong-Yoo. "Credible plea bargaining." European Journal of Law and Economics 29.3 (2010): 279-293.

[16] Kim, Jeong-Yoo, and Nathan Berg. "Plea Bargaining with Multiple Defendants and Its Deterrence Effect". Working paper (2016).

[17] Kobayashi, Bruce H. "Deterrence with multiple defendants: an explanation for" Unfair" plea bargains." The RAND Journal of Economics (1992): 507-517.

[18] Lewis, Tracy R., and David EM Sappington. "Motivating wealth-constrained actors." American Economic Review (2000): 944-960.

[19] Midjord, Rune. "Competitive Pressure and Job Interview Lying: A Game Theoretical Analysis". Working paper (2013).

[20] Motta, Massimo, and Michele Polo. "Leniency programs and cartel prosecution." International journal of industrial organization 21.3 (2003): 347-379.

[21] Myerson, Roger B. "Incentive compatibility and the bargaining problem." Econometrica (1979): 61-73.

[22] Mylovanov, Tymofiy, and Andriy Zapechelnyuk. "Mechanism Design with ex-post Verification and Limited Punishments", *mimeo* (2014).

[23] Posner, Richard A. "An economic approach to the law of evidence." Stanford Law Review (1999): 1477-1546.

[24] Siegel, Ron and Bruno Strulovici. "Judicial Mechanism Design", *Working paper* (2018).

[25] Spagnolo, G. "Leniency and whistleblowers in antitrust." *Handbook of antitrust economics*. MIT Press (2008): Chpt 12

[26] Tor, Avishalom, Oren Gazal-Ayal, and Stephen M. Garcia. "Fairness and the willingness to accept plea bargain offers." Journal of Empirical Legal Studies 7.1 (2010): 97-116.