INSTITUTO DE ECONOMÍA

DOCUMENTO de TRABAJO

500

2017

If we Confess our Sins

Francisco Silva

www.economia.puc.cl • ISSN (edición impresa) 0716-7334 • ISSN (edición electrónica) 0717-7593

If we Confess our Sins[*]

Francisco Silva[†]

November 28, 2017

Abstract

I consider the problem a social planner faces in constructing a criminal justice system which addresses two needs: to protect the innocent and to punish the guilty. I characterize the socially optimal criminal justice system under various assumptions with respect to the planner's ability to commit. In the optimal system, before a criminal investigation is initiated, all members of the community are given the opportunity to confess to having committed the crime in exchange for a smaller than socially optimal punishment, independent of any future evidence. Leniency for confessing agents is efficient because there are informational externalities to each confession.

JEL classification: D82, K14

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

[*] I would like to thank David Abrams, Mustafa Dogan, Selman Erol, 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

In this paper, I study how to design a criminal justice system in order to most efficiently collect the necessary information to identify and appropriately punish those who are guilty of committing a crime. I consider a scenario where there is a community of N agents and a principal, who is thought of as some kind of planner or benevolent decision maker. She is responsible for administering criminal justice, which means that, whenever there is a suspicion that a crime has been committed, it is her responsibility to select whom to punish and the extent of that punishment. In a perfect world, she would punish only agents who are guilty of committing the crime but, of course, the problem is that the principal does not know who is guilty and who is innocent. And, knowing that the principal is interested in punishing those agents who are guilty makes them reluctant to announce their guilt. I study the principal's problem of creating a mechanism that, to the extent that is possible, punishes those who are guilty while protecting the rights of the innocent.

The traditional solution for this problem is a "trial system". In a trial system, if the principal suspects the crime has been committed, she initiates an investigation aimed at obtaining evidence. Based on the evidence, the principal forms beliefs about the guilt of each agent and chooses punishments accordingly. The merit of this system is that the evidence is more likely to point to guilt if the agent is indeed guilty than if he is not.

In this paper, however, I argue that trial systems are not optimal. There are other systems which generate a larger social welfare, which will be understood as a weighted average between society's desire to appropriately punish those who are guilty and to protect those who are innocent. In particular, the optimal system will be a "confession inducing system" (CIS). A CIS has two stages. In the first stage, before the investigation begins, all agents are given the opportunity to confess the crime, in exchange for a constant and guaranteed punishment, independent of any evidence which might be gathered in the future. In the second stage, if necessary, the principal conducts an investigation, and, based on the information gathered, chooses the punishments, if any, to apply to agents who chose not to confess in the first stage.

Variants of this system exist already in American law. The closest system to the one this paper suggests is "self-reporting" in environment law. The idea behind self-reporting 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. Another system with some similarities is plea bargaining in criminal law, where defendants are given the chance to confess to having committed the crime in exchange for a reduced sentence.

This type of systems has received quite a lot of attention in the literature on the economics of crime and law enforcement, which has highlighted some of its advantages.[1] This paper contributes to this literature in two ways.

First, in its approach. Unlike most of the literature, which performs pairwise comparisons between the trial system and an alternative system (like plea bargaining or self-reporting), I use some of the techniques from mechanism design to find the optimal system.[2] I believe this is an important contribution in that it makes unnecessary the pursuit of a better system, at least in the context of my model.

Second, I identify an informational advantage that these systems have which has not yet been accounted for. The main idea that I explore in the paper is that there are information externalities to each confession. By confessing, an agent provides the principal with information not only about himself but indirectly about the other agents: that they are more or less likely to be guilty. The principal is then able to use this information to make better judgments about the *other* agents' guilt. For most crimes, by their nature, one would expect that the principal's prior belief is that there is at most one criminal, whose identity is unknown to her. In such a scenario, a credible confession by some agent would allow the principal to become aware that all other agents are very likely to be innocent and thus avoid making the mistake of punishing them. It is also easy to think of crimes where there is positive correlation between the agents' guilt. For example, the confession from some firm's owner of having reported a smaller profit when filling out the previous year's tax returns, in order to pay less taxes, might inform the principal that other firms in the same sector might have followed a similar practice.

Naturally, the informational benefit of CIS's depends on there being multiple agents who could have conceivably committed the crime. This is always going to be the case if the opportunity to confess is given early enough in the criminal process,

[1] See the related literature section for an overview.

[2] In an independent work, Siegel and Strulovici (2016) follow a similar approach, which I discuss in more detail in the related literature section.

even before an investigation has been initiated, because, at such an early stage, essentially everyone is a suspect. As a result, a system like "self-reporting", where the initiative to confess comes from the agents, is more suited to explore these information externalities than a system like plea bargaining, where it is the prosecutor who, further along in the criminal process, approaches the typically single agent to seek a confession.

In the first part of the paper, I assume that the principal has commitment power and that there is an upper bound on the punishments that she can choose. In such a framework, the revelation principle holds, and so it follows that there is an optimal mechanism where each agent finds it desirable to report their "type" - innocent or guilty. I show that the punishment after a report of "guilty" (a confession) is constant, which implies that there is a CIS which is optimal. Furthermore, I characterize how punishments are allocated to each agent in the optimal CIS depending on the prior belief of the principal and on the evidence gathered. In particular, in the optimal CIS, if an agent refuses to confess, his punishment is the one that maximizes the principal's expected utility, conditional on all the information that she can gather from the evidence and from the reports of all *other* agents. If an agent confesses, the constant punishment he receives leaves him indifferent to refusing to confess if and only if he is guilty. Innocent agents refuse to confess because they know that they are innocent and have more reason than guilty agents to believe that the evidence will support their claim of innocence.

Having commitment power allows the principal to i) impose small punishments on agents who are known to be guilty (those who confess), and ii) punish agents who are known to be innocent (those who refuse to confess). Assumption ii) seems particularly problematic. Notice that, simply by observing that the agent has chosen not to confess, the principal is able to infer that he is likely to be innocent. And yet, the optimal CIS often requires the principal to ignore this belief and punish the agent (if refusing to confess never led to any punishment no one would ever confess). In light of this, in the second part of the paper, I consider the principal's problem when she has limited commitment power. I consider two cases.

First, I consider the class of renegotiation proof mechanisms, where only i) is permitted.[3] The problem of characterizing the optimal mechanism becomes considerably

[3] I call these mechanisms renegotiation proof because if the principal is supposed to punish an

harder as the revelation principle no longer holds, so one cannot simply restrict attention to revelation mechanisms, especially in a context with multiple agents.[4] I show that there is a CIS that is approximately optimal among renegotiation proof mechanisms provided that the prior probability that more than one agent has committed the crime is sufficiently small. Furthermore, for any prior belief about the agents' guilt, it is always possible to construct a renegotiation proof CIS that does better than any trial system, which shows that the superiority of CIS's with respect to trial systems does not depend on assumption ii).

Finally, I consider sequentially optimal mechanisms, where the principal has no commitment power and so neither i) nor ii) are permitted. In this setup, I show that it is not possible to improve upon the trial system as confessions are no longer sustainable, because an agent who is revealed to be guilty is shown no leniency.

The paper is organized as follows. In section 2, I provide an overview of the related literature. In section 3, I describe the model. In section 4, I discuss a simple example. In section 5, I study trial mechanisms. In section 6, I study the principal's problem of finding the optimal mechanism when she has commitment power, while, in section 7, I consider the case where her commitment power is limited. In section 8, I conclude.

In the appendix, I consider several extensions: in appendix A1, I study what changes if it is possible that the crime is committed by an organized group of agents; in appendix A2, I study the implications of assuming that agents are risk averse (as opposed to risk neutral); in appendix A3, I show that there is no loss of generality in not modelling each agent's decision of whether to commit the crime; in appendix A4, I consider a different timing, where the principal is only able to negotiate with the agents after the investigation has been completed; and, finally, in appendix A5, I study what happens if additional heterogeneity among the agents is added to the model.

agent she knows is innocent, both her and the agent would have an incentive to renegotiate such punishment, as they would both prefer a smaller one.

[4] See Bester and Strausz (2000).

2 Related Literature

There is a considerable amount of literature in economics that argues for the use of variants of CIS's in law enforcement. Kaplow and Shavell (1994) add a stage, where agents can confess to be guilty, to a standard model of negative externalities and argue that this improves social welfare because it saves monitoring costs. By setting the punishment after a confession to be 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 (2016) 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, at least in main text, 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 CIS's explore the correlation between the agents' innocence.

A feature common to these papers is that they have assumed that the law enforcer has commitment power. There have been different articles, particularly in the plea bargaining literature, that have discussed the implications of limiting that commitment power. Baker and Mezzetti (2001) assume that the 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. This is because, if all guilty agents confessed, there would be no incentive for the prosecutor to exert any effort, which, in turn, would induce the guilty agents not to confess. This type of equilibrium is a common occurrence when limited commitment power is assumed - see for example Kim (2010), Franzoni (1999) or Bjerk (2007). In section 7, I consider the implications of reducing the principal's commitment power and find that the optimal mechanism has this same feature: in equilibrium, there is a positive probability that guilty agents prefer not to confess.

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. There is also a literature on industrial organization that considers the design of leniency programs in Antitrust law which also considers multiple agents - see Spagnolo (2006) for a literature review.

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.[5]

3 Model

3.1 Fundamentals

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 $T = T_1 \times ... \times T_N$ 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, so that if $t \in T$, then $t_{-n} = (t_1, ..., t_{n-1}, t_{n+1}, ..., t_N) \in T_{-n}$. The ex-ante probability that vector t is realized is denoted by $\pi(t) > 0$ for all $t \in T$ and is assumed to be common knowledge.

This description implicitly assumes that each agent knows only whether he is innocent or guilty, and has no other relevant information about the other agents' innocence. Thus, it rules out crimes which are likely to have been committed by an organized group of agents (conspiracy crimes). For example, imagine that agents 1 and 2 rob a bank together. It would be very likely that agent 1 would know that both him and agent 2 are guilty of committing the crime. In Appendix A1, I extend the model in order to consider this type of information structure and show that the same intuition of the simpler model considered in the main text carries through.

Evidence: After t has been drawn, each agent n is randomly assigned an evidence level $\theta_n \in [0, 1]$. Let $\Theta_n = [0, 1]$ and $\Theta = [0, 1]^N$ denote the set of all possible evidence vectors, while Θ_{-n} denotes the set of all possible evidence vectors that exclude only agent n's evidence level. The evidence vector $\theta = (\theta_1, ..., \theta_N)$ is made of exogenous

[5] 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 under the assumption that the principal has commitment power.

signals correlated with the agents' guilt and is interpreted as the product of a criminal investigation.

I assume that each θ_n only depends on agent n's innocence - $\theta_n|t_n$ is independent of t_{-n} - and denote the conditional probability density function (pdf) of θ_n by $\pi\left(\theta_n|t_n\right)$, while the joint conditional pdf of θ given t is denoted by $\pi\left(\theta|t\right) = \prod\limits_{n=1}^{N} \pi\left(\theta_n|t_n\right)$. (For expositional purposes, I have abused notation by using π to represent probability measures over different spaces, but this will lead to no confusion).

Even though I have assumed that each agent n generates its own signal θ_n, this does not mean that every agent in the community is personally investigated. For example, gathering evidence can be checking for fingerprints near the crime scene. Even if the fingerprints of agent n are not found, this information is still contained in θ_n. Also, the assumption of conditional independence of $\theta_n|t_n$ is mostly made out of expositional simplicity as no result depends on it. In particular, notice that it does not imply that θ_n is independent of θ_{-n}.

Let $l\left(\theta_n\right) = \frac{\pi(\theta_n|t_n=g)}{\pi(\theta_n|t_n=i)}$ be the evidence likelihood ratio. I assume that l is differentiable and strictly increasing. This implies that the larger the realized θ_n is, the more likely it is that agent n is guilty. I also assume that $\lim_{\theta_n \to 0} l\left(\theta_n\right) = 0$ and $\lim_{\theta_n \to 1} l\left(\theta_n\right) = \infty$, which means that, as long as the principal is not completely certain of agent n's guilt, there is always some evidence level θ_n that changes his mind - there is always some θ_n such that the posterior probability of guilt can be made arbitrarily close to either 0 or 1.

Preferences: I assume that each agent n's utility is given by $u^a\left(x_n\right) = -x_n$, where $x_n \in \mathbb{R}_+$ represents the punishment agent n receives - it could be time in prison, community service time, physical punishment or a monetary fine. Each agent simply wants to minimize the punishments inflicted upon him. I make the assumption that agents are risk neutral in order to distinguish my argument from the one, for example, of Grossman and Katz (1983) (see the related literature section). In Appendix A2, I analyze the case where agents are risk averse.

As for the principal, she is thought of as a sort of social planner or benevolent decision maker and her preferences are supposed to represent society's preferences. Following Grossman and Katz (1983), I assume that her utility depends not only on the punishment she inflicts but also on whether the agent who receives it is innocent or guilty. In particular, I assume that the principal's utility function is given by

$$u^p\left(t,x\right)=\sum_{n=1}^{N}u_n^p\left(t_n,x_n\right) \text{ for all } t\in T \text{ and } x=\left(x_1,..,x_N\right)\in\mathbb{R}_+^N\text{, where}$$

$$u_n^p\left(t_n,x_n\right)=\begin{cases} -\infty \text{ if } x_n>\phi \\ -\alpha x_n \text{ if } t_n=i \text{ and } x_n\leq\phi \\ -\left|1-x_n\right| \text{ if } t_n=g \text{ and } x_n\leq\phi \end{cases}$$

with $\alpha>0$ and $\phi\geq1$. Parameter ϕ serves as an upper bound on the punishments that the principal may impose. The principal is assumed to be either unable or unwilling to impose punishments larger than ϕ.[6]

If agent n is innocent, the principal prefers to acquit him, while if he is guilty, the principal prefers to punish him to the extent of the crime, which I normalize to 1. In either case, deviations from the preferred punishment induce a linear cost to the principal. This punishment of 1 that "fits the crime" is exogenous to the model and is likely to be influenced by the nature of the crime - the punishment that fits the crime of murder is larger than the punishment that fits the crime of minor theft. The parameter α captures the potentially different weights that these interests may have - α is large if the principal is more concerned with wrongly punishing innocent agents and is small if she is more concerned with wrongly acquitting guilty agents.

The way the utility 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. Essentially, it is assumed that it is desirable to punish only the guilty agents at a level proportional to the severity of the crime. This is compatible, for example, with the idea that the State wishes to punish guilty agents to prevent them from committing further crimes. It is also compatible with wanting to punish guilty agents to satisfy a public need of revenge. And, as I make explicit in Appendix A3, it is also possible to interpret the principal's preferences as representing society's desire to deter crime. In particular, in Appendix A3, I show that there is no loss in generality in not modelling the agents' decision of whether to commit a crime. The basic logic is that, if the principal wants to deter crime, she should want to maximize the punishments of those who are guilty and minimize the punishments of those who are innocent, which can be accomplished by making $\phi=1$ and properly choosing α.

[6] In the text, I characterize the optimal mechanism for any level of ϕ, including the limit case where $\phi\rightarrow\infty$.

3.2 Definitions

The timing is the following. Before any evidence is generated (and, so, before an investigation has been initiated), the principal selects a mechanism. Given the mechanism, each agent n chooses to send a message $m_n \in M_n$ to the principal, where M_n is assumed to be arbitrarily large for all n. Let $M = M_1 \times ... \times M_N$ and interpret M_{-n} in the usual way.[7] Each agent is assumed to know the distribution of θ but not the actual realization of θ before choosing his message.[8]

A mechanism $x : M \times \Theta \to \mathbb{R}_+^N$ is a map from the agents' messages and evidence to non-negative punishments to be allocated to each agent. Each agent's strategy is a probability distribution over his message space M_n for each type, which I denote by $\sigma_n(t_n, \cdot)$ for $t_n \in \{i, g\}$. Vector $\sigma = (\sigma_1, ..., \sigma_N)$ represents the strategy profile of the N agents, while the set of all of strategy profiles is denoted by Φ. Each profile (x, σ) is called a system.

Definition 1 *A trial system (x, σ) is such that x is independent of $m \in M$.*

In a trial system, the punishment that each agent receives does not depend on his report but only on the evidence produced.

Let c and $\overline{c}$ be two messages from set M_n for each agent n, i.e. $c \in M_n$ and $\overline{c} \in M_n$ for all n.

Definition 2 *A CIS (x, σ) is such that*
i) $x_n(c, m_{-n}, \theta)$ is independent of $m_{-n} \in M_{-n}$ and $\theta \in \Theta$ for all n.
ii) $\sigma_n(t_n, m_n) = 0$ for all $m_n \notin \{c, \overline{c}\}$, for $t_n \in \{i, g\}$ and for all n.

One can interpret a CIS as having two stages. In the first stage, before any investigation has been initiated, an agent is given the opportunity to confess to be guilty (and send message c). If the agent confesses, he receives a constant punishment. If

[7] In appendix A4, I consider a slightly different timing where the mechanism is selected after the evidence has been generated and show, that, if possible, the principal would prefer to select the mechanism before an investigation is initiated.

[8] In Appendix A5, I relax the assumption that every agent and the principal have the same prior knowledge about the evidence.

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 (x, σ) is a CIS, I say that x is a confession inducing mechanism.

Definition 3 *A system (x, σ) is renegotiation proof if and only if, for all $m \in M$, $\theta \in \Theta$ and n,*

$$x_n(m, \theta) \leq \min \left\{ \arg\max_{x_n \geq 0} \ E^{\sigma}\left(u_n^p(t_n, x_n) | m, \theta\right) \right\}$$

The set

$$\arg\max_{x_n \geq 0} \ E^{\sigma}\left(u_n^p(t_n, x_n) | m, \theta\right)$$

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 any (m, θ), the principal (weakly) prefers to choose a smaller punishment, because such a reduction would be promptly accepted by the agent. If there is more than one maximizer, it is assumed that the punishment must be smaller than the smallest maximizer.[9] If system (x, σ) is renegotiation proof, then I say that mechanism x is renegotiation proof.

Definition 4 *A system (x, σ) is sequentially optimal if and only if, for all $m \in M$, $\theta \in \Theta$ and n„*

$$x_n(m, \theta) \in \arg\max_{x_n \geq 0} \ E^{\sigma}\left(u_n^p(t_n, x_n) | m, \theta\right)$$

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 $x_n(m, \theta)$. If system (x, σ) is sequentially optimal, then I say that mechanism x is sequentially optimal.

Definition 5 *A system (x, σ) is incentive compatible if and only if σ is a Bayes-Nash equilibrium of the game induced by mechanism x.*

[9] The motivation is that, otherwise, the principal would not be made worse off by a reduction in the punishment but the agent would be strictly better off. Nevertheless, the events where there are multiple maximizers has measure 0.

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

$$\begin{aligned} & -\int\limits_{\theta\in\Theta}\int\limits_{m_{-n}\in M_{-n}} \pi^{\sigma}(m_{-n},\theta|t_n)\, x_n(m_n, m_{-n},\theta)\, dm_{-n} d\theta \\ \geq & -\int\limits_{\theta\in\Theta}\int\limits_{m_{-n}\in M_{-n}} \pi^{\sigma}(m_{-n},\theta|t_n)\, x_n(m_n', m_{-n},\theta)\, dm_{-n} d\theta \end{aligned}$$

for all $m_n' \in M_n$, where $\pi^{\sigma}(m_{-n},\theta|t_n)$ represents the conditional joint density of $(m_{-n},\theta) \in M_{-n}\times\Theta$, given agent n's type t_n and strategy profile σ.

4 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 and her prior beliefs about each of the two agents' guilt is given by

π	$t_2 = i$	$t_2 = g$
$t_1 = i$	$\frac{1}{10}+\delta$	$\frac{4}{10}$
$t_1 = g$	$\frac{4}{10}$	$\frac{1}{10}-\delta$

for $\delta \in \left[0, \frac{1}{10}\right)$. Each entry in the table represents the prior probability that event $t = (t_1, t_2)$ is realized. So, for example, the prior probability that both agents are innocent and that the fire was simply an accident is $\pi(i,i) = \frac{1}{10}+\delta$.

Assume that the conditional distribution of each evidence level $\theta_n \in [0,1]$ is given by

$$\pi(\theta_n|t_n) = \begin{cases} 2(1-\theta_n) \text{ if } t_n = i \\ 2\theta_n \text{ if } t_n = g \end{cases}$$

and that $\alpha = 1$.

Consider first the case where the principal is restricted to trial systems and recall that, in a trial system, punishments are independent of the agents' reports. Let x^{Tr}

denote the optimal trial mechanism so that

$$x_n^{Tr}(m,\theta) \in \arg\max_{x_n \geq 0} \; E\left(u_n^p(t_n, x_n)\,|\theta\right)$$

for all $m \in M$, $\theta \in \Theta$, and n and so

$$x_n^{Tr}(m,\theta) \in \arg\max_{x_n \geq 0} \; \left\{-\pi(t_n = g|\theta)\,|1 - x_n| - \pi(t_n = i|\theta)\, x_n\right\}$$

If follows that

$$x_n^{Tr}(m,\theta) = \begin{cases} 0 \text{ if } \pi(t_n = g|\theta) \leq \frac{1}{2} \\ 1 \text{ otherwise} \end{cases}$$

If one assumes that $\delta = 0$, then

$$x_n^{Tr}(m,\theta) = \begin{cases} 0 \text{ if } \theta_n \leq \frac{1}{5} + \frac{3}{5}\theta_{-n} \\ 1 \text{ otherwise} \end{cases}$$

for all $m \in M$, $\theta \in \Theta$, and n.[10]

Given that the principal is risk neutral, he finds a bang-bang type of system optimal within trial systems. Notice also that the choice of ϕ does not influence the optimal trial system. In this system, the expected punishment of each agent n when innocent (which I denote by B_n^i) is equal to 0.22 while, if guilty (which I denote by B_n^g) it is equal to 0.78.

Now, suppose that the principal is not restricted to trial systems and can select any system, provided it is incentive compatible. If there is no upper bound on the punishments that the principal can inflict, she is actually able to achieve the first best solution. In particular, consider the following confession inducing mechanism $\widehat{x}$:

$$\widehat{x}_n(m,\theta) = \begin{cases} 1 \text{ if } m_n = c \\ 0 \text{ if } m_n = \overline{c} \text{ and } \theta_n \leq \varepsilon(\phi) \\ \phi \text{ if } m_n = \overline{c} \text{ and } \theta_n > \varepsilon(\phi) \end{cases}$$

where $\varepsilon(\phi) \in (0,1)$ is chosen so that, if agent n is guilty, he is indifferent between

[10] Without loss of generality, in the example and throughout the paper, I assume that ties are broken in favor of an acquittal.

choosing messages c and $\overline{c}$:

$$1 = \phi \int_{\varepsilon(\phi)}^{1} 2\theta_n d\theta_n \Leftrightarrow \varepsilon(\phi) = \frac{\sqrt{\phi(\phi-1)}}{\phi}$$

If the agent is innocent, he prefers to choose message $\overline{c}$, because the evidence is less likely to incriminate him than if he was guilty. So, in this system, guilty agents confess and always receive a punishment of 1, while innocent agents refuse to confess and their expected punishment is given by

$$\phi \int_{\varepsilon(\phi)}^{1} 2(1-\theta_n) d\theta_n \to 0 \text{ as } \phi \to \infty$$

This solution relies on the principal being able and willing to impose arbitrarily large punishments. If that is not the case, she will not be able to attain the first best. Imagine that $\phi = 1$ so that the principal's preferences can simply be stated as wanting to maximize the expected punishments of the guilty and minimize the expected punishments of the innocent.[11] Consider the following alternative to the optimal trial system. Before initiating an investigation and collecting the evidence, suppose the principal approaches agent 1 and makes him the following offer: if he confesses, he receives a certain punishment of 0.78; if he refuses, he faces the same lottery as in the trial system, i.e. he receives a punishment of 1 if $\theta_1 > \frac{1}{5} + \frac{3}{5}\theta_2$ and receives no punishment otherwise. After observing the choice made by agent 1, the principal investigates and collects evidence and then uses all information available at that time to select the punishment of agent 2.

Given that each agent is assumed to be risk neutral, it follows that agent 1 has just enough incentives to confess when guilty but prefers not to when innocent. So, under this new alternative, agent 1's expected punishment is unchanged. But now, consider what happens to agent 2. When the principal is making her decision about agent 2's punishment she will have observed θ, just like in the trial system, but she will also have observed the decision of agent 1. Therefore, in addition to the evidence, the principal will know whether agent 1 is innocent or guilty: he is guilty if and only

[11] In the text, I characterize the optimal system for any ϕ.

he chose to accept the offer. Being better informed allows the principal to select agent 2's punishment more accurately. In particular, in this alternative system, agent 2 is punished in 1 if and only if $\pi\left(t_2 = g|\theta, t_1\right) > \frac{1}{2}$. Therefore, this new system is clearly preferred to the trial system because, while agent 1's expected punishment is kept intact, the expected punishment of agent 2 decreases when he is innocent (B_2^i goes from 0.22 in the trial system to 0.16 in the new system) and increases when he is guilty (B_2^g goes from 0.78 to 0.84).

In the optimal incentive compatible system, rather than asking only one of the agents to confess in exchange for a constant punishment that leaves him indifferent only when guilty, the principal asks *both* agents. In this way, the principal is able to use the information each agent provides her by confessing (or refusing to confess) in determining the punishment of the other agent. In particular, when $\delta = 0$, the optimal system is a CIS $\left(x^{SB}, \sigma^{SB}\right)$ such that

$$x_n^{SB}\left(m, \theta\right) = \begin{cases} 0.84 \text{ if } m_n = c \\ 0 \text{ if } m_n = \overline{c} \text{ and } \pi\left(t_n = g|\theta, m_{-n}\right) \leq \frac{1}{2} \\ 1 \text{ if } m_n = \overline{c} \text{ and } \pi\left(t_n = g|\theta, m_{-n}\right) > \frac{1}{2} \end{cases}$$

and

$$\sigma_n^{SB}\left(t_n, m_n\right) = \begin{cases} 1 \text{ if } \left(t_n, m_n\right) = \left(g, c\right) \text{ or } \left(t_n, m_n\right) = \left(i, \overline{c}\right) \\ 0 \text{ otherwise} \end{cases}$$

for all $m \in M$, $\theta \in \Theta$ and n.

The problem with system $\left(x^{SB}, \sigma^{SB}\right)$ is that agents report truthfully. If an agent is guilty he confesses, and if he is innocent he does not. This means that, simply by observing the choice of the agent, the principal is able to infer his type, making it possible that the principal knows that the agent is innocent but is required by the mechanism to punish him. As a result, system $\left(x^{SB}, \sigma^{SB}\right)$ is not renegotiation proof.

The challenge of finding the optimal renegotiation proof mechanism is that the revelation principle need not hold in a context with multiple agents (Bester and Strausz (2000)). Nevertheless, in section 7, I show that one can build a renegotiation proof CIS that is approximately optimal, provided the probability that there is more than one guilty agent is small (if δ is close to $\frac{1}{10}$ in this example). This CIS is different than $\left(x^{SB}, \sigma^{SB}\right)$ in two ways. First, each guilty agent n does not confess with probability 1 but only with a probability $\tau_n \in (0, 1)$. And second, if an agent refuses to confess, his punishment is sequentially optimal, i.e., it is the punishment that the principal finds optimal given *all* the information he will have after observing everyone's report and the evidence. In other words, if an agent refuses to confess, his punishment is equal to 1 if the posterior probability that his type is "guilty", conditional on message m and evidence θ, is larger than $\frac{1}{2}$ and is 0 otherwise.

5 Trial System

Recall that, in any trial system, the reports of the agents are irrelevant as the punishments are only a function of the evidence. Let x^{Tr} denote the optimal trial mechanism such that

$$x_n^{Tr}(m,\theta) \in \arg\max_{x_n \geq 0} \; E\left(u_n^p(t_n, x_n)\,|\theta\right)$$

for all $m \in M$, $\theta \in \Theta$ and n. By following the steps of the previous example, one gets that

$$x_n^{Tr}(m,\theta) = \begin{cases} 0 \text{ if } \pi(t_n = g|\theta) \leq \alpha\pi(t_n = i|\theta) \\ \quad\quad\quad 1 \text{ otherwise} \end{cases}$$

Proposition 6 *For all $m \in M$, $\theta \in \Theta$ and n,*

$$x_n^{Tr}(m,\theta) = \begin{cases} 0 \textit{ if } \theta_n \leq \theta_n^{Tr}(\theta_{-n}) \\ \quad 1 \textit{ otherwise} \end{cases}$$

where threshold $\theta_n^{Tr}(\theta_{-n}) \in (0,1)$ is completely characterized in the proof.

Proof. The optimal trial system is a simple threshold rule because the principal is risk neutral and because of the monotone likelihood ratio assumption on the distribution of the evidence. Threshold θ_n^{Tr} is characterized in Appendix B1. ■

As one might have expected, in a trial system, parameter α determines the standard of proof: if α is large, the threshold will be larger and so the evidence must be more conclusive for the agent to be punished. The threshold θ^{Tr} depends on θ_{-n} because the agents' innocence might be correlated: agent n's type is correlated with agent $\widehat{n}$'s type which, in turn, is correlated with $\theta_{\widehat{n}}$.

6 A simple case: full commitment power

In this section, I find the second best system: the principal's most preferred incentive compatible system. Let the principal's expected utility given a system (x, σ) be denoted $V(x, \sigma)$ and notice that $V(x, \sigma) = \sum_{n=1}^{N} V_n(x_n, \sigma)$, where

$$V_n(x_n, \sigma) = \int_{m \in M} \int_{\theta \in \Theta} \sum_{t \in T} \pi^{\sigma}(m|t)\, \pi(\theta|t)\, u_n^p(t_n, x_n(m, \theta))$$

where $\pi^{\sigma}(m|t)$ represents the density of message m, given type vector t and strategy profile σ.

By the revelation principle (see, for example, Myerson (1979)), it follows that it is enough to focus on revelation mechanisms that induce truthful reporting in order to maximize the principal's expected utility. Therefore, without loss of generality, I only consider systems (x, σ) where guilty agents confess and send message c, while innocents agents refuse to confess and send message $\overline{c}$. In this context, there are only two incentive constraints for each agent: one for the guilty (if the agent is guilty he must prefer to report c than to report $\overline{c}$) and one for the innocent (if the agent is innocent he must prefer to report $\overline{c}$ than to report c).

The general problem of finding the optimal incentive compatible mechanism can be transformed into N independent problems where, in each n^{th} problem, one chooses $x_n : M \times \theta \rightarrow \mathbb{R}_+$ subject to the two incentive constraints with respect to agent n. I now proceed to solve each n^{th} problem.

Lemma 7 *The innocent's incentive constraint does not bind.*

Proof. Consider the less constrained problem of maximizing V_n subject only to the guilty's incentive constraint. I claim that the solution x_n' of that problem must still satisfy the innocent's incentive constraint, which shows the statement. Suppose not. This would mean that if the agent is innocent, he would be strictly better off reporting c and pretending to be guilty, than to report $\overline{c}$. Given that the interests of the principal and of the innocent are perfectly aligned, it would actually be in the interest of the principal that the innocent agent did report c. Therefore, if one considers x''_n where

$$x''_n\left(c, m_{-n}, \theta\right) = x'_n\left(c, m_{-n}, \theta\right) \text{ and } x''_n\left(\bar{c}, m_{-n}, \theta\right) = x'_n\left(c, m_{-n}, \theta\right)$$

for all $m_{-n} \in M_{-n}$ and $\theta \in \Theta$, it follows that x''_n is strictly preferred to x'_n by the principal and is also incentive compatible, which is a contradiction to x'_n being the solution of the less constrained problem. ■

Lemma 8 *Punishments after confessions (message c) never exceed* 1*.*

Proof. Given the previous lemma, the problem of the principal is less constrained. It is simply to maximize V_n subject to the guilty's incentive constraint. Thus, increasing punishments on guilty agents (those who confess) in more than 1 is not optimal as it decreases the principal's expected utility (remember that the principal prefers to impose a punishment of 1 on guilty agents) and makes it harder to satisfy the guilty's incentive constraint. ■

Let $B_n^g\left(x_n\right)$ and $B_n^i\left(x_n\right)$ denote the expected punishment of agent n if guilty and innocent respectively under mechanism x:

$$B_n^g\left(x_n\right) = \int\limits_{m_{-n}\in M_{-n}} \int\limits_{\theta\in\Theta} \pi^\sigma\left(m_{-n}, \theta | t_n = g\right) x_n\left(c, m_{-n}, \theta\right) d\theta dm_{-n}$$

and

$$B_n^i\left(x_n\right) = \int\limits_{m_{-n}\in M_{-n}} \int\limits_{\theta\in\Theta} \pi^\sigma\left(m_{-n}, \theta | t_n = i\right) x_n\left(\bar{c}, m_{-n}, \theta\right) d\theta dm_{-n}$$

where $\pi^\sigma\left(m_{-n}, \theta | t_n\right)$ represents the joint density of $\left(m_{-n}, \theta\right)$ given the agent's type t_n and strategy profile σ. Lemmas 7 and 8, combined with the fact that it is not optimal to punish agents in more than ϕ in any circumstance, imply that the problem of the principal is that of maximizing

$$-\pi\left(t_n = g\right)\left(1 - B_n^g\left(x_n\right)\right) - \alpha\pi\left(t_n = i\right) B_n^i\left(x_n\right) \tag{1}$$

subject to the guilty's incentive constraint, which can be written as

$$B_n^g(x_n) \leq \int\limits_{m_{-n}\in M_{-n}} \int\limits_{\theta\in\Theta} \pi^\sigma(m_{-n},\theta|t_n = g)\, x_n(\bar{c}, m_{-n},\theta)\, d\theta dm_{-n} \tag{2}$$

and upper bound constraints

$$x_n(\bar{c}, m_{-n},\theta) \leq \phi \text{ for all } m_{-n} \in M_{-n} \text{ and } \theta \in \Theta \tag{3}$$

and

$$x_n(c, m_{-n},\theta) \leq 1 \text{ for all } m_{-n} \in M_{-n} \text{ and } \theta \in \Theta \tag{4}$$

Notice that the choice of punishments after a confession only affects (1) or (2) through its impact on B_n^g. In other words, only the expected punishment for guilty agents after confessions affects (1) and (2), not the whole distribution of punishments that follow a confession. Furthermore, the upper bound constraint (4) is more easily satisfied if punishments after confessions are constant.[12]

As a result, it follows that CIS's are optimal (even though not necessarily uniquely optimal) because there is an optimal system which satisfies the two requirements: each agent sends at most two messages and punishments after confessions are constant.

Proposition 9 *There is a CIS* (x^{SB}, σ^{SB}) *which is an optimal incentive compatible system.*

Strategy profile σ^{SB} has already been described:

$$\sigma_n^{SB}(i, m_n) = \begin{cases} 1 \text{ if } m_n = \bar{c} \\ 0 \text{ otherwise} \end{cases} \text{ and } \sigma_n^{SB}(g, m_n) = \begin{cases} 1 \text{ if } m_n = c \\ 0 \text{ otherwise} \end{cases}$$

In the remainder of this section, I characterize x^{SB} and compare it with the optimal trial mechanism.

[12] This statement follows because

$$\left(\max_{(m_{-n},\theta)\in M_{-n}\times\Theta} x_n(c, m_{-n},\theta)\right) \leq 1 \Rightarrow E_{(m_{-n},\theta)}(x_n(c, m_{-n},\theta)) \leq 1$$

Notice that, in any solution of the principal's problem, it must be that if the agent is guilty he is indifferent between confessing and not confessing, i.e. condition (2) holds with equality. The reason is that, if not, the principal would be able to improve the solution by reducing the punishments that follow a refusal to confess, which would reduce the expected punishment of the agent when innocent. Therefore, one can plug condition (2) into (1) and into (4) to obtain a problem which depends only on the punishments that follow message $\overline{c}$. In this reduced problem, the principal must choose $x_n(\overline{c}, m_{-n}, \theta) \in [0, \phi]$ for each $m_{-n} \in M_{-n}$ and $\theta \in \Theta$ in order to maximize

$$\int_{m_{-n}\in M_{-n}} \int_{\theta\in\Theta} \left(\pi^{\sigma^{SB}}(t_n = g|m_{-n}, \theta) - \alpha\pi^{\sigma^{SB}}(t_n = i|m_{-n}, \theta)\right)\pi^{\sigma^{SB}}(m_{-n}, \theta)\, x_n(\overline{c}, m_{-n}, \theta)\, d\theta dm_{-n} \tag{5}$$

subject to

$$\int_{m_{-n}\in M_{-n}} \int_{\theta\in\Theta} \frac{\pi^{\sigma^{SB}}(t_n = g|m_{-n}, \theta)\,\pi^{\sigma^{SB}}(m_{-n}, \theta)}{\pi(t_n = g)} x_n(\overline{c}, m_{-n}, \theta)\, d\theta dm_{-n} \leq 1 \tag{6}$$

Notice that the left hand side of (6) represents the expected punishment of the agent if guilty, given that he is indifferent between confessing and refusing to confess.

Proposition 10 *For all $m_{-n} \in M_{-n}$, $\theta \in \Theta$ and n,*

$$x_n^{SB}(\overline{c}, m_{-n}, \theta) = \begin{cases} 0 \text{ } if\ \theta_n \leq \theta_n^{SB}(m_{-n}, \theta_{-n}) \\ \phi \text{ } otherwise \end{cases}$$

and

$$x_n^{SB}(c, m_{-n}, \theta) = \int_{m_{-n}\in M_{-n}} \int_{\theta\in\Theta} \frac{\pi^{\sigma^{SB}}(t_n = g|m_{-n}, \theta)\,\pi^{\sigma^{SB}}(m_{-n}, \theta)}{\pi(t_n = g)} x_n^{SB}(\overline{c}, m_{-n}, \theta)\, d\theta dm_{-n}$$

where threshold $\theta_n^{SB}(m_{-n}, \theta_{-n}) \in (0, 1)$ is completely characterized in the proof.

Proof. The solution of the reduced problem is a threshold rule because the problem is linear on the choice variable and because of the monotone likelihood ratio assumption on the evidence generating process. The threshold θ_n^{SB} is calculated in Appendix B2. ■

Larger values of ϕ make the principal strictly better off as the problem becomes less constrained and the constraint (that punishments must be lower than ϕ) binds. Therefore, the optimal expected utility of the principal is the smallest when $\phi = 1$. In that case, constraint (6) does not bind, because, if all punishments are bounded by 1, their weighted average must also be bounded by 1. Therefore, it follows directly from (5) that the optimal punishment after an agent refuses to confess (sends message $\overline{c}$) is 1 if

$$\pi^{\sigma^{SB}} (t_n = g|m_{-n}, \theta) > \alpha\pi^{\sigma^{SB}} (t_n = i|m_{-n}, \theta) \tag{7}$$

and 0 otherwise. So, when deciding whether or not to punish an agent who has refused to confess, the principal is supposed to use all information available to him that comes from a different source than the agent himself: the evidence and the other agents' reports. Given that all agents report truthfully, their reports are perfectly informative about their type, which makes them particularly valuable for the principal, assuming that the agents' innocence is correlated. By contrast, in the optimal trial system, agents are punished according to the evidence alone, making it more likely that the principal ends up making mistakes: convicting those who are innocent or acquitting those who are guilty.

Proposition 11 *The trial system is an optimal incentive compatible mechanism if and only if $\phi = 1$ and the agents' types are independent.*

There is one interesting property of this system which I believe is worth emphasizing. Condition (7) represents the optimal decision regarding agent n that the principal is able to make, given the evidence and the information provided by all other agents. The principal obtains this information from the agents through the promise that a confession does not increase the agent's expected punishment. In other words, a guilty agent chooses to confess because he knows that this piece of information he provides (the fact that he is guilty) will not be used against him when determining what punishment he is supposed to receive. This seems to be in contrast with the American criminal law practice of the *Miranda warnings*, or, at least, with the part where an agent is told that everything he says might be used against him in court. According to this analysis, the principal should be doing the exact opposite: she should be providing a guarantee that she will **not** use this information against the agent, which, in the current legislation, is actually achieved by purposefully not reading the Miranda warnings.

If ϕ is "close" to 1, it will still be the case that constraint (6) does not bind. In this case, if ϕ was to increase, the threshold θ_n^{SB} would not change as it would be determined by condition (7). Therefore, increases in ϕ would only lead to larger punishments in the event that an agent who refuses to confess is convicted, and, as a result, the expected punishment of the innocent and the guilty would also increase with ϕ. This would go on until the expected punishment of the agent when guilty hits 1 (until constraint (6) becomes binding), which happens when $\phi = \overline{\phi}_n$, calculated in Appendix B2. If $\phi > \overline{\phi}_n$, constraint (6) always binds and, so, the expected utility of the guilty is fixed at the level 1. As a result, if there is an increase of ϕ when $\phi > \overline{\phi}_n$, because the punishment in the event of a conviction after a refusal to confess is equal to ϕ and is, thus, increasing, the threshold θ_n^{SB} must also be increasing for the expected punishment of the guilty to remain equal to 1. In other words, if an agent refuses to confess he is less likely to be punished, but if he is punished, his punishment will be larger. This ends up being beneficial for the innocent as he is less likely to generate incriminating evidence than the guilty. In the limit, as shown in the example of section 4, it is possible to achieve the first best if there is no upper bound on punishments:

Proposition 12 *For all n,* $\lim_{\phi\to\infty} \left(B_n^i\left(x^{SB}\right), B_n^g\left(x^{SB}\right)\right) = (0,1)$.

Proof. See Appendix B3. ■

Figure 1 presents a graphical representation of how the expected punishment of the agent when innocent and guilty ($B_n^i\left(x^{SB}\right)$ and $B_n^g\left(x^{SB}\right)$) evolves with ϕ.

There is one last comment I would like to make in this section, with respect to whether or not system $\left(x^{SB}, \sigma^{SB}\right)$ is uniquely optimal. If ϕ is small (if $\phi < \overline{\phi}_n$) and, as a result, constraint (6) holds strictly, there are other systems that are also optimal. All optimal mechanisms have in common the punishments that innocent agents receive but not necessarily the punishments that guilty agents receive, because only the expected punishment of guilty agents matters. However, if $\phi > \overline{\phi}_n$, the expected punishment of a guilty agent must be equal to 1 for the system to be

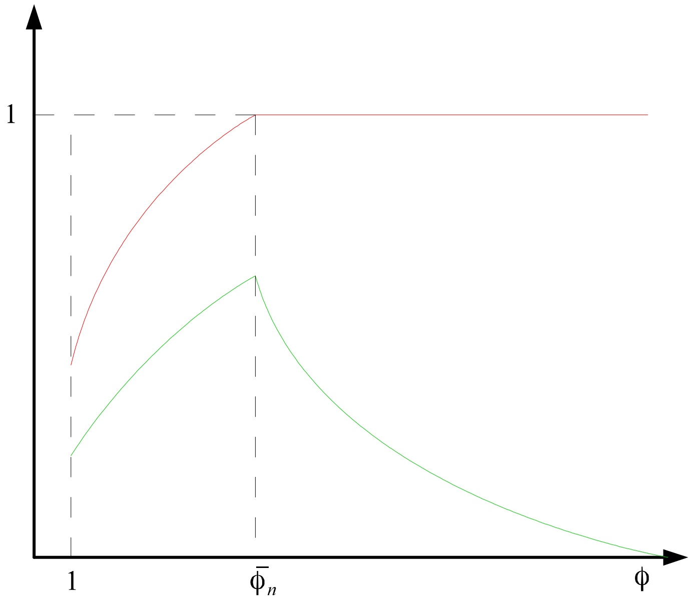

Figure 1: Evolution of the agent's expected punishment as a function of ϕ. The red and green curves represent the expected punishment when the agent is guilty and innocent respectively.

optimal. Seeing as it is never optimal to impose punishments on guilty agents that exceed 1 (by Lemma 8), it follows that the only way that the expected punishment of guilty agents is 1 is if all punishments after a confession are equal to 1.[13] Therefore, if $\phi > \overline{\phi}_n$, $\left(x^{SB}, \sigma^{SB}\right)$ is uniquely optimal. Furthermore, even if $\phi < \overline{\phi}_n$, as long as agents are even slightly risk averse, there is a CIS, constructed in essentially the same way as $\left(x^{SB}, \sigma^{SB}\right)$, which is uniquely optimal. I make this point in detail in Appendix A2, but the idea is that if an agent is risk averse, he is more willing to confess when guilty if his punishment is constant than if it is not.

The problem of excessive commitment power

The CIS $\left(x^{SB}, \sigma^{SB}\right)$ is based on the assumption that the principal is able to commit not to alter any punishment from the mechanism even after observing the agents' reports and evidence. That assumption allows the principal i) to punish the guilty agents in less than 1 once they confess, and ii) to punish innocent agents even with the knowledge that they are indeed innocent.

[13] Naturally, there could be deviations from this description on events with measure 0, which is something that I am overlooking.

As for i), only guilty agents confess the crime in the CIS $\left(x^{SB}, \sigma^{SB}\right)$. Hence, upon hearing a confession, the principal would prefer to renege his promise and choose a punishment of 1. Of course, knowing this, a guilty agent would not confess. Is it reasonable to believe that the principal can commit not to punish more harshly the confessing agents? Currently, there are several examples where the law protects agents that confess a crime in exchange for a softer punishment.[14] It seems that, by regulating these confession inducing contracts through law, the principal is able to credibly commit to leniency towards confessing agents, and breaches to these contracts by the principal are deemed unacceptable.

Implication ii) seems more unreasonable. In the mechanism described, only innocent agents choose not to confess to having committed the crime. However, the principal will still punish some of them in some circumstances to deter guilty agents from misreporting. Hence, the principal must be able to commit to punish agents she knows are innocent. This seems harder to accept as, not only does the principal prefer to go back on her promise of punishment, but also the agent prefers she does, i.e. both parties prefer to renegotiate the confession inducing contract, once an agent has not confessed. Knowing this, guilty agents would not confess, in the hope that the promise of punishment would be reneged by the principal. Even if the principal employed such a system through law, it still seems unlikely to me that society would be willing to accept that agents who are known to be innocent are punished.

7 Limited Commitment Power

In this section, I restrict the set of systems that the principal can choose from. I divide the section into two parts. In the first part, I consider renegotiation proof systems, where the principal is able to commit to show leniency towards agents she believes are guilty but is not able to commit to punishing them if she believes they are innocent. In the second part, I consider sequentially optimal systems, where the principal is not able to commit.

[14] See Kaplow and Shavell (1994) for a description of some of the regulations in environmental law like, for example, the Compreehensive Environmental Response, Compensation and Liability Act; and, with respect to plea bargaining, Rule 11 of the Federal Rules of Criminal Procedure regulates the process under which the prosecutor and the defendants reach a plea deal.

7.1 Renegotiation Proof Mechanisms

Recall from definition 3 that a system (x, σ) is renegotiation proof if and only if, for all $m \in M$, $\theta \in \Theta$ and n,

$$x_n(m, \theta) \leq \min \left\{ \arg \max_{x_n \geq 0} \; E^{\sigma} \left(u_n^p (t_n, x_n) \, | m, \theta \right) \right\}$$

It is convenient to define

$$\gamma_n^{\sigma}(m, \theta) = \min \left\{ \arg \max_{x_n \geq 0} \; E^{\sigma} \left(u_n^p (t_n, x_n) \, | m, \theta \right) \right\}$$

Notice that, because the principal is risk neutral,

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

If $x_n(m, \theta) > \gamma_n^{\sigma}(m, \theta)$ - if the punishment imposed on agent n is larger than the punishment the principal would rather impose - both the principal and agent n have an incentive to reduce the punishment at least to $\gamma_n^{\sigma}(m, \theta)$. However, if $x_n(m, \theta) \leq \gamma_n^{\sigma}(m, \theta)$, the principal is no longer willing to accept a smaller punishment.

The first thing to notice is that, in any renegotiation proof system, there are no punishments that exceed 1, simply because, if there was such a punishment, both the principal and the agent would be better off by reducing it to 1. Therefore, the optimal renegotiation proof system is independent of the parameter ϕ.

Notice also that the CIS $\left(x^{SB}, \sigma^{SB}\right)$, which was optimal among incentive compatible systems, is not renegotiation proof. The strategy profile σ^{SB} involves agents reporting truthfully - all guilty agents confess while all innocent agents do not. This means that, upon observing that an agent has not confessed, the principal believes he is innocent, and so will not be willing to punish him.

Recall that, if the principal has limited commitment power, the revelation principle need not hold. Therefore, in principle, it is possible that the optimal renegotiation proof system is something quite intricate where agents send multiple messages. The next lemma allows me to restrict attention to a particular message set:

Lemma 13 *Without loss of generality, it is possible to set* $M_n = \mathbb{R}_+ \cup \{c\}$ *for all* n.

Proof. See Appendix B4. ■

The meaning of a message is given by the belief that the principal forms when she receives it. In Lemma 13, I show that, if two distinct messages sent by some agent lead to the same belief by the principal, then one of the messages is redundant and can be eliminated, i.e. there is an equivalent system where only one of the two messages is sent. It follows that, without loss of generality, there is an optimal renegotiation proof system where each message leads to a different posterior belief. In particular, let

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

for all n. If there are two message m_n' and m_n'' such that $r_n(m_n') = r_n(m_n'')$ they are essentially the same message and one of them is redundant. So, the set M_n only has to be large enough to accommodate all elements of the range of $r_n(m_n)$. Message c is interpreted as a confession and is only sent by guilty agents in any given incentive compatible system (x, σ), and so $r_n(c) = \infty$.

I divide the problem of finding the optimal incentive compatible renegotiation proof system $\left(x^{RP}, \sigma^{RP}\right)$ into two parts. In the first part, in Lemma 14, I fix any strategy profile σ and find the best possible mechanism x^σ that makes system (x^σ, σ) incentive compatible and renegotiation proof. So, for all σ, $V(x^\sigma, \sigma) \geq V(x, \sigma)$ for any mechanism x such that (x, σ) is incentive compatible and renegotiation proof. Given the first step, the problem simply becomes one of choosing a σ that maximizes $V(x^\sigma, \sigma)$.

For a fixed σ, let m_n^σ denote the message after which the principal believes agent n is more likely to be innocent. More rigorously, let m_n^σ be such that, for all n,

$$r_n(m_n^\sigma) = \inf\{r_n(m_n) \text{ for all } m_n \in \mathbb{R}_+ : \sigma_n(i, m_n) > 0\}$$

Lemma 14 *For all* n, *and for all* $m_{-n} \in M_{-n}$ *and* $\theta \in \Theta$,

$$\begin{cases} x_n^\sigma(m_n, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta) \text{ for all } m_n \in \mathbb{R}_+ \\ x_n^\sigma(c, m_{-n}, \theta) = \varphi_n \end{cases}$$

where

$$\varphi_n = \int\limits_{(m_{-n},\theta)\in M_{-n}\times\Theta} \pi^{\sigma}\left(m_{-n},\theta|t_n = g\right)\gamma_n^{\sigma}\left(m_n^{\sigma}, m_{-n},\theta\right)d\left(m_{-n},\theta\right)$$

Mechanism x^{σ} has three main features. First, the lottery of punishments that follows any non-confessing message ($m_n \neq c$) is always the same. This means that, conditional on not confessing, it is the same for agent n what message he ends up sending. However, it is not the same for other agents, as the choice of agent n influences other agents' punishments.

Second, the punishment that an agent receives if he refuses to confess is the sequentially optimal punishment if the message chosen had been m_n^{σ}. The idea is that, whenever an agent refuses to confess, the principal pretends the agent sent message m_n^{σ}, the message after which the agent "appears more innocent", and chooses what would have been her preferred punishment, given the evidence and the other agents' reports.

Third, if the agent chooses to confess, he receives a constant punishment which leaves him indifferent to refusing to confess only if he is guilty.

Proof of Lemma 14. Take some incentive compatible renegotiation proof system (x,σ) and fix some agent n. First, if σ is such that m_n^{σ} is only sent by type $t_n = i$, the principal will be convinced that agent n is innocent upon observing m_n^{σ} and, as a result, will be obliged to choose a punishment of 0. Thus, in order to satisfy incentive compatibility, it must be that $x_n\left(m,\theta\right) = 0$ for any (m,θ).

Consider, instead, the case where m_n^{σ} is sent by both types of agent n. By incentive compatibility, it must be that each type t_n is indifferent between sending any message m_n for which $\sigma_n\left(t_n, m_n\right) > 0$ and m_n^{σ}. But then, consider the following mechanism $\widehat{x}_n$ where $\widehat{x}_n\left(m,\theta\right) = x_n\left(m_n^{\sigma}, m_{-n},\theta\right)$ for all (m,θ). In words, make all messages return the same lottery of punishments as message m_n^{σ}. The new system $(\widehat{x},\sigma)$ is trivially incentive compatible and satisfies the renegotiation proof condition because that constraint is the most tight after message m_n^{σ}, which is when the posterior probability that agent n is innocent is the largest. Furthermore, the principal is indifferent between systems (x,σ) and $(\widehat{x},\sigma)$ because she only cares about each type's expected punishment, which is the same in both systems.

The only issue for the principal is to determine what $x_n\left(m_n^{\sigma}, m_{-n},\theta\right)$ should be for each (m_{-n}, θ). She would like to choose $z_n(m_{-n}, \theta)$, where

$$z_n(m_{-n}, \theta) \in \arg \max_{z_n \in [0,1]} \left(\pi^\sigma(t_n = g | m_{-n}, \theta) - \alpha \pi^\sigma(t_n = i | m_{-n}, \theta)\right) z_n$$

but the problem is that $z_n(m_{-n}, \theta) \geq \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$, because, by definition, the posterior probability that agent n is innocent is the largest after m_n^σ. As a result, it is optimal for the principal to set

$$x_n^\sigma(m_n, m_{-n}, \theta) = \gamma_n^\sigma(m_n^\sigma, m_{-n}, \theta)$$

Finally, the principal is able to make $x_n^\sigma(c, m_{-n}, \theta) = \varphi_n$ for all (m_{-n}, θ) rather than equal to $x_n^\sigma(m_n^\sigma, m_{-n}, \theta)$ because, by definition of φ_n, the guilty type (who is the only one who sends c) is made indifferent between reporting c and not reporting c, while the innocent type prefers not to report c, given that the lottery of punishments that follows m_n^σ is more appealing if the agent is innocent than if he is guilty (this last part of the argument is shown formally in Appendix B5). ■

If a system (x^σ, σ) is a CIS, it must be that σ is such that only two messages are sent with positive probability by each agent n: c and m_n^σ, which one can label as $\bar{c}$. If agent n is innocent, he sends message $\bar{c}$, while, if he is guilty, he randomizes between c and $\bar{c}$, instead of always confessing as was the case when the principal had commitment power. A second difference between a CIS (x^σ, σ) and the optimal CIS of the previous section is that the punishments that follow a refusal to confess are sequentially optimal, while, in the previous section, they were chosen regardless of the perceived guilt of the agent. In particular, even though the principal was able to identify all innocent agents, as they were the only ones who did not confess, she was still supposed to punish them. She was only able do this because she was able to commit, which would presumably require having a set of laws and regulations for judges, lawyers and jurors to follow, which would not necessarily be designed to assess the agents' guilt. But under this new CIS this is no longer necessary. Its implementation requires only the guarantee that the rights of confessing agents are protected.

If the optimal renegotiation proof system is a CIS $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$, then

$$\sigma^{CIS} \in \arg\max_{\sigma \in \Phi} \; V\left(x^{\sigma}, \sigma\right)$$

In general, in environments where the principal has limited commitment power and there are multiple agents, it is possible that the number of necessary messages for optimality exceeds the number of types (in this case, two), which suggests that CISs are not always optimal.[15] Nevertheless, in Proposition 15, I show that this is the case in this framework, provided that the probability that more than one agent is guilty is sufficiently small.

Let

$$\xi_n = \pi\left(t_n = g, t_{-n} = (i, ..., i)\right) \text{ for all } n$$

represent the prior probability that only agent n is guilty and let

$$\widehat{T} = \{t \in T : \nexists n', n'' \text{ such that } t_{n'} = t_{n''} = g \text{ and } n' \neq n''\}$$

denote the set of type vectors for which there is, at most, one guilty agent.

Proposition 15 *For all $\gamma > 0$ and $(\xi_1, ..., \xi_N) \in (0,1)^N$ such that $\sum_{n=1}^{N} \xi_n < 1$, there is $\delta > 0$ and a CIS $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$ such that, if $\sum_{t \notin \widehat{T}} \pi(t) < \delta$, then*

$$\max_{\sigma \in \Phi} \; \{V\left(x^{\sigma}, \sigma\right)\} - V\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right) \leq \gamma$$

Proof. See Appendix B6. ■

To understand the argument it is simpler if one assumes that there are only two agents: 1 and 2. Consider some σ such that agent 1 sends 3 messages with positive probability: messages c, m_1' and $\overline{c}$ such that

$$r_1\left(\overline{c}\right) < r_1\left(m_1'\right) < \infty$$

[15] See Bester and Strausz (2000).

Now consider the following change to σ: shift weight $v > 0$ from $\sigma_1(g, m_1')$ to $\sigma_1(g, c)$ enough so that

$$\frac{\sigma_1(g, m_1') - v}{\sigma_1(i, m_1')} = r_1(\bar{c})$$

as shown in Figure 2. The idea is to make message m_1' equivalent to message $\bar{c}$ and then merge the two.

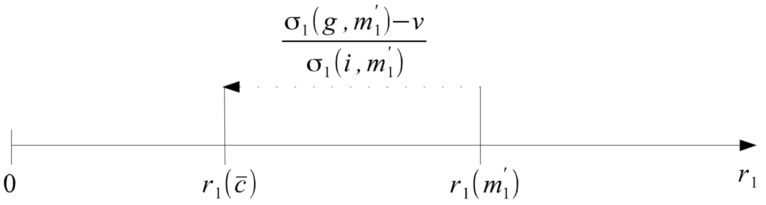

Figure 2: Shift from $r_1(m_1')$ to $r_1(\bar{c})$

Agent 1's expected utility is unchanged for either type and, as a result, so is the expected utility that the principal obtains from agent 1. However, this change effectively gives the principal more information. It is as if, after receiving message m_1', the principal is able to receive a second signal $y \in \{0, 1\}$. If agent 1 is innocent, he sends only $y = 0$ but if he is guilty he may also send $y = 1$ with some probability so that the posterior belief after signal $y = 0$ is exactly the same as if agent 1 sent message m_1^σ to begin with. Hence, one would think that such added information would actually help the principal when choosing the punishments to inflict on agent 2. However, more information is not necessarily beneficial for the principal because of her limited commitment power.

Recall that, by Lemma 14, in the optimal system, agent 2 cannot be discriminated against as a result of the message he chooses to send. Given this, the principal would actually prefer to ignore the message agent 2 sends and just use m_1 and the evidence θ to select agent 2's punishment. However, in order to get agent 2 to confess with some positive probability, a message m_2^σ must be generated, which is more likely to be sent if agent 2 is innocent. Because the principal has limited commitment power, the existence of this message constrains her in that if forces the principal to choose punishments that are smaller than what she would like. In particular, the fact that the punishments that follow m_2^σ determine the utility of agent 2 generates a trade-off for the principal: in order to get agent 2 to confess, she must choose his punishments as if she had a bias towards his innocence. Because of this bias, more information might not always be good as I illustrate below.

Fix any θ for which $x_2^g(m_2^g, m_1', \theta) = 0$. In this case, receiving the additional signal y after message m_1' does not make the principal worse off because she can always keep the punishment at 0 after observing the additional signal - the commitment problem she faces is that she cannot punish those who she believes are innocent and not the other way around.

Suppose instead that $x_2^g(m_2^g, m_1', \theta) = 1$ and that $x_2^g(m_2^g, m_1 = c, \theta) = 0$ and consider the problem of the principal after observing signal $y = 1$, which reveals that agent 1 is guilty. In this case, the principal is not able to choose a punishment of 1 anymore because

$$x_2^g(m_2^g, m_1 = c, \theta) = \gamma_2^g(m_2^g, m_1 = c, \theta) = 0$$

So, the principal is not able to do as before having the extra signal y. And, because m_2^g is the message after which agent 2 appears more likely to be innocent, it is possible that the principal would prefer to inflict a punishment of 1 in such an event but is prevented from doing so because of her lack of commitment power.

Now, think about what changes if the principal knows (or is very certain) that there is at most one guilty agent. Consider the same scenario: $x_2^g(m_2^g, m_1', \theta) = 1$ but $x_2^g(m_2^g, m_1 = c, \theta) = 0$. After observing $y = 1$, she gets to know that agent 2 is innocent and, as a result, would like to acquit him. So, the fact that she acts as if she has a bias towards the innocence of agent 2 does not matter - she would like to acquit agent 2 after any message m_2 and not only after message m_2^g. As a result, under the assumption that there is at most one guilty agent, the knowledge that agent 1 is guilty is always beneficial for the principal, which means that the change displayed in figure 2 would actually be good for her.

While CISs might not be optimal for a general prior, it is the case that, for any prior, there is always a CIS that performs better than the optimal trial system.

Proposition 16 *The trial system is **not** an optimal incentive compatible renegotiation proof CIS, unless agents' have independent types.*

Proof. See Appendix B7. ■

Take a trial system and consider a marginal deviation from player 1 - suppose he confesses with a very small probability, if he is guilty. The direct impact of this change is that, when other agents are taken to trial and agent 1 is guilty, the principal is more likely to be aware of it (because it is more likely that agent 1 confesses) and, so, she is able to choose more appropriate punishments to impose on them (in this case, more information is always better for the principal as the other agents do not confess). There is also an indirect impact in that, in the event that agent 1 does not confess, the principal believes he is more likely to be innocent than before. As a result, the punishments that she chooses for agent 1 will be slightly smaller than what she would like. However, proposition 16 shows that, if the probability of confession is sufficiently small, it is possible to guarantee that the direct impact dominates.

7.2 Sequentially Optimal Mechanisms

CIS's are based on the assumption that the principal is able to partially forgive a guilty agent who confesses, precisely in order for him to confess. However, knowing that only guilty agents confess, it is not ex-post optimal for the principal to show leniency towards them. Hence, if the principal does not have commitment power, she will be unable to implement such confession inducing mechanisms. In this section, I analyze what mechanism should the principal implement if she has no commitment power.

Recall that a system (x, σ) is sequentially optimal if and only if, for all $m \in M$, $\theta \in \Theta$ and n,

$$x_n(m, \theta) \in \arg\max_{x_n \geq 0} \; E^{\sigma}\left(u_n^p(t_n, x_n) \,|m, \theta\right)$$

By eliminating the commitment power of the principal, one also eliminates her ability to collect any information from the agents. Imagine that agent n is sending two distinct messages a and b. For these messages to convey any information to the principal, it must be that they are sent with different probabilities by the innocent and the guilty types. Suppose a is more likely to have been sent by the innocent type than b. Knowing this, the principal has no choice but to be more lenient towards agents that have sent message a. But then, no agent would ever send message b. It follows that, if the principal is unable to recover any information from the agents, all we are left with is the trial system.

Proposition 17 *The optimal trial system is an optimal incentive compatible sequentially optimal system.*

7.3 How much commitment power does the principal have?

This paper characterizes the principal's preferred mechanism under three different assumptions with respect to her commitment power: full commitment power, no commitment power and an in-between assumption, where the principal is only unable to commit not to renegotiate. But which of the three assumptions is more reasonable?

One way to approach the problem of analyzing what an optimal criminal justice system should look like is to imagine that society is ruled by a benevolent dictator who is granted the exclusive responsibility of administering criminal justice and make her the principal in the model. But if the benevolent dictator is the principal, she should be unable to commit. To have the ability to commit is to be able to write contracts that some exogenous entity will enforce. Parties follow the contract for, if not, that exogenous source of authority punishes them heavily. But if the benevolent dictator is one of the parties, then, by definition, there is no other source of authority that rules over her. So she is unable to write any contracts in the sense that there is no entity that enforces them. Hence, it would follow that the principal should not be able to commit and the trial system would be the only alternative.

However, looking at contemporaneous societies one can see that there are several examples where leniency is shown towards agents who confess to having committed a crime. The method modern societies seem to follow, in order to commit to such leniency, is to use law. For example, plea bargain deals are protected under Rule 11 of the Federal Rules of Criminal Procedure, which ensures that the prosecutor cannot go back on his word once he has obtained the confession from the agent. But if societies can use law as a commitment device, one could argue that the relevant analysis should be the one that assumes full commitment power by the principal. The problem with this argument has to do with the human element that is present in judging an agent's guilt. Consider the optimal CIS under full commitment power. In this system, agents who refuse to confess are to be punished if their evidence level is too low. By the nature of this system, it is known that the only agents who refuse to confess are innocent and yet the law would require the law enforcement institutions to punish them. But these law enforcement institutions are the ones that collect (in the case of the police) and assess (in the case of the judge or jury) the evidence. If they know that the agent is innocent (by observing that the agent chose not to confess to having committed the crime), it seems reasonable to think that they would claim that the evidence level is low even if it is high, in order to avoid convicting the agent they know is innocent.

In the American criminal justice system there are some examples of this phenomenon, where there seems to be an attempt to condition the way jurors appreciate the defendant's guilt. One such example is the inadmissibility of plea discussions in court, according to Rule 410 of the Federal Rules of Criminal Procedure. Another debated issue concerns the orders given to jurors at criminal trials by the judge to disregard some prosecutorial elements of the case - for example they are told they should not infer anything from the fact that the agent chose not to testify. As Laudan (2006) points out, this practice precludes important information from the trial and conditions how jury members assess the defendant's guilt. Whether these recommendations are indeed taken into account by the jurors is a matter of discussion: Laudan (2006) cites Posner (1999) on this matter: "Judges who want jurors to take seriously the principle that guilt should not be inferred from a refusal to waive the privilege against self-incrimination will have to come up with a credible explanation for why an innocent person might fear the consequences of testifying".

In my opinion, the proper assumption over the principal's commitment power depends very much on how one feels about these attempts at conditioning guilt assessment. If one believes that police, judges and jurors always follow the law and enforce punishments they know are unfair, then the relevant assumption should be of full commitment power. If not, then one accepts that the principal has some limited commitment power and the relevant analysis is that of section 7.

8 Concluding remarks

The idea that I explore in the paper is that CISs are, in general, a good idea because there are information externalities to each confession: when an agent confesses to be guilty he is providing the principal with the information that other agents are likely to be innocent. It then follows from my analysis that all members of the community should be allowed to confess the crime in exchange for a constant punishment, even before any investigation has been initiated (or as soon as possible). Even though this might appear as a radical suggestion, there are variants of CIS's already in American law. Self-reporting in environmental law works in very much the same way, even though it is mostly motivated by an attempt to reduce monitoring costs. In that context, agents are firms which are able to confess to having broken environmental regulations in exchange for smaller punishments. And even in criminal law, plea bargaining also allows agents to confess. In this case, agents are defendants and, typically, the bargaining occurs only when there is a single defendant, which largely defeats the purpose of having agents confess, according to my analysis. A confession produces no information externalities if there are no other agents to consider. In that sense, this paper can be seen as providing an argument for plea discussions to occur earlier in the criminal process, at a time when there are several suspects of committing the crime.

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 CIS 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 CIS 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.

9 Appendix

9.1 Appendix A - Extensions

9.1.1 Appendix A1: Conspiracies

In the main text, I have maintained the assumption that each agent knows only whether they are innocent or guilty and have no other information about the crime. By making this assumption, I have implicitly ruled out criminal conspiracies. When a group of agents commits a crime together, it seems reasonable to expect them to know the identity of the remaining conspirators. For example, if a group of 3 agents robs a bank, it is very likely that each of them will know the identity of the others. In this section, I extend the model to accommodate for this possibility and investigate how the optimal mechanism changes if the principal believes that a criminal conspiracy might be behind the crime.

I assume that, for each event $t \in T$, there is a commonly known probability $p(t) \in [0,1]$ that each guilty agent knows the identity of the remaining criminals (and so knows vector t). So, for example, if $N = 3$ and $p((g,g,i)) = 0.75$, it means that, when the crime is committed by agents 1 and 2, there is a 75% chance that the agents committed the crime together and know each other's identity. Hence, in that case, agents 1 and 2 would know that vector (g,g,i) had been realized. With 25% chance, agents 1 and 2 committed the crime independently and do not know whether any of the other agents is also guilty. In either case, agent 3 is innocent and forms beliefs about agents 1 and 2's guilt as before.

In this setup, because agents' beliefs do not depend only on whether they are innocent or guilty, it is necessary to enlarge the set of types that each agent might have. Let $\widehat{t}_n \in \widehat{T}_n$ denote agent n's *extended* type, where $\widehat{T}_n = \{i\} \cup \{\widehat{g}\} \cup T$. If $\widehat{t}_n = i$, then the agent is innocent; if $\widehat{t}_n = \widehat{g}$, then the agent is guilty but does not know t; and, finally, if $\widehat{t}_n = t \in T$, then the agent is guilty and knows that vector t has been realized.

For simplicity, I consider only the case of $\phi = 1$ and assume that the principal has commitment power. Also, without loss of generality, I assume that $M_n = \widehat{T}_n$ for all n.

Let $L \subset \widehat{T} = \widehat{T}_1 \times ... \times \widehat{T}_N$ be the set of extended types that do not have a strictly positive measure. For example, in the case of $N = 2$, $\widehat{t} = ((g,g), i) \in L$ because if agent 1 is guilty and part of a conspiracy with agent 2, it must be that agent 2's extended type is (g,g).

Let the mechanism $\widehat{x}^{SB} : \widehat{T} \times \Theta \rightarrow \mathbb{R}_+^N$ be defined as follows. For all $\widehat{t} \in L$, $\theta \in \Theta$ and for all n, $\widehat{x}_n^{SB}\left(\widehat{t}, \theta\right) = 1$. For all $\widehat{t} \in \widehat{T} \backslash L$ and for all $\widehat{t}_{-n} \in \widehat{T}_{-n}$ (where $\widehat{T}_{-n}$ is defined as usual), $\theta \in \Theta$, and for all n,

$$\begin{cases} \widehat{x}_n^{SB}\left(i, \widehat{t}_{-n}, \theta\right) = \begin{cases} 1 \text{ if } \pi\left(t_n = g | \widehat{t}_{-n}, \theta\right) > \alpha\pi\left(t_n = i | \widehat{t}_{-n}, \theta\right) \\ 0 \text{ otherwise} \end{cases} \\ \widehat{x}_n^{SB}\left(\widehat{t}_n, \widehat{t}_{-n}, \theta\right) = \widehat{\varphi}_n\left(\widehat{t}_{-n}\right) \text{ for all } \widehat{t}_n \neq i \end{cases}$$

where $\widehat{\varphi}_n$ is characterized in the proof of Proposition 18.

Proposition 18 *Mechanism $\widehat{x}^{SB}$ is optimal within the set of incentive compatible mechanisms.*

Proof. See Appendix B8. ■

If agents produce a report $\widehat{t} \in L$, the principal realizes one of them is lying. So, in order to induce truthful reporting, it is in her best interest to punish the agents as much as possible. The rest of the mechanism is constructed using the same principle as in the main text. The principal is able to get agents to report to be guilty by guaranteeing that such information will not be used against them but only against other agents. In the first stage of the "extended" CIS, and in the same way as in the standard CIS, agents are given the opportunity to confess. However, they are also asked to report any other information they might have, in particular, whether there are other guilty agents and their identity. By construction of $\widehat{x}^{SB}$, guilty agents are indifferent between confessing and refusing to confess, while innocent agents prefer the latter. These proceed to the second stage and are judged only with the information the principal can gather from other agents and the evidence. Another feature of this system is that agents who confess no longer receive a constant punishment. With this information structure, guilty agents might have different beliefs about the guilt or innocence of other agents. This means that a constant punishment which leaves a guilty agent of extended type $\widehat{g}$ indifferent might not leave him indifferent if he has some other extended type. However, these different extended types of guilty agents all have the same beliefs with respect to the evidence the agent himself generates. Therefore, the punishment an agent receives when he confesses only depends on the type of information that other agents grant the principal $(\widehat{t}_{-n})$ and not on the evidence.

So far in the paper, I have assumed that, given a mechanism, the principal can choose the strategy profile if there are multiple ones which are a Bayes Nash Equilibrium of the game induced by the mechanism. The implicit justification for this approach is that, if there is some equilibrium more likely than the others, the equilibrium where guilty agents confess and innocents do not seems like a natural candidate. However, when extending the analysis to consider conspiracies, this approach is more problematic as it is possible that the conspirators could coordinate on a deviation that would make them both better off. I illustrate with the example of section 4.

Example 19 *Consider the same parameters of section* 4 *with* $\delta = 0$*, but extend it to consider conspiracies. Assume that* $p(t) = 0$ *for all* $t \in T$ *except* $p((g,g)) = \frac{1}{2}$*. Basically, the setting is as before except that, if both agents are guilty, there is a* 50% *probability that they committed the crime together.*

Consider what would happen to agent $n = 1$ *with type* $\widehat{t}_1 = (g,g)$ *in the CIS described in proposition* 18*, when agents report truthfully. Given that agent* 2 *will report* $\widehat{t}_2 = (g,g)$*, anything that agent* 1 *reports will lead to a punishment of* 1*: if he reports some* $\widehat{t}_1' \neq (g,g)$*, the principal will know one of the agents is lying and will give both a punishment* 1*, while if he reports* $\widehat{t}_n = (g,g)$*, his punishment is based on the other agent's report, which incriminates him. Therefore, the agent's expected punishment would be* 1*.*

An alternative option for agent 1 *is to reach out to agent* 2*, who he knows is guilty, and convince him that it would be a good idea if they both claimed to be innocent. In this case, given the mechanism* $\widehat{x}^{SB}$*, agent* 1 *would receive a punishment of* 1 *if*

$$\pi(t_1 = g | t_2 = i, \theta) > \alpha \pi(t_1 = i | t_2 = i, \theta) \tag{8}$$

and would be acquitted otherwise. Notice that (8) can be written simply as

$$\theta_1 > \frac{1}{5}$$

so the expected punishment of agent 1 *in this equilibrium, if his type is* (g,g) *is*

$$\int_{\frac{1}{5}}^{1} 2\theta_1 d\theta_1 = \frac{24}{25} < 1$$

The same would be true for agent 2*, so they would both be better off reporting to be innocent. However, it is possible to alter the mechanism* $\widehat{x}^{SB}$ *in a way that eliminates this undesirable equilibrium, while keeping intact the equilibrium where agents report truthfully. The idea is to change what happens if agents are "caught lying".*

In the mechanism $\widehat{x}^{SB}$*, the punishment for an agent who confesses to be part of a conspiracy is always* 1*. Take mechanism* $\widehat{x}'$ *to be the same as mechanism* $\widehat{x}^{SB}$ *except that*

$$\widehat{x}'_n\left((g,g),i\right) = \begin{cases} 1 \text{ } if\, \theta_{-n} < d_n \\ 0 \text{ } otherwise \end{cases} \quad for\, n = 1, 2$$

where $d_n \in (0,1)$*. In other words, if agent* n *confesses to be guilty and incriminates agent* $-n$*, he gets to be acquitted if and only if the evidence supports his claim that the other agent is guilty (if* $\theta_{-n} \geq d_n$*). By choosing* d_n *properly, it is possible to make it so that agent* n *only incriminates the other agent when he knows him to be guilty.*

In general, by making similar changes to the punishments after reports that contradict each other ($\widehat{t} \in L$), it is possible to transform the extended CIS in order to eliminate the incentives that conspiracy members have in colluding in the report they submit to the principal.

9.1.2 Appendix A2: Risk Aversion

One of the assumptions of this paper is that agents are risk neutral. This might lead the reader to inquire on whether CIS's would still be appealing if agents were risk averse. The concern might be that agents choose to confess because they are risk averse and not because they are guilty. In order to address this issue, in this section, I extend the analysis to consider arbitrary levels of risk aversion for the agents and for the principal in a setup close to that of the independent work of Siegel and Strulovici (2016). Proposition 20 corroborates that paper's result in arguing that enlarging the set of possible verdicts increases the expected utility of the principal, while Proposition 21 can be understood as a special case of their analysis where, given a specific functional form for the agents' utility function, I show that a CIS is still an optimal system if the principal has commitment power.

Recall that $u^i(\cdot)$, $u^g(\cdot)$ denote the agent's utility if he is innocent and guilty respectively and $u_n^p(t_n, \cdot)$ is the principal's utility when the agent is of type t_n. In this section, I assume that $u^i(x_n) = -x_n^{\omega_i}$ and $u^g(x_n) = -x_n^{\omega_g}$, where $\omega_i > 1$ and $\omega_g > 1$, so that each agent is risk averse. Furthermore, I assume that, for all n, $u_n^p(i, \cdot)$ is strictly decreasing, $u_n^p(g, \cdot)$ is single peaked around 1 and both are strictly concave and differentiable.

Let $\widetilde{x}^{Tr}$ denote the optimal trial mechanism.

Proposition 20 *For all n, if $\frac{\partial u_n^p(i,0)}{\partial x_n} = 0$, then $\widetilde{x}_n^{Tr}(\theta)$ is continuous, strictly increasing with θ_n and is such that, for all θ_{-n}, $\lim_{\theta_n \to 0} \widetilde{x}_n^{Tr}((\theta_n, \theta_{-n})) = 0$ and $\lim_{\theta_n \to 1} \widetilde{x}_n^{Tr}((\theta_n, \theta_{-n})) = 1$.*

Proof. See Appendix B9. ■

In the optimal trial system, punishments are determined only by the preferences of the principal. If the principal is risk averse, then she prefers to smooth punishments rather than adopt a "bang-bang" solution like in the main text. In particular, the punishment the principal imposes is strictly increasing with her belief about each agent's guilt.

Now, consider the problem of finding the optimal incentive compatible system. Without loss of generality, let $M_n = T_n$ for all n and consider only the strategy profile where agents report truthfully. Let $\widetilde{x}^{SB}$ denote the optimal incentive compatible mechanism.

Proposition 21 *For all n, if $u_n^p(i, x_n) = \alpha u^i(x_n)$ for all x_n and for some $\alpha > 0$, then $\widetilde{x}_n^{SB}(g, t_{-n}, \theta)$ is independent of t_{-n} and θ and equal to*

$$\sum_{t_{-n} \in T_{-n}} \int_{\theta} \frac{\pi(g, t_{-n}, \theta)}{\pi(t_n = g)} u^g\left(\widetilde{x}_n^{SB}(i, t_{-n}, \theta)\right) d\theta$$

Proof. See Appendix B10. ■

Recall that, in this paper, the principal is interpreted as being benevolent - similar to a social planner - and so, it seems reasonable to me to assume that, if the principal faces an innocent agent, she would want to maximize his expected utility. Assuming that $u_n^p(i, \cdot)$ is proportional to $u^i(\cdot)$ implies precisely that - the principal has the same preferences of the innocent agent when she knows him to be innocent. This assumption is convenient in that it guarantees that innocents' incentive constraints do not bind.

Proposition 21 implies that, if the agents and the principal are risk averse, the optimal mechanism is a CIS, where guilty agents confess the crime and receive a constant punishment in return. The intuition for the result is as follows. In the optimal mechanism, guilty agents must be indifferent between reporting truthfully and reporting to be innocent (for otherwise the principal could reduce the punishments innocent agents receive) and must be receiving punishments that never exceed 1 (for otherwise those punishments could be reduced to 1, which would increase the principal's expected utility and give more incentives for guilty agents to report truthfully). Suppose that, in the optimal mechanism, a guilty agent receives a lottery of distinct punishments. Because the guilty agent is risk averse, he would be willing to accept, as an alternative, a constant punishment larger than the expected punishment of the lottery. The principal would strictly prefer this alternative, as long as she is (weakly) risk averse.

Notice that, if agents and principal are risk averse, the case for CIS's is even stronger, because, even if there is only one agent ($N = 1$) and even if punishments cannot exceed 1, it is still strictly better to have CIS's than to have any other system. In particular, it is not the case that, if agents are made more and more risk averse, they eventually confess regardless of their guilt. That argument assumes that the principal is unaware of how risk averse the agents are. If the principal knows the agents' preferences, she is able to select punishments in such a way that only guilty agents choose to confess, by using the fact that guilty agents are more afraid that future evidence and other agents might incriminate them.

The following proposition characterizes how the optimal mechanism depends on the risk aversion level of innocent and guilty agents.

Proposition 22 *For all n, if $u_n^p(i, x_n) = \alpha u^i(x_n)$ for all x_n and for any $\alpha > 0$, then*

i) If $\omega_i > \omega_g$ (innocent agents are more risk averse than guilty agents) then

$$\widetilde{x}_n^{SB}(i, t_{-n}, \theta) = \begin{cases} \phi \textit{ if } \theta_n > \widetilde{\theta}_n^{SB(i)}(t_{-n}) \\ \psi_n^{SB}(\theta_n, t_{-n}) \textit{ otherwise} \end{cases}$$

where $\psi_n^{SB}(\theta_n, t_{-n})$ is continuous and strictly increasing with θ_n.

ii) If $\omega_i \leq \omega_g$ (guilty agents are more risk averse than innocent agents) then

$$\widetilde{x}_n^{SB}(i, t_{-n}, \theta) = \begin{cases} \phi \textit{ if } \theta_n > \widetilde{\theta}_n^{SB(g)}(t_{-n}) \\ 0 \textit{ otherwise} \end{cases}$$

Expressions $\widetilde{\theta}_n^{SB(i)}(t_{-n})$, $\widetilde{\theta}_n^{SB(g)}(t_{-n})$ and $\psi_n^{SB}(\theta_n, t_{-n})$ are characterized in the proof.

Proof. See Appendix B11. ■

When the principal is determining the optimal punishments to impose on innocent agents she faces a trade-off. On the one hand, she would like to select small punishments in order to spare the innocents as much as possible. But on the other hand, those punishments determine the punishment that guilty agents receive in equilibrium. So, the principal wants to construct a lottery of punishments that is very appealing for those who are innocent but very unappealing for those who are guilty. If innocent agents are more risk averse than guilty agents, then smoothing punishments is relatively better for them, which is why, if $\omega_i > \omega_j$, the punishments innocent agents receive are strictly increasing and continuous until hitting the upper bound of ϕ. If, on the contrary, guilty agents are more risk averse, following a similar strategy would be relatively better to guilty agents. Therefore, even though agents are strictly risk averse regardless of whether they are innocent or guilty, if $\omega_i \leq \omega_j$, it is still better for the principal to impose a risky lottery of punishments, where agents are punished very harshly only for very high levels of evidence, and are acquitted otherwise.

9.1.3 Appendix A3: Deterrence

In the main text, unlike some of the literature on law enforcement, I do not explicitly model the agents' decision of committing the crime.[16] 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.[17] 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 the crime is the difference between the expected punishment that the agent will receive if he commits a crime (and becomes guilty) and if he does not (and becomes innocent). Let B_n^g and B_n^i denote the expected punishment of agent n if he chooses to commit the crime and 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

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

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

$$\sum_{n=1}^{N} \left(B_n^g - B_n^i \right) \tag{9}$$

[16] 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.

[17] See Garoupa (1997).

Consider first the case where there is no upper bound on punishments. In this case, it is possible to achieve a world with virtually no crime, and so

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

The idea is to build a mechanism that requires a very large standard of proof but then imposes very large punishments (as one of the mechanisms in the Example of section 4). For example, consider mechanism $\widehat{x}$ such that

$$\widehat{x}_n(m,\theta) = \begin{cases} k \text{ if } m_n = c \\ 0 \text{ if } m_n = \overline{c} \text{ and } \theta_n \leq \widehat{\varepsilon}(\phi) \\ \phi \text{ if } m_n = \overline{c} \text{ and } \theta_n > \widehat{\varepsilon}(\phi) \end{cases}$$

where $k > 0$ and $\widehat{\varepsilon}(\phi) \in (0,1)$ is such that, in the event that agent n commits the crime, he is indifferent between reporting message c and message $\overline{c}$, which implies that $B_n^g = k$. It follows, by the same argument as in the Example of section 4, that

$$\lim_{\phi\to\infty} B_n^i = 0$$

Seeing as one can choose k to be arbitrarily large, expression (9) can also be made arbitrarily large.

Say that, for some reason other than the principal's preferences, as those are simply expressed by condition (9), there is an upper bound on punishments. What would be the optimal incentive compatible system in this case?

Notice that, without loss of generality, one can normalize the upper bound on punishments to 1. So, in this interpretation, a punishment of 1 is the largest available punishment. It is **not** the punishment that fits the crime, as there is no such thing if the preferences of the principal are given by (9). However, it is possible to use the work done in section 6 to answer this question.

Recall that, in section 6, I showed that, if the preferences of the principal were not given by (9) but were instead given by

$$u^p = \sum_{n=1}^{N} u_n^p$$

the optimal incentive compatible system was a CIS for any value of $\phi \geq 1$ and $\alpha > 0$. In particular, this was also true if $\phi = 1$. But if $\phi = 1$, it follows that

$$E\left(u_n^p\right) = -\pi\left(t_n = g\right)\left(1 - B_n^g\right) - \alpha\pi\left(t_n = i\right) B_n^i$$

and if

$$\alpha = \frac{\pi\left(t_n = g\right)}{\pi\left(t_n = i\right)}$$

then

$$E\left(u_n^p\right) = \pi\left(t_n = g\right)\left(B_n^g - B_n^i\right) - \pi\left(t_n = g\right) \tag{10}$$

As is clear, maximizing (10) for each agent n is exactly the same as maximizing (9).[18] Therefore, the optimal incentive compatible mechanism if there is an upper bound on punishments and the preferences of the principal are given by (9) is a CIS, where guilty agents confess and receive a constant punishment, while innocent agents never confess but may be punished with the largest available punishment, depending on the evidence.

The analysis of section 7 about limited commitment power does not apply if the only goal of the criminal justice system is to deter crime. The idea behind deterrence is that society punishes those who are believed to be guilty not because it has any desire to do so, but because it generates the belief that anyone who is thought to be guilty in the future shall receive a similar punishment. In other words, if society could not commit, no one would ever be punished, as as there is no intrinsic desire to punish those who are guilty.

[18] The only issue is that, as defined, there is only one α for all agents, which makes it impossible to set

$$\alpha = \frac{\pi\left(t_n = g\right)}{\pi\left(t_n = i\right)} = \frac{\pi\left(t_{n'} = g\right)}{\pi\left(t_{n'} = i\right)}$$

unless

$$\frac{\pi\left(t_n = g\right)}{\pi\left(t_n = i\right)} = \frac{\pi\left(t_{n'} = g\right)}{\pi\left(t_{n'} = i\right)}$$

However, this problem is easily bypassed by simply assuming that there are potentially different α_n for each agent. All results from the previous sections hold as each nth problem is independently solved.

9.1.4 Appendix A4: Timing

In this extension, I consider a different timing of events and assume that the principal selects the mechanism after an investigation has taken place, and, so, after having observed evidence θ, which becomes her private information. The main point is to show that, if she is able to, the principal is better off by proposing the mechanism before the investigation.[19] I assume that $\phi = 1$ and only consider the case where the principal has commitment power for simplicity.

The sequence of events is as follows. First, evidence θ is realized and privately observed by the principal. After observing θ, the principal selects a mechanism $y_\theta : T \times \Theta \to [0,1]^N$, which maps the agents' types (without loss of generality, by the revelation principle) and evidence to punishments. Following the announcement of the mechanism, each agent chooses what to report and, finally, punishments are allocated according to the announced mechanism.

A strategy y for the principal is a specification of y_θ for all θ. Knowing y, the agents are now able to infer about the realized θ through the principal's specific proposal y_θ. The principal will then face a dilemma. She would prefer to tailor her proposal y_θ to the evidence gathered θ, but doing so runs the risk of allowing the agent to infer θ from the proposal itself, which might be detrimental to her.

The relevant solution concept in this framework is Perfect Bayesian Equilibrium (PBE), where i) given their beliefs, each agent prefers to report truthfully after the principal's proposal and given that all other agents do so; ii) after each θ and given the agents' beliefs, the principal prefers to select y_θ and not some other mechanism $\widetilde{y}_\theta : T \times [0,1]^N \to [0,1]^N$ for which it is a (Bayes-Nash) equilibrium for agents to report truthfully given their beliefs; and iii) agents' beliefs are consistent with Bayes' rule.

Proposition 23 *The principal is better off proposing the mechanism before the investigation.*

Proof. See Appendix B12. ■

[19] This type of problem is referred to in the literature as the informed principal problem - some of the classic references are Myerson (1983) and Maskin and Tirole (1990).

The intuition for this result is very similar to the inscrutability principle of Myerson (1983). Say that the principal only acts after observing the evidence and there is an equilibrium where agents prefer to report truthfully no matter what the proposal y_θ. That means that, on average, agents prefer to report truthfully, where the average is taken with respect to θ. But then, if the principal just made a single proposal $x_y : T \times \Theta \rightarrow \mathbb{R}_+^N$ before observing evidence, such that $x_y\left(t,\theta\right) = y_\theta\left(t,\theta\right)$ for all $t \in T$ and $\theta \in \Theta$, agents would report truthfully and the punishment allocation would be exactly the same as the one induced by y.

9.1.5 Appendix A5: Heterogeneous agents

In the main text, I have assumed that the distribution of the evidence level of each agent only depended on the guilt of that agent. However, it is likely the case that guilty agents are better informed about the distribution of the evidence than the principal. It could be that a given guilty agent is more skilled in the art of committing crimes and, so, is less likely to produce incriminating evidence. It can also be that agents are unlucky and leave some evidence behind - maybe someone who has robbed a bank has dropped their wallet in the escape. Even innocent agents are likely to have some private information as to whether the evidence is more or less likely to incriminate them. For example, it could be that, even though an agent is innocent, he was at the crime scene only a few moments before the crime and there is a considerable probability that his fingerprints will be found. One way to extend the model to allow for this type of heterogeneity is to assume that each agent n is privately informed of a random variable $\beta_n \in [0,1]$, which determines the distribution of the evidence. In particular, let

$$\pi\left(\theta_n|\beta_n\right) = \beta_n f^g\left(\theta_n\right) + \left(1-\beta_n\right) f^i\left(\theta_n\right)$$

denote the conditional distribution of θ_n given the agent's β_n where $\frac{f^g(\theta_n)}{f^i(\theta_n)} = l\left(\theta_n\right)$ for all θ_n. The idea is that β_n and $\left(1-\beta_n\right)$ are the weights put on the distributions f^g and f^i respectively. In the main text, the assumption was that, if agent n was guilty, then $\beta_n = 1$ while, if he was innocent, then $\beta_n = 0$ and this was commonly known. In this extension, I assume that β_n is only privately known by each agent and its distribution depends only on whether agent n is innocent or guilty. By assuming that $\frac{\pi(\beta_n|t_n=g)}{\pi(\beta_n|t_n=i)}$ is strictly increasing for all $\beta_n \in [0,1]$, it is possible to recover the idea that guilty agents are more likely to draw worse evidence, because they are more likely to generate a larger β_n. I also assume, for simplicity, that $\pi(\beta_n|t_n)$ has full support, is continuous and differentiable for $t_n = i, g$.

Proposition 24 below characterizes how each agent acts in the optimal CIS when $\phi = 1$ and the principal has commitment power.

Proposition 24 *For all n, there is $(\beta_n^i, \beta_n^g) \in [0,1]^2$ such that, for all $t_n \in \{i, g\}$ and $\beta_n \in [0,1]$,*

$$s_n(t_n, \beta_n) = \begin{cases} c \text{ if } \beta_n \geq \beta_n^{t_n} \\ \overline{c} \text{ otherwise} \end{cases}$$

where $s_n(t_n, \beta_n) \in \{c, \overline{c}\}$ represents the message that agent n with type t_n and β_n chooses.

Proof. See Appendix B13. ■

Agents that have a larger β_n are more likely to generate more incriminating evidence. Hence, they have a larger incentive to confess (and select message c) than those with a smaller β_n. If the agents' innocence is not independent, it is easy to show that $\beta_n^i > \beta_n^g$ - for a given β_n the agent has more incentives to confess if he guilty than if he is innocent. This is because he is more afraid that the other agents' reports and evidence might incriminate him.

If there are homogeneous types as in the main text, $\beta_n^i = 1$ while $\beta_n^g = 0$ so that only guilty agents confess. However, in general, it is not in the best interest of the principal to do this if the agents are heterogeneous. Suppose the principal wants to guarantee that the agent confesses if he is guilty no matter what β_n he draws. For this to be possible, it must be that the punishment upon a confession is small enough that even if the guilty agent draws $\beta_n = 0$, he still prefers to confess. But establishing such a small punishment leads to innocent agents confessing. For example, if there is no correlation between the agents' innocence (and so a guilty and an innocent agent have the same beliefs, conditional on drawing the same β_n), the agent also confesses when he is innocent, regardless of β_n.

Finally, notice that a CIS might not be optimal in this setting. Consider a given set of parameters for which the optimal CIS is such that $\beta_n^i = 1$ for all n so that guilty agents are the only ones who confess (the following argument could also be made if only a small fraction of innocent agents confesses). Of these, only a small fraction is made indifferent (which has a 0 measure) - the pair (g, β_n^g) for each agent n. This means that anytime a guilty agent draws $\beta_n > \beta_n^g$ and chooses to confess, he is strictly better off than choosing not to. Thus, a more successful mechanism would be to punish agents that confess as if they did not. The principal would still solicit a report from the agents on whether they are innocent or guilty, and punishments that follow an innocent report would still be the same as in the optimal CIS. The difference would be that agents who confess would also face the same lottery of punishments as if they chose not to. They would still have enough incentives to confess (because they would be indifferent), but their expected punishment would be larger.

Of course, a problem with this system is whether it is robust enough. In this alternative system, someone who is guilty receives exactly the same lottery of punishments regardless of whether he confesses or not. So, the agent might be inclined to claim to be innocent in the hope that, if is there is some error in the implementation of the mechanism, it would favor those who claim to be innocent. In the CIS this is not a problem as only a small fraction of agents are actually indifferent. And even when agents are homogeneous (when guilty agents have $\beta_n = 1$ and innocent agents have $\beta_n = 0$) and the optimal CIS is such that all guilty agents are made indifferent, it is easy to accommodate for these types of concerns by simply decreasing the punishment that follows a confession in a small amount so that guilty agents are no longer indifferent but rather strictly prefer to confess.

9.2 Appendix B - Proofs

9.2.1 Appendix B1

Notice that, for all $m \in M$, $\theta \in \Theta$ and n, $x_n^{Tr}(m,\theta) = 1$ if and only if

$$\begin{aligned}
\pi(t_n = g|\theta) &> \alpha\pi(t_n = i|\theta) \Leftrightarrow \\
\sum_{t_{-n}\in T_{-n}} \pi(g,t_{-n})\,\pi(\theta|g,t_{-n}) &> \alpha \sum_{t_{-n}\in T_{-n}} \pi(i,t_{-n})\,\pi(\theta|i,t_{-n}) \Leftrightarrow \\
\pi(\theta_n|t_n = g) \sum_{t_{-n}\in T_{-n}} \pi(g,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}}) &> \alpha\pi(\theta_n|t_n = i) \sum_{t_{-n}\in T_{-n}} \pi(i,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}}) \Leftrightarrow \\
l(\theta_n) &> \alpha \frac{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(i,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})}{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(g,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})} \Leftrightarrow \\
\theta_n &> l^{-1}\left(\alpha \frac{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(i,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})}{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(g,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})}\right) \Leftrightarrow
\end{aligned}$$

and so

$$\theta_n^{Tr}(\theta_{-n}) = l^{-1}\left(\alpha \frac{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(i,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})}{\displaystyle\sum_{t_{-n}\in T_{-n}} \pi(g,t_{-n}) \prod_{\widetilde{n}\neq n} \pi(\theta_{\widetilde{n}}|t_{\widetilde{n}})}\right)$$

9.2.2 Appendix B2

Denote by $\widehat{\lambda}_n \geq 0$ the lagrange multiplier associated with the constraint (6) and let $\widehat{\zeta}(m_{-n},\theta) \geq 0$ and $\widehat{\eta}(m_{-n},\theta) \geq 0$ be the multipliers associated with $x_n(i,m_{-n},\theta) \leq \phi$ and $x_n(i,m_{-n},\theta) \geq 0$ respectively. It follows that in any solution for this problem,

$$\begin{aligned}
&\left(\pi^{\sigma^{SB}}(t_n = g|m_{-n},\theta) - \alpha\pi^{\sigma^{SB}}(t_n = i|m_{-n},\theta)\right)\pi^{\sigma^{SB}}(m_{-n},\theta) + \widehat{\eta}(m_{-n},\theta) \quad (11)\\
= \;& \widehat{\lambda}_n \frac{\pi^{\sigma^{SB}}(t_n = g|m_{-n},\theta)\,\pi^{\sigma^{SB}}(m_{-n},\theta)}{\pi(t_n = g)} + \widehat{\zeta}(m_{-n},\theta)
\end{aligned}$$

Let $q_{\widehat{n}} : T_{\widehat{n}} \rightarrow M_{\widehat{n}}$ for any $\widehat{n}$ be a one to one function such that

$$q_{\widehat{n}}\left(t_{\widehat{n}}\right) = \begin{cases} c \text{ if } t_{\widehat{n}} = g \\ \overline{c} \text{ if } t_{\widehat{n}} = i \end{cases}$$

and $q_{-n}\left(t_{-n}\right) = \left(q_1\left(t_1\right), ..., q_{n-1}\left(t_{n-1}\right), q_{n+1}\left(t_{n+1}\right), ..., q_N\left(t_N\right)\right)$. Given that agents report truthfully, one can write (11) as

$$\begin{aligned} & \left(\pi^{\sigma^{SB}}\left(t_n = g|q_{-n}^{-1}\left(m_{-n}\right), \theta\right) - \alpha\pi^{\sigma^{SB}}\left(t_n = i|q_{-n}^{-1}\left(m_{-n}\right), \theta\right)\right)\pi^{\sigma^{SB}}\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right) + \widehat{\eta}\left(m_{-n}, \theta\right) \\ = \; & \widehat{\lambda}_n \frac{\pi^{\sigma^{SB}}\left(t_n = g|q_{-n}^{-1}\left(m_{-n}\right), \theta\right)\pi^{\sigma^{SB}}\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right)}{\pi\left(t_n = g\right)} + \widehat{\zeta}\left(m_{-n}, \theta\right) \end{aligned}$$

which, in turn, can be simplified to

$$\pi\left(g, q_{-n}^{-1}\left(m_{-n}\right)\right) l\left(\theta_n\right)\left(1 - \lambda_n\right) - \alpha\pi\left(i, q_{-n}^{-1}\left(m_{-n}\right)\right) = \zeta\left(m_{-n}, \theta\right) - \eta\left(m_{-n}, \theta\right) \quad (12)$$

where

$$\lambda_n = \frac{\widehat{\lambda}_n}{\pi\left(t_n = g\right)},$$

$$\zeta\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right) = \frac{\widehat{\zeta}\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right)}{\pi\left(\theta_n|t_n = i\right)\prod_{\widetilde{n}\neq n}\pi\left(\theta_{\widetilde{n}}|q_{\widetilde{n}}^{-1}\left(m_{\widetilde{n}}\right)\right)}$$

and

$$\eta\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right) = \frac{\widehat{\eta}\left(q_{-n}^{-1}\left(m_{-n}\right), \theta\right)}{\pi\left(\theta_n|t_n = i\right)\prod_{\widetilde{n}\neq n}\pi\left(\theta_{\widetilde{n}}|q_{\widetilde{n}}^{-1}\left(m_{\widetilde{n}}\right)\right)}$$

Notice that, for a fixed $m_{-n} \in M_{-n}$, the LHS of (12) is strictly increasing with θ_n, which means that there is a threshold $\theta_n^{SB}\left(m_{-n}, \theta_{-n}\right)$ such that

$$x_n^{SB}\left(\overline{c}, m_{-n}, \theta\right) = \begin{cases} \phi \text{ if } \theta_n > \theta_n^{SB}\left(m_{-n}, \theta_{-n}\right) \\ 0 \text{ otherwise} \end{cases}$$

where ties are resolved in favor of an acquittal. The threshold $\theta_n^{SB}\left(t_{-n}\right)$ is such that

$$\pi\left(g, q_{-n}^{-1}\left(m_{-n}\right)\right) l\left(\theta_n^{SB}\left(m_{-n}, \theta_{-n}\right)\right)\left(1 - \lambda_n\right) - \alpha\pi\left(i, q_{-n}^{-1}\left(m_{-n}\right)\right) = 0$$

and so

$$\theta_n^{SB}\left(m_{-n},\theta_{-n}\right)=l^{-1}\left(\frac{\alpha}{1-\lambda_n}\frac{\pi\left(i,q_{-n}^{-1}\left(m_{-n}\right)\right)}{\pi\left(g,q_{-n}^{-1}\left(m_{-n}\right)\right)}\right)$$

and is independent of θ_{-n} (other agents' evidence is irrelevant if the principal has the information about the other agents' types).

As for λ_n, it is equal to 0 whenever the constraint does not bind. Let

$$B_n\left(\phi,\lambda_n\right)=\phi\sum_{t_{-n}\in T_{-n}}\frac{\pi\left(g,t_{-n}\right)}{\pi\left(t_n=g\right)}\int\limits_{l^{-1}\left(\frac{\alpha}{1-\lambda_n}\frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)}\right)}^{1}\pi\left(\theta_n|t_n=g\right)d\theta_n$$

which represents the expected punishment of the guilty agent under threshold θ_n^{SB}, given that he is indifferent between reporting truthfully and misreporting. Then, it follows that

$$\lambda_n=\begin{cases}0\text{ if }B_n\left(\phi,0\right)\leq 1\\ \lambda_n^*\text{ otherwise}\end{cases}$$

where λ_n^* is such that $B_n\left(\phi,\lambda_n^*\right)=1$. Notice that, for any ϕ, λ_n always exists and is strictly increasing for all $\phi\geq\overline{\phi}_n>1$ where $\overline{\phi}_n$ is such that $B_n\left(\overline{\phi}_n,0\right)=1$.

9.2.3 Appendix B3

Recall that from Appendix B2, the threshold $\theta_n^{SB}\left(m_{-n},\theta_{-n}\right)$ is independent of θ_{-n}, so, without loss of generality, I refer to it as simply $\theta_n^{SB}\left(m_{-n}\right)$.

Let $\overline{\phi}=\max\left\{\overline{\phi}_n\right\}_{n=1}^{N}$, so that, for all $\phi>\overline{\phi}$ and for all n,

$$B_n^g\left(x_n^{SB}\right)=\phi\sum_{t_{-n}\in T_{-n}}\frac{\pi\left(g,t_{-n}\right)}{\pi\left(t_n=g\right)}\int\limits_{\theta_n^{SB}\left(q_{-n}\left(t_{-n}\right)\right)}^{1}\pi\left(\theta_n|t_n=g\right)d\theta_n=1 \qquad (13)$$

and

$$B_n^i\left(x_n^{SB}\right)=\phi\sum_{t_{-n}\in T_{-n}}\frac{\pi\left(i,t_{-n}\right)}{\pi\left(t_n=i\right)}\int\limits_{\theta_n^{SB}\left(q_{-n}\left(t_{-n}\right)\right)}^{1}\pi\left(\theta_n|t_n=i\right)d\theta_n \qquad (14)$$

Given (13) we have that (14) is equivalent to

$$\begin{aligned}
B_n^i\left(x_n^{SB}\right) &= \frac{\phi \sum\limits_{t_{-n}\in T_{-n}} \frac{\pi(i,t_{-n})}{\pi(t_n=i)} \int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=i\right)d\theta_n}{\phi \sum\limits_{t_{-n}\in T_{-n}} \frac{\pi(g,t_{-n})}{\pi(t_n=g)} \int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=g\right)d\theta_n} \\
&= \frac{\pi\left(t_n=g\right)}{\pi\left(t_n=i\right)} \frac{\sum\limits_{t_{-n}\in T_{-n}} \pi\left(i,t_{-n}\right) \int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=i\right)d\theta_n}{\sum\limits_{t_{-n}\in T_{-n}} \pi\left(g,t_{-n}\right) \int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=g\right)d\theta_n} \\
&< \frac{\pi\left(t_n=g\right)}{\pi\left(t_n=i\right)} \sum_{t_{-n}\in T_{-n}} \left(\frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)} \frac{\int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=i\right)d\theta_n}{\int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \pi\left(\theta_n|t_n=g\right)d\theta_n} \right) \\
&< \frac{\pi\left(t_n=g\right)}{\pi\left(t_n=i\right)} \sum_{t_{-n}\in T_{-n}} \left(\frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)} \int\limits_{\theta_n^{SB}(q_{-n}(t_{-n}))}^{1} \frac{1}{l\left(\theta_n\right)}d\theta_n \right) \\
&< \frac{\pi\left(t_n=g\right)}{\pi\left(t_n=i\right)} \sum_{t_{-n}\in T_{-n}} \frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)} \frac{1}{l\left(\theta_n^{SB}\left(q_{-n}\left(t_{-n}\right)\right)\right)}
\end{aligned}$$

where the last inequality follows from the monotone likelihood ratio property on l. The last step is to realize that $\lim_{\phi\to\infty}\ \theta_n^{SB}\left(q_{-n}\left(t_{-n}\right)\right) = 1$ for all $t_{-n} \in T_{-n}$ (for otherwise the expected punishments would become arbitrarily large, violating the upper bound constraints), which implies that $\lim_{\phi\to\infty}\ l\left(\theta_n^{SB}\left(q_{-n}\left(t_{-n}\right)\right)\right) = \infty$, and so $\lim_{\phi\to\infty}\ B_n^i\left(x_n^{SB}\right) = 0$ for all n.

9.2.4 Appendix B4

Take any system (x, σ) where, for some n, there are m_n' and m_n'' such that

$$r_n\left(m_n'\right) \equiv \frac{\sigma_n\left(g, m_n'\right)}{\sigma_n\left(i, m_n'\right)} = \frac{\sigma_n\left(g, m_n''\right)}{\sigma_n\left(i, m_n''\right)} \equiv r_n\left(m_n''\right)$$

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 $r_n(\cdot)$ is $\mathbb{R}_+$ to which one adds the confessing message c.

Consider the alternative system $(\overline{x}, \overline{\sigma})$ that is equal to (x, σ) except that:

$$\begin{cases} i)\ \overline{\sigma}_n\left(t_n, m_n'\right) = \sigma_n\left(t_n, m_n'\right) + \sigma_n\left(t_n, m_n''\right) \text{ for } t_n = i, g \\ ii)\ \overline{\sigma}_n\left(t_n, m_n''\right) = 0 \text{ for } t_n = i, g \\ ii)\ \overline{x}\left(m_n', m_{-n}, \theta\right) = \begin{pmatrix} \frac{\sigma_n\left(t_n, m_n'\right)}{\sigma_n\left(t_n, m_n'\right) + \sigma_n\left(t_n, m_n''\right)} x\left(m_n', m_{-n}, \theta\right) \\ + \frac{\sigma_n\left(t_n, m_n''\right)}{\sigma_n\left(t_n, m_n'\right) + \sigma_n\left(t_n, m_n''\right)} x\left(m_n'', m_{-n}, \theta\right) \end{pmatrix} for\ t_n = i, g \\ iii)\ \overline{x}\left(m_n'', m_{-n}, \theta\right) = (1, ..., 1) \end{cases}$$

The new system merges the two messages and effectively eliminates message m_n'' by making it undesirable to agent n. I want to show that the new system $(\overline{x}, \overline{\sigma})$ is still incentive compatible, renegotiation proof and leaves the expected utility of the principal unchanged.

Let $B_n^{t_n}(x, \sigma)$ denote the expected punishment of agent n, type t_n, under system (x, σ).

Notice that

$$\begin{aligned} & B_n^{t_n}\left(\overline{x}, \overline{\sigma}\right) \\ = & \frac{\sigma_n\left(t_n, m_n'\right)}{\sigma_n\left(t_n, m_n'\right) + \sigma_n\left(t_n, m_n''\right)} B_n^{t_n}(x, \sigma) + \frac{\sigma_n\left(t_n, m_n''\right)}{\sigma_n\left(t_n, m_n'\right) + \sigma_n\left(t_n, m_n''\right)} B_n^{t_n}(x, \sigma) \\ = & B_n^{t_n}(x, \sigma) \end{aligned}$$

for $t_n = i, g$.

As for $\widehat{n} \neq n$, notice that we can write,

$$B_{\widehat{n}}^{t_{\widehat{n}}}\left(\overline{x}, \overline{\sigma}\right) = \int\limits_{\theta \in \Theta} \int\limits_{m_{-\widehat{n}} \in M_{-\widehat{n}}} \pi^{\overline{\sigma}}\left(m_{-\widehat{n}}, \theta | t_{\widehat{n}}\right) \overline{x}_n\left(m_{\widehat{n}}, m_{-\widehat{n}}, \theta\right) dm_{-\widehat{n}} d\theta$$

for some $m_{\widehat{n}}$ such that $\sigma_{\widehat{n}}\left(t_{\widehat{n}}, m_{\widehat{n}}\right) > 0$. Notice also that $\pi^{\overline{\sigma}}\left(m_n', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\right)$ is equal to

$$\sum_{t_n \in \{i,g\}} \left[\overline{\sigma}_n\left(t_n, m_n'\right) \pi\left(\theta_n | t_n\right) \pi\left(\theta_{\widehat{n}} | t_{\widehat{n}}\right) \sum_{t_{-\widehat{n},n}} \pi\left(t_{\widehat{n}}, t_n, t_{-\widehat{n},n} | t_{\widehat{n}}\right) \prod_{\widetilde{n} \neq n, \widehat{n}} \pi\left(\theta_{\widetilde{n}} | t_{\widetilde{n}}\right) \sigma_{\widetilde{n}}\left(m_{\widetilde{n}}, t_{\widetilde{n}}\right) \right]$$

Given that

$$\begin{aligned} & \begin{pmatrix} \pi^{\overline{\sigma}}\left(m_n', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\right) \overline{x}_n\left(m_{\widehat{n}}, m_n', m_{-\widehat{n},n}, \theta\right) \\ +\pi^{\overline{\sigma}}\left(m_n'', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\right) \overline{x}_n\left(m_{\widehat{n}}, m_n'', m_{-\widehat{n},n}, \theta\right) \end{pmatrix} \\ = & \begin{pmatrix} \pi^{\sigma}\left(m_n', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\right) x_n\left(m_{\widehat{n}}, m_n', m_{-\widehat{n},n}, \theta\right) \\ +\pi^{\sigma}\left(m_n'', m_{-\widehat{n},n}, \theta | t_{\widehat{n}}\right) x_n\left(m_{\widehat{n}}, m_n'', m_{-\widehat{n},n}, \theta\right) \end{pmatrix} \end{aligned}$$

it follows that $B_{\widehat{n}}^{t_{\widehat{n}}}\left(\overline{x}, \overline{\sigma}\right) = B_{\widehat{n}}^{t_{\widehat{n}}}\left(x, \sigma\right)$ for all $t_{\widehat{n}}$ and for all $\widehat{n} \neq n$, which implies that $V\left(\overline{x}, \overline{\sigma}\right) = V\left(x, \sigma\right)$.

The system $\left(\overline{x}, \overline{\sigma}\right)$ is incentive compatible as sending message m_n'' is not strictly preferred to any other message and the expected punishment of sending any other message remains unchanged. It is also renegotiation proof because, for all m_{-n}, θ and for all $\widehat{n}$ (including n)

$$\begin{aligned} & \overline{x}_{\widehat{n}}\left(m_n', m_{-n}, \theta\right) \\ \leq\ & \max\left\{x_{\widehat{n}}\left(m_n', m_{-n}, \theta\right), x_{\widehat{n}}\left(m_n'', m_{-n}, \theta\right)\right\} \\ \leq\ & \gamma_{\widehat{n}}^{\sigma}\left(m_n', m_{-n}, \theta\right) \\ =\ & \gamma_{\widehat{n}}^{\overline{\sigma}}\left(m_n', m_{-n}, \theta\right) \end{aligned}$$

9.2.5 Appendix B5

I show that, for all σ, (x^{σ}, σ) is incentive compatible and renegotiation proof. Notice that all non-confessing reports involve the same punishment, which means that agents are indifferent between sending any non-confessing message. By the definition of φ_n, guilty agents are indifferent between confessing and not confessing. Hence, it is only necessary to show that innocent agents do not strictly prefer to confess, which is equivalent to showing that the innocent's expected punishment of sending message m_n^{σ} is smaller or equal than the guilty's expected punishment of sending message m_n^{σ}.

Notice that it is possible to write

$$x_n^{\sigma}\left(m_n^{\sigma}, m_{-n}, \theta\right) = \begin{cases} 1 \text{ if } \alpha \frac{\sigma_n(i, m_n^{\sigma})}{\sigma_n(g, m_n^{\sigma})} \frac{\pi(t_n=i)}{\pi(t_n=g)} \frac{\pi(m_{-n}, \theta | t_n=i)}{\pi(m_{-n}, \theta | t_n=g)} < 1 \\ 0 \text{ otherwise} \end{cases}$$

Define

$$E_n \equiv \left\{(m_{-n}, \theta) \in M_{-n} \times [0,1]^N : x_n^{\sigma}\left(m_n^{\sigma}, m_{-n}, \theta\right) = 1\right\}$$

If $E_n = \varnothing$ or $\rceil E_n = \varnothing$, then the expected punishment of the agent when sending message m_n^{σ} is independent of his type.

If $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)} < 1$ for all $e_n \in E_n$ then

$$\int\limits_{e_n \in E_n} \pi\left(e_n | t_n = g\right) de_n > \int\limits_{e_n \in E_n} \pi\left(e_n | t_n = i\right) de_n$$

and so the expected punishment of the agent when sending message m_n^{σ} is larger if he is guilty.

Finally, if there is $e_n' \in E_n$ such that $\frac{\pi(e_n'|t_n=i)}{\pi(e_n'|t_n=g)} \geq 1$ and given that $x_n^{\sigma}\left(m_n^{\sigma}, m_{-n}, \theta\right)$ is decreasing with $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)}$, then it must be that $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)} > 1$ for all $e_n \notin E_n$. Hence,

$$\int\limits_{e_n \notin E_n} \pi\left(e_n | t_n = g\right) de_n < \int\limits_{e_n \notin E_n} \pi\left(e_n | t_n = i\right) de_n$$

which implies that

$$\int_{e_n \in E_n} \pi\left(e_n | t_n = g\right) de_n > \int_{e_n \in E_n} \pi\left(e_n | t_n = i\right) de_n$$

and so, also in this case, the expected punishment of the agent when sending message m_n^σ is larger if he is guilty. Hence, it follows that the system (x^σ, σ) is incentive compatible.

To guarantee that the system is renegotiation proof, I set the beliefs after any message that is not sent in equilibrium to be as if the message sent was m_n^σ, except for message c, where the agent is always believed to be guilty with certainty. Hence, it follows that the system is renegotiation proof because $\gamma_n^\sigma\left(m_n^\sigma, m_{-n}, \theta\right) \leq \gamma_n^\sigma\left(m_n, m_{-n}, \theta\right)$ for all $m_n \in \mathbb{R}_+$.

9.2.6 Appendix B6

Let $\widetilde{V}(\sigma, \pi)$ denote the expected utility of the principal under prior distribution π and given strategy profile σ, when $x = x^\sigma$. Additionally, define

$$\sigma^*(\pi) \in \arg\max_{\sigma \in \Phi} \widetilde{V}(\sigma, \pi)$$

and notice that, because $\widetilde{V}$ is continuous, $\sigma^*(\pi)$ is continuous.

Fix any $(\xi_1, ..., \xi_N)$ such that $\sum_{n=1}^{N} \xi_n < 1$, any $\gamma > 0$ and let $\overline{\pi}$ be such that $\overline{\pi}(t) = 0$ for all $t \notin \widehat{T}$. It follows that there is $\delta \in (0, 1)$ such that, if $\sum_{t \notin \widehat{T}} \pi(t) \leq \delta$,

$$\widetilde{V}(\sigma^*(\pi), \pi) - \widetilde{V}(\sigma^*(\overline{\pi}), \pi) \leq \gamma$$

The proof is complete by showing that $\sigma^*(\overline{\pi})$ is a CIS.

I claim that, if $\pi = \overline{\pi}$, there is a CIS which is an optimal incentive compatible renegotiation proof system. Fix some optimal system $\left(x^{RP}, \sigma^{RP}\right)$. Notice that Lemmas 13 and 14 still hold when $\pi = \overline{\pi}$. Therefore, without loss of generality (because of Lemma 14), assume that

$$x_n^{RP}\left(m_n, m_{-n}, \theta\right) = \gamma_n^{\sigma^{RP}}\left(m_n^{\sigma^{RP}}, m_{-n}, \theta\right)$$

for any $(m_n, m_{-n}) \in M_n \times M_{-n}$ sent with positive probability, and for any θ and n.

Consider an alternative system $(\widehat{x}, \widehat{\sigma})$ where the message set of each agent is given by $\widehat{M_n} = M_n \times \{0, 1\}$. System $(\widehat{x}, \widehat{\sigma})$ is as follows:

i)

$$\widehat{x}_n\left(\widehat{m}_n, \widehat{m}_n, \theta\right) = \gamma_n^{\widehat{\sigma}}\left(\widehat{m}_n^{\widehat{\sigma}}, \widehat{m}_{-n}, \theta\right) \text{ for all } \left(\widehat{m}_n, \widehat{m}_{-n}, \theta\right) \text{ and } n$$

ii)

$$\widehat{\sigma}_n\left(i, (m_n, a)\right) = \begin{cases} \sigma_n^{RP}\left(i, m_n\right) \text{ if } a = 0 \\ 0 \text{ if } a = 1 \end{cases} \quad \text{for all } m_n \in M_n \text{ and for all } n$$

iii)

$$\widehat{\sigma}_n\left(g, (m_n, a)\right) = \begin{cases} \sigma_n^{RP}\left(g, m_n\right) - v_n\left(m_n\right) \text{ if } a = 0 \text{ and } m_n \neq c \\ \sigma_n^{RP}\left(g, m_n\right) \text{ if } a = 0 \text{ and } m_n = c \\ v_n\left(m_n\right) \text{ if } a = 1 \text{ and } m_n \neq c \\ 0 \text{ if } a = 1 \text{ and } m_n = c \end{cases}$$

where

$$v_n\left(m_n\right) : \frac{\sigma_n^{RP}\left(g, m_n\right) - v_n\left(m_n\right)}{\sigma_n^{RP}\left(g, m_n\right)} = \frac{\sigma_n^{RP}\left(g, m_n^{\sigma^{RP}}\right)}{\sigma_n^{RP}\left(g, m_n^{\sigma^{RP}}\right)}$$

Basically, in system $(\widehat{x}, \widehat{\sigma})$, the probability that the guilty type of agent n sends non-confessing message m_n is reduced by $v_n(m_n)$, which is chosen so that the posterior belief after message m_n is the same as after message $m_n^{\sigma^{RP}}$. That extra weight is then allocated to message $(m_n, 1)$, which is essentially a confession because only the guilty type sends it. Notice that, in system $(\widehat{x}, \widehat{\sigma})$, all non-confessing messages induce the same posterior and, therefore, could be merged. The same happens for all confessing messages (c and all $(m_n, 1)$ messages). As a result, the expected utility for the principal of system $(\widehat{x}, \widehat{\sigma})$ is equivalent to that of the corresponding CIS. It is convenient for exposition not to actually merge the messages.

I show that

$$V\left(x^{RP},\sigma^{RP}\right)\leq V\left(\widehat{x},\widehat{\sigma}\right)$$

which completes the proof. WLOG, I focus on the impact on the expected utility the principal gets from agent 1 and show that

$$V_1\left(x^{RP},\sigma^{RP}\right)-V_1\left(\widehat{x},\widehat{\sigma}\right)\leq 0$$

Notice that

$$V_1\left(x^{RP},\sigma^{RP}\right)=\int\limits_{(m_{-1},\theta)\in M_{-1}\times\Theta}\chi^{RP}\left(m_{-1},\theta\right)d\left(m_{-1},\theta\right)$$

where

$$\begin{aligned}\chi^{RP}\left(m_{-1},\theta\right) &= \max_{x\in\left[0,\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}},m_{-1},\theta\right)\right]}\left\{\begin{pmatrix}\pi^{\sigma^{RP}}\left(t_1=g,m_{-1},\theta\right)\\-\alpha\pi^{\sigma^{RP}}\left(t_1=i,m_{-1},\theta\right)\end{pmatrix}x\right\}\\ &= \begin{pmatrix}\pi^{\sigma^{RP}}\left(t_1=g,m_{-1},\theta\right)\\-\alpha\pi^{\sigma^{RP}}\left(t_1=i,m_{-1},\theta\right)\end{pmatrix}\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}},m_{-1},\theta\right)\end{aligned}$$

By definition of $(\widehat{x},\widehat{\sigma})$,

$$\chi^{RP}\left(m_{-1},\theta\right)=\left[\begin{array}{c}\begin{pmatrix}\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},0_{-1}\right),\theta\right)\\-\alpha\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},0_{-1}\right),\theta\right)\end{pmatrix}\\+\sum\limits_{a_{-1}\neq 0_{-1}}\begin{pmatrix}\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},a_{-1}\right),\theta\right)\\-\alpha\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},a_{-1}\right),\theta\right)\end{pmatrix}\end{array}\right]\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}},m_{-1},\theta\right)$$

where

$$\left(m_{-1},0_{-1}\right)=\left(\left(m_2,0\right),...,\left(m_N,0\right)\right)$$

and

$$\left(m_{-1},a_{-1}\right)=\left(\left(m_2,a_2\right),...,\left(m_N,a_N\right)\right)$$

Given that

$$\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},a_{-1}\right),\theta\right)=0\text{ for any }a_{-1}\neq 0_{-1}$$

it follows that

$$\chi^{RP}\left(m_{-1},\theta\right)=\left[\begin{array}{c}\left(\begin{array}{c}\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},0_{-1}\right),\theta\right)\\-\alpha\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},0_{-1}\right),\theta\right)\end{array}\right)\\-\alpha\sum\limits_{a_{-1}\neq 0_{-1}}\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},a_{-1}\right),\theta\right)\end{array}\right]\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}},m_{-1},\theta\right)$$

Notice also that

$$V_1\left(\widehat{x},\widehat{\sigma}\right)=\int\limits_{\left(m_{-1},\theta\right)\in M_{-1}\times\Theta}\widehat{\chi}\left(m_{-1},\theta\right)d\left(m_{-1},\theta\right)$$

where

$$\begin{aligned}\widehat{\chi}\left(m_{-1},\theta\right) &= \left\{\begin{array}{c}\max_{x\in\left[0,\gamma_1^{\widehat{\sigma}}\left(\left(m_1^{\sigma^{RP}},0\right),\left(m_{-1},0_{-1}\right),\theta\right)\right]}\left\{\left(\begin{array}{c}\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},0_{-1}\right),\theta\right)\\-\alpha\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},0_{-1}\right),\theta\right)\end{array}\right)x\right\}+\\-\alpha\sum\limits_{a_{-1}\neq 0_{-1}}\max_{x\in\left[0,\gamma_1^{\widehat{\sigma}}\left(\left(m_1^{\sigma^{RP}},0\right),\left(m_{-1},a_{-1}\right),\theta\right)\right]}\left\{\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},a_{-1}\right),\theta\right)x\right\}\end{array}\right\}\\ &= \max_{x\in\left[0,\gamma_1^{\widehat{\sigma}}\left(\left(m_1^{\sigma^{RP}},0\right),\left(m_{-1},0_{-1}\right),\theta\right)\right]}\left\{\left(\begin{array}{c}\pi^{\widehat{\sigma}}\left(t_1=g,\left(m_{-1},0_{-1}\right),\theta\right)\\-\alpha\pi^{\widehat{\sigma}}\left(t_1=i,\left(m_{-1},0_{-1}\right),\theta\right)\end{array}\right)x\right\}\end{aligned}$$

so that

$$V_1\left(x^{RP},\sigma^{RP}\right)-V_1\left(\widehat{x},\widehat{\sigma}\right)=\int\limits_{\left(m_{-1},\theta\right)\in M_{-1}\times\Theta}\left(\chi^{RP}\left(m_{-1},\theta\right)-\widehat{\chi}\left(m_{-1},\theta\right)\right)d\left(m_{-1},\theta\right)$$

Notice that, by definition, $\widehat{\chi}\left(m_{-1},\theta\right)\geq 0$ for all $\left(m_{-1},\theta\right)\in M_{-1}\times\Theta$. As a result, it follows that

$$\begin{aligned}&V_1\left(x^{RP},\sigma^{RP}\right)-V_1\left(\widehat{x},\widehat{\sigma}\right)\\ \leq\ &\int\limits_{\left(m_{-1},\theta\right):\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}},m_{-1},\theta\right)=1}\left(\chi^{RP}\left(m_{-1},\theta\right)-\widehat{\chi}\left(m_{-1},\theta\right)\right)d\left(m_{-1},\theta\right)\end{aligned}$$

Take any $(m_{-1},\theta) \in M_{-1} \times \Theta$ such that

$$\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}}, m_{-1}, \theta\right) = 1$$

This implies that

$$\begin{aligned}
& \sum_{t_{-1}} \pi\left(g, t_{-1}\right) l\left(\theta_1\right) \frac{\sigma_1^{RP}\left(g, m_1^{\sigma^{RP}}\right)}{\sigma_1^{RP}\left(i, m_1^{\sigma^{RP}}\right)} \prod_{n \neq 1} \pi\left(\theta_n | t_n\right) \sigma_n^{RP}\left(t_n, m_n\right) \\
\geq \quad & \alpha \sum_{t_{-1}} \pi\left(i, t_{-1}\right) \prod_{n \neq 1} \pi\left(\theta_n | t_n\right) \sigma_n^{RP}\left(t_n, m_n\right)
\end{aligned}$$

which can be written as

$$\xi_1 l\left(\theta_1\right) \frac{\sigma_1^{RP}\left(g, m_1^{\sigma^{RP}}\right)}{\sigma_1^{RP}\left(i, m_1^{\sigma^{RP}}\right)} \geq \alpha \left[\sum_{\widehat{n} \neq 1}^{N} \xi_{\widehat{n}} l\left(\theta_{\widehat{n}}\right) \frac{\sigma_{\widehat{n}}^{RP}\left(g, m_{\widehat{n}}\right)}{\sigma_{\widehat{n}}^{RP}\left(i, m_{\widehat{n}}\right)} + \left(1 - \sum_{n=1}^{N} \xi_n\right)\right]$$

which implies that

$$\xi_1 l\left(\theta_1\right) \frac{\sigma_1^{RP}\left(g, m_1^{\sigma^{RP}}\right)}{\sigma_1^{RP}\left(i, m_1^{\sigma^{RP}}\right)} \geq \alpha \left[\sum_{\widehat{n} \neq 1}^{N} \xi_{\widehat{n}} l\left(\theta_{\widehat{n}}\right) \frac{\sigma_{\widehat{n}}^{RP}\left(g, m_{\widehat{n}}\right) - v_n\left(m_{\widehat{n}}\right)}{\sigma_{\widehat{n}}^{RP}\left(i, m_{\widehat{n}}\right)} + \left(1 - \sum_{n=1}^{N} \xi_n\right)\right]$$

Therefore,

$$\gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}}, m_{-1}, \theta\right) = 1 \Rightarrow \gamma_1^{\sigma^{RP}}\left(m_1^{\sigma^{RP}}, m_{-1}^{\sigma^{RP}}, \theta\right) = 1$$

where

$$m_{-1}^{\sigma^{RP}} = \left(m_2^{\sigma^{RP}}, ..., m_N^{\sigma^{RP}}\right)$$

so that

$$\begin{aligned}
\widehat{\chi}\left(m_{-1}, \theta\right) &= \max_{x \in [0,1]} \left\{ \begin{pmatrix} \pi^{\widehat{\sigma}}\left(t_1 = g, \left(m_{-1}, 0_{-1}\right), \theta\right) \\ -\alpha \pi^{\widehat{\sigma}}\left(t_1 = i, \left(m_{-1}, 0_{-1}\right), \theta\right) \end{pmatrix} x \right\} \\
&\geq \begin{pmatrix} \pi^{\widehat{\sigma}}\left(t_1 = g, \left(m_{-1}, 0_{-1}\right), \theta\right) \\ -\alpha \pi^{\widehat{\sigma}}\left(t_1 = i, \left(m_{-1}, 0_{-1}\right), \theta\right) \end{pmatrix} \geq \chi^{RP}\left(m_{-1}, \theta\right)
\end{aligned}$$

which implies that

$$V_1\left(x^{RP}, \sigma^{RP}\right) - V_1\left(\widehat{x}, \widehat{\sigma}\right) \leq 0$$

9.2.7 Appendix B7

Consider a CIS $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$ and let $\tau \in [0,1]^N$ be such that $\sigma_n^{CIS}(g,c) = \tau_n$ for all n. Also, let $\overline{V}(\tau)$ denote the corresponding expected utility of the principal. System $\left(x^{\sigma^{CIS}}, \sigma^{CIS}\right)$ is a trial system if and only if $\tau = \underline{\tau} \equiv (0, ..., 0)$.

I show the statement by showing that

$$\frac{\partial \overline{V}_n}{\partial \tau_n}(\underline{\tau}) = 0 \text{ for all } n \tag{15}$$

and

$$\frac{\partial \overline{V}_{\widehat{n}}}{\partial \tau_n}(\underline{\tau}) \geq 0 \text{ for all } n \text{ and } \widehat{n} \tag{16}$$

with the inequality being strict for at least one pair $(\widehat{n}, n)$, unless the types of the agents are independent.

In a CIS, if agent n refuses to confess, his punishment will be given by

$$x_n^{CIS}(\overline{c}, m_{-n}, \theta_{-n}) = \begin{cases} 1 \text{ if } \theta_n > \theta_n^{CIS}(m_{-n}, \theta_{-n}) \\ 0 \text{ otherwise} \end{cases}$$

where threshold $\theta_n^{CIS}(m_{-n}, \theta_{-n})$ is chosen to maximize the principal expected utility, conditional on message $m = (\overline{c}, m_{-n})$ and evidence θ. Therefore, it is possible to write $\overline{V}(\tau) = \sum_{n=1}^{N} \overline{V}_n(\tau)$ where

$$\overline{V}_n(\tau) = \int_{m_{-n} \in M_{-n}} \int_{\theta_{-n} \in \Theta_{-n}} \int_{\theta_n^{CIS}(m_{-n}, \theta_{-n})}^{1} A^{CIS}(m_{-n}, \theta_n, \theta_{-n})\, d\theta_n d\theta_{-n} dm_{-n}$$

where

$$\begin{aligned} & A^{CIS}(m_{-n}, \theta_n, \theta_{-n}) \\ = & \sum_{t_{-n} \in T_{-n}} \left(\begin{array}{c} \pi(g, t_{-n})\, \pi(\theta_n | t_n = g) - \\ \alpha \pi(i, t_{-n})\, \pi(\theta_n | t_n = i) \end{array} \right) \prod_{\widetilde{n} \neq n} \pi(\theta_{\widetilde{n}} | t_{\widetilde{n}})\, \sigma_{\widetilde{n}}^{CIS}(t_{\widetilde{n}}, m_{\widetilde{n}}) \end{aligned}$$

and

$$\theta_n^{CIS}\left(m_{-n},\theta_{-n}\right)=l^{-1}\left(\frac{\alpha}{1-\tau_n}\frac{\sum\limits_{t_{-n}\in T_{-n}}\pi\left(i,t_{-n}\right)\prod\limits_{\widetilde{n}\neq n}\pi\left(\theta_{\widetilde{n}}|t_{\widetilde{n}}\right)\sigma_{\widetilde{n}}^{CIS}\left(t_{\widetilde{n}},m_{\widetilde{n}}\right)}{\sum\limits_{t_{-n}\in T_{-n}}\pi\left(g,t_{-n}\right)\prod\limits_{\widetilde{n}\neq n}\pi\left(\theta_{\widetilde{n}}|t_{\widetilde{n}}\right)\sigma_{\widetilde{n}}^{CIS}\left(t_{\widetilde{n}},m_{\widetilde{n}}\right)}\right)$$

Notice that $\frac{\partial\overline{V}_n}{\partial\tau_n}$ is given by

$$-\int\limits_{m_{-n}\in M_{-n}}\int\limits_{\theta_{-n}\in\Theta_{-n}}A^{CIS}\left(m_{-n},\theta_n^{CIS}\left(m_{-n},\theta_{-n}\right),\theta_{-n}\right)\frac{d\theta_n^{CIS}\left(m_{-n},\theta_{-n}\right)}{d\tau_n}d\theta_{-n}dm_{-n}$$

Given that, when $\tau_n=0$,

$$A^{CIS}\left(m_{-n},\theta_n^{CIS}\left(m_{-n},\theta_{-n}\right),\theta_{-n}\right)=0$$

it must be that $\frac{\partial\overline{V}_n}{\partial\tau_n}\left(\underline{\tau}\right)=0$, which shows condition (15).

Now, consider (16). Notice that one can write $\overline{V}_{\widehat{n}}\left(\tau\right)$ as

$$\int\limits_{m_{-\widehat{n},n}\in M_{-\widehat{n},n}}\int\limits_{\theta_{-\widehat{n}}\in\Theta_{-\widehat{n}}}\left[\begin{array}{c}\int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{1}A^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right)d\theta_{\widehat{n}}\\+\int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{1}A^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right)d\theta_{\widehat{n}}\end{array}\right]d\theta_{-\widehat{n}}dm_{-\widehat{n},n}$$

Therefore, $\frac{\partial \overline{V}_n}{\partial \tau_{\widehat{n}}}$ is equal to

$$\int\limits_{m_{-\widehat{n},n}\in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}}\in\Theta_{-\widehat{n}}} \left[\begin{array}{c} -A^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}\right) * \\ \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}{d\tau_n} \\ -A^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}\right) * \\ \frac{d\theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}{d\tau_n} \end{array}\right] d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \tag{17}$$

$$+ \int\limits_{m_{-\widehat{n},n}\in M_{-\widehat{n},n}} \int\limits_{\theta_{-\widehat{n}}\in\Theta_{-\widehat{n}}} \left[\begin{array}{l} \int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{\widehat{n}}, \theta_{-\widehat{n}}\right)}{d\tau_n} d\theta_{\widehat{n}} \\ + \int\limits_{\theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)}^{1} \frac{dA^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{\widehat{n}}, \theta_{-\widehat{n}}\right)}{d\tau_n} d\theta_{\widehat{n}} \end{array}\right] d\theta_{-\widehat{n}} dm_{-\widehat{n},n} \tag{18}$$

Notice that (17) is equal to 0 when $\tau_{\widehat{n}} = 0$ given that

$$\begin{aligned} A^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}\right) &= \\ A^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right), \theta_{-\widehat{n}}\right) &= 0 \end{aligned}$$

so we are only left with analyzing (18).

Let

$$\underline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}} = \left\{\theta_{-\widehat{n}} \in \Theta_{-\widehat{n}} : \theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right) < \theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)\right\}$$

and

$$\overline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}} = \left\{\theta_{-\widehat{n}} \in \Theta_{-\widehat{n}} : \theta_{\widehat{n}}^{CIS}\left(m_n = c, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right) > \theta_{\widehat{n}}^{CIS}\left(m_n = \overline{c}, m_{-\widehat{n},n}, \theta_{-\widehat{n}}\right)\right\}$$

Then, condition (18) can be written as

$$\int_{m_{-\widehat{n},n}\in M_{-\widehat{n},n}} \int_{\theta_{-\widehat{n}}\in \underline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}} \int_{\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)} B^{CIS}\left(m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right) d\theta_{\widehat{n}} d\theta_{-\widehat{n}} dm_{-\widehat{n},n}$$

$$- \int_{m_{-\widehat{n},n}\in M_{-\widehat{n},n}} \int_{\theta_{-\widehat{n}}\in \overline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}} \int_{\theta_{\widehat{n}}^{CIS}\left(m_n=\overline{c},m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)}^{\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)} B^{CIS}\left(m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right) d\theta_{\widehat{n}} d\theta_{-\widehat{n}} dm_{-\widehat{n},n}$$

where $B^{CIS}\left(m_{-\widehat{n},n},\theta_{\widehat{n}},\theta_{-\widehat{n}}\right)$ is equal to

$$\sum_{t_{-\widehat{n},n}\in T_{-\widehat{n},n}} \left(\begin{array}{c} \pi\left(g,g,t_{-\widehat{n},n}\right)\pi\left(\theta_{\widehat{n}}|t_{\widehat{n}}=g\right)\pi\left(\theta_n|t_n=g\right)- \\ \alpha\pi\left(i,g,t_{-\widehat{n},n}\right)\pi\left(\theta_{\widehat{n}}|t_{\widehat{n}}=i\right)\pi\left(\theta_n|t_n=g\right) \end{array} \right) \prod_{\widetilde{n}\neq\widehat{n},n} \pi\left(\theta_{\widetilde{n}}|t_{\widetilde{n}}\right)\sigma_{\widetilde{n}}^{CIS}\left(t_{\widetilde{n}},m_{\widetilde{n}}\right)$$

which is strictly positive when $\tau_{\widehat{n}}=0$, given that

$$B^{CIS}\left(\theta_{\widehat{n}},\theta_{-\widehat{n}},m_{-\widehat{n},n}\right)>0 \text{ if and only if } \theta_n>\theta_{\widehat{n}}^{CIS}\left(m_n=c,m_{-\widehat{n},n},\theta_{-\widehat{n}}\right)$$

This implies that $\frac{\partial\overline{V}_n}{\partial\tau_{\widehat{n}}}\left(\underline{\tau}\right)>0$ unless $\underline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}$ and $\underline{\Theta}_{-\widehat{n}}^{m_{-\widehat{n},n}}$ are empty for all $m_{-\widehat{n},n}\in M_{-\widehat{n},n}$. But if that happens for all n, then the agents' types must be independent.

9.2.8 Appendix B8

An optimal mechanism which induces truthful reporting must maximize the principal's expected utility subject to the agents' incentive constraints. Unlike in the main text, there are many incentive constraints per agent as the number of extended types is now larger. My approach to solving this problem is to relax some of the incentive constraints and show that the solution of the relaxed problem satisfies the relaxed constraints. In particular, the relaxed problem is to select a mechanism $x:\widehat{T}\times\Theta\rightarrow[0,1]^N$ in order to maximize the principal's expected utility subject to the constraint that, for all n and for all $\widehat{t}_n\neq i$,

$$B_n^{\widehat{t}_n}\leq\int_{\theta\in\Theta}\sum_{\widehat{t}_{-n}\in\widehat{T}_{-n}}\pi\left(\widehat{t}_{-n},\theta|\widehat{t}_n\right)x_n\left(i,\widehat{t}_{-n},\theta\right)d\theta$$

Each constraint states that the guilty agent of extended type $\widehat{t}_n$ does not want to report to be innocent.

Notice that, by definition, any $\widehat{t} \in L$ does not enter the principal's expected utility function. Therefore, punishments that follow reports belonging to L should be chosen to minimize deviations which is achieved by setting them to 1.

A lot of the next steps are the same as in the main text. First, transform the problem into N independent problems. Second, all constraints must hold with equality for otherwise it would be possible to increase $B_n^{\widehat{t}_n}$ on the constraint that holds with strict inequality and make the strictly principal better off while still satisfying that constraint. This means that it is possible to write the problem solely in terms of the punishment that innocent agents receive. Guilty agents simply need to be made indifferent between reporting truthfully and reporting to be innocent. Hence, the new nth problem becomes one of selecting $x_n\left(i, \widehat{t}_{-n}, \theta\right) \in [0,1]$ for all $\widehat{t}_{-n} \in \widehat{T}_{-n}$ and $\theta \in \Theta$ in order to maximize

$$\int\limits_{\theta \in \Theta} \sum_{\widehat{t}_{-n} \in \widehat{T}_{-n}} \left(\sum_{\widehat{t}_n \neq i} \pi\left(\widehat{t}_n, \widehat{t}_{-n}, \theta\right) - \alpha \pi\left(i, \widehat{t}_{-n}, \theta\right) \right) x_n\left(i, \widehat{t}_{-n}, \theta\right) d\theta$$

which implies that it is optimal to select $x_n\left(i, \widehat{t}_{-n}, \theta\right) = \widehat{x}_n^{SB}\left(i, \widehat{t}_{-n}, \theta\right)$. By definition of $\widehat{x}_n^{SB}\left(i, \widehat{t}_{-n}, \theta\right)$ and for each $\widehat{t}_{-n}$ and θ there is $\overline{\theta}_n\left(\widehat{t}_{-n}\right) \in [0,1]$ such that

$$\widehat{x}_n^{SB}\left(i, \widehat{t}_{-n}, \theta\right) = \begin{cases} 1 \text{ if } \theta_n > \overline{\theta}_n\left(\widehat{t}_{-n}\right) \\ 0 \text{ otherwise} \end{cases}$$

Notice that $\overline{\theta}_n\left(\widehat{t}_{-n}\right)$ does not depend on θ_{-n} because it is not informative given the principal also knows $\widehat{t}_{-n}$.

In order to guarantee that guilty agents are indifferent to reporting to be innocent it is enough to set

$$\widehat{\varphi}_n\left(\widehat{t}_{-n}\right) = \int\limits_{\overline{\theta}_n\left(\widehat{t}_{-n}\right)}^{1} \pi\left(\theta | t_n = g\right) d\theta_n$$

so that, for all $\widehat{t}_n$,

$$B_n^{\widehat{t}_n} = \int\limits_{\theta\in\Theta} \sum_{\widehat{t}_{-n}\in\widehat{T}_{-n}} \pi\left(\widehat{t}_{-n},\theta|\widehat{t}_n\right)\widehat{x}_n^{SB}\left(i,\widehat{t}_{-n},\theta\right) = \sum_{\widehat{t}_{-n}\in\widehat{T}_{-n}} \pi\left(\widehat{t}_{-n}|\widehat{t}_n\right)\widehat{\varphi}_n\left(\widehat{t}_{-n}\right)$$

As for the relaxed incentive constraints it is easy to see that they are satisfied under allocation $\widehat{x}^{SB}$. In particular, the punishment a guilty agent receives is independent of his own report, which means that he has no strict incentive to deviate.

9.2.9 Appendix B9

The problem the principal faces is one of selecting $x_n(\theta) \in \mathbb{R}_+$ for all n and $\theta \in \Theta$ in order to maximize

$$\int\limits_{\theta\in\Theta} \left(\pi\left(t_n = i,\theta\right)u_n^p\left(i,x_n\left(\theta\right)\right) + \pi\left(t_n = g,\theta\right)u_n^p\left(g,x_n\left(\theta\right)\right)\right)d\theta$$

The derivative of the objective function with respect to $x_n(\theta)$ is given by

$$\pi\left(t_n = i,\theta\right)\frac{\partial u_n^p\left(i,x_n\left(\theta\right)\right)}{\partial x_n} + \pi\left(t_n = g,\theta\right)\frac{\partial u_n^p\left(g,x_n\left(\theta\right)\right)}{\partial x_n}$$

Given that both $u_n^p(i,\cdot)$ and $u_n^p(g,\cdot)$ are strictly concave and that

$$\pi\left(t_n = i,\theta\right)\frac{\partial u_n^p\left(i,0\right)}{\partial x_n} + \pi\left(t_n = g,\theta\right)\frac{\partial u_n^p\left(g,0\right)}{\partial x_n} > 0$$

it follows that $x_n^{Tr}(\theta)$ is such that

$$\pi\left(t_n = i,\theta\right)\frac{\partial u_n^p\left(i,x_n^{Tr}\left(\theta\right)\right)}{\partial x_n} + \pi\left(t_n = g,\theta\right)\frac{\partial u_n^p\left(g,x_n^{Tr}\left(\theta\right)\right)}{\partial x_n} = 0$$

and so it is continuous. Notice that the previous equation can be rewritten as

$$\frac{\partial u_n^p\left(i,x_n^{Tr}\left(\theta\right)\right)}{\partial x_n} + \frac{\pi\left(t_n = g\right)}{\pi\left(t_n = i\right)}\frac{\pi\left(\theta_{-n}|t_n = g\right)}{\pi\left(\theta_{-n}|t_n = i\right)}l\left(\theta_n\right)\frac{\partial u_n^p\left(g,x_n^{Tr}\left(\theta\right)\right)}{\partial x_n} = 0$$

Given that $l(\theta_n)$ is strictly increasing it follows that $x_n^{Tr}(\theta)$ is strictly increasing. Furthermore, given that $\lim_{\theta_n \to 0} l(\theta_n) = 0$ it must be that, for all $\theta_{-n} \in \Theta_{-n}$, $\lim_{\theta_n \to 0} x_n^{Tr}((\theta_n, \theta_{-n})) = 0$ and given that $\lim_{\theta_n \to 1} l(\theta_n) = \infty$ it must be that, for all $\theta_{-n} \in \Theta_{-n}$, $\lim_{\theta_n \to 1} x_n^{Tr}((\theta_n, \theta_{-n})) = 1$.

9.2.10 Appendix B10

If $\alpha u^i(x_n) = u_n^p(i, x_n)$ the innocent's incentive constraints do not bind for the same reason as in the main text. Hence, the problem becomes one of maximizing

$$\sum_{t_{-n} \in T_{-n}} \int_{\theta \in \Theta} \pi(g, t_{-n}, \theta) \, u_n^p(g, x_n(g, t_{-n}, \theta)) + \alpha \pi(i, t_{-n}, \theta) \, u^i(i, x_n(i, t_{-n}, \theta)) \, d\theta$$

subject

$$\sum_{t_{-n} \in T_{-n}} \int_{\theta} \pi(g, t_{-n}, \theta) \, u^g(x_n(g, t_{-n}, \theta)) \, d\theta \geq \sum_{t_{-n} \in T_{-n}} \int_{\theta} \pi(g, t_{-n}, \theta) \, u^g(x_n(i, t_{-n}, \theta)) \, d\theta$$

where the constraint must bind for otherwise the first best solution would be incentive compatible. The first order condition with respect to $x_n(g, t_{-n}, \theta)$ can be written as

$$\pi(g, t_{-n}, \theta) \frac{\partial u_n^p(g, x_n(g, t_{-n}, \theta))}{\partial x_n} + \lambda_n \pi(g, t_{-n}, \theta) \frac{\partial u^g(g, x_n(g, t_{-n}, \theta))}{\partial x_n} = \zeta_n^g(t_{-n}, \theta) - \eta_n^g(t_{-n}, \theta)$$

where $\lambda_n > 0$ denotes the lagrange multiplier associated with the constraint above, while $\zeta_n^g(t_{-n}, \theta) \geq 0$ and $\eta_n^g(t_{-n}, \theta) \geq 0$ denote the lagrange multiplier associated with $\{x_n(g, t_{-n}, \theta) \geq 0\}$ and $\{x_n(g, t_{-n}, \theta) \leq \phi\}$.

Given that

$$\frac{\partial^2 u_n^p(g, \cdot)}{\partial (x_n)^2} + \lambda_n \frac{\partial^2 u^g(g, \cdot)}{\partial (x_n)^2} < 0$$

and

$$\frac{\partial u_n^p(g, 1)}{\partial x_n} + \lambda_n \frac{\partial u^g(g, 1)}{\partial x_n} < 0$$

and

$$\frac{\partial u_n^p(g, 0)}{\partial x_n} + \lambda_n \frac{\partial u^g(g, 0)}{\partial x_n} > 0$$

it follows that $\widetilde{x}_n^{SB}(g, t_{-n}, \theta)$ uniquely solves

$$\frac{\partial u_n^p\left(g, \widetilde{x}_n^{SB}(g, t_{-n}, \theta)\right)}{\partial x_n} + \lambda_n \frac{\partial u^g\left(g, \widetilde{x}_n^{SB}(g, t_{-n}, \theta)\right)}{\partial x_n} = 0$$

Hence, $\widetilde{x}_n^{SB}(g, t_{-n}, \theta)$ is independent of t_{-n} and θ and must be equal to

$$\sum_{t_{-n} \in T_{-n}} \int_{\theta} \frac{\pi(g, t_{-n}, \theta)}{\pi(t_n = g)} u^g\left(\widetilde{x}_n^{SB}(i, t_{-n}, \theta)\right) d\theta$$

because the incentive constraint binds.

9.2.11 Appendix B11

The first order condition with respect to $x_n(i, t_{-n}, \theta)$ (of the problem described in Appendix B10) is given by

$$\alpha\pi(i, t_{-n}, \theta) \frac{\partial u^i(x_n(i, t_{-n}, \theta))}{\partial x_n} - \lambda_n \pi(g, t_{-n}, \theta) \frac{\partial u^g(x_n(i, t_{-n}, \theta))}{\partial x_n} = \zeta_n^g(t_{-n}, \theta) - \eta_n^g(t_{-n}, \theta)$$

where λ_n, ζ_n^g and η_n^g are as in Appendix B10, which can be written as

$$\left(\begin{array}{c} -\alpha\pi(i, t_{-n})\, \pi(\theta_n | t_n = i)\, \omega_i\, (x_n(i, t_{-n}, \theta))^{\omega_i - 1} \\ +\lambda_n \pi(g, t_{-n})\, \pi(\theta_n | t_n = g)\, \omega_g\, (x_n(i, t_{-n}, \theta))^{\omega_g - 1} \end{array} \right) = \frac{\zeta_n^g(t_{-n}, \theta) - \eta_n^g(t_{-n}, \theta)}{\pi(\theta_{-n} | t_{-n})}$$

Let $\psi_n(t_{-n}, \theta_n)$ be the unique value of $x_n(i, t_{-n}, \theta)$ such that the left hand side is equal to 0, i.e.

$$\psi_n(t_{-n}, \theta_n) = \left(\frac{\lambda_n \omega_g}{\alpha \omega_i} \frac{\pi(g, t_{-n})}{\pi(i, t_{-n})} l(\theta_n) \right)^{\frac{1}{\omega_i - \omega_g}}$$

Notice that

$$\alpha\pi(i, t_{-n}, \theta) \frac{\partial^2 u^i(\psi_n(t_{-n}, \theta_n))}{\partial (x_n)^2} - \lambda_n \pi(g, t_{-n}, \theta) \frac{\partial^2 u^g(\psi_n(t_{-n}, \theta_n))}{\partial (x_n)^2}$$

is strictly negative if and only if $\omega_i > \omega_g$ in which case $\widetilde{x}_n^{SB}(i, t_{-n}, \theta) = \psi_n(t_{-n}, \theta_n)$ if $\psi_n(t_{-n}, \theta_n) \leq \phi$. Otherwise, $\widetilde{x}_n^{SB}(i, t_{-n}, \theta) = \phi$. It follows that $\widetilde{\theta}_n^{SB(i)}$ is such that

$\psi_n\left(t_{-n},\widetilde{\theta}_n^{SB(i)}\right)=\phi$. In particular, $\widetilde{\theta}_n^{SB(i)}$ is such that

$$\widetilde{\theta}_n^{SB(i)}=l^{-1}\left(\phi^{\omega_i-\omega_g}\frac{\alpha\omega_i}{\lambda_n\omega_g}\frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)}\right)$$

This shows i).

If $\omega_i\leq\omega_g$, then it follows that $\widetilde{x}_n^{SB}\left(i,t_{-n},\theta\right)$ is a corner and so it is either 0 or ϕ. In particular, it is ϕ if and only if

$$\alpha\pi\left(i,t_{-n},\theta\right)u^i\left(\phi\right)-\lambda_n\pi\left(g,t_{-n},\theta\right)u^g\left(\phi\right)>0$$

which implies that

$$\theta_n>l^{-1}\left(\frac{\alpha}{\lambda_n}\frac{\pi\left(i,t_{-n}\right)}{\pi\left(g,t_{-n}\right)}\phi^{\omega_i-\omega_g}\right)\equiv\widetilde{\theta}_n^{SB(g)}$$

Therefore, ii) follows.

The variable λ_n is such that

$$\widetilde{\varphi}_n=\sum_{t_{-n}\in T_{-n}}\int_{\theta}\frac{\pi\left(g,t_{-n},\theta\right)}{\pi\left(t_n=g\right)}u^g\left(\widetilde{x}_n^{SB}\left(i,t_{-n},\theta\right)\right)d\theta$$

holds where $\widetilde{\varphi}_n$ is such that

$$\frac{\partial u_n^p\left(g,\widetilde{\varphi}_n\right)}{\partial x_n}+\lambda_n\frac{\partial u^g\left(g,\widetilde{\varphi}_n\right)}{\partial x_n}=0$$

and $\widetilde{x}_n^{SB}\left(g,t_{-n},\theta\right)=\widetilde{\varphi}_n$.

9.2.12 Appendix B12

Suppose that the principal waits until he receives evidence θ and then makes a proposal $y_\theta:T\times\Theta\rightarrow R_+^N$ such that it is a Bayes-Nash equilibrium for all agents to tell the truth. Let $x_y:T\times\Theta$ denote the allocation which is implemented by the PBE: $x_y\left(t,\theta\right)=y_\theta\left(t,\theta\right)$ for all t and θ. I show that, if the principal proposed mechanism x_y before observing the evidence it would be incentive compatible.

Given each proposal y_θ and their type own type t_n, agents form some posterior belief about t and θ whose joint density I denote by $\pi^{y_\theta}(t,\theta|t_n)$. Given that agents prefer to report truthfully after each proposal y_θ, for all θ, it must be that, for all $\widehat{\theta}$, $t_n \in \{i,g\}$ and n, for all t_n',

$$-\sum_{t\in T}\int_{\theta\in\Theta}\pi^{y_{\widehat{\theta}}}(t,\theta|t_n)\,y_{\widehat{\theta}}(t_n,t_{-n},\theta)\,d\theta \geq -\sum_{t\in T}\int_{\in\Theta}\pi^{y_{\widehat{\theta}}}(t,\theta|t_n)\,y_{\widehat{\theta}}(t_n',t_{-n},\theta)\,d\theta$$

Given that the previous expression holds for all $\widehat{\theta}$, it follows that, for all t_n',

$$\begin{aligned}
&-\int_{\widehat{\theta}\in\Theta}\pi\left(\widehat{\theta}|t_n\right)\sum_{t\in T}\int_{\theta\in\Theta}\pi^{y_{\widehat{\theta}}}(t,\theta|t_n)\,y_{\widehat{\theta}}(t_n,t_{-n},\theta)\,d\theta d\widehat{\theta}\\
\geq\ &-\int_{\widehat{\theta}\in\Theta}\pi\left(\widehat{\theta}|t_n\right)\sum_{t\in T}\int_{\theta\in\Theta}\pi^{y_{\widehat{\theta}}}(t,\theta|t_n)\,y_{\widehat{\theta}}(t_n',t_{-n},\theta)\,d\theta d\widehat{\theta}
\end{aligned}$$

where $\pi\left(\widehat{\theta}|t_n\right)$ refers to the density of $\widehat{\theta}$ conditional of the agent's type t_n. Now, I want to group into disjoint sets the evidence that, given the strategy of the principal, induces the same posterior on the agent. More formally denote by $\chi_{\widehat{\theta}} \equiv \left\{\theta\in\Theta : y_\theta = y_{\widehat{\theta}}\right\}$ and $\widehat{\Theta} \equiv \left\{\widehat{\theta}\in\Theta : \text{for all } \theta \text{ such that } \pi_{\widehat{\theta}} = \pi_\theta \text{ then } \widehat{\theta} \prec_l \theta\right\}$ where $\prec_l$ denotes the lexicographic ordering.[20] Finally, let $\Upsilon = \left\{\chi_{\widehat{\theta}} \text{ for } \widehat{\theta}\in\widehat{\Theta}\right\}$. Notice that Υ represents a set of disjoint sets of $\widehat{\theta}$, where each set contains elements that induce the same posterior. It follows that the left hand side of the inequality above can be written as

$$\begin{aligned}
&-\sum_{t\in T}\int_{\chi_{\widehat{\theta}}\in\Upsilon}\pi\left(\theta\in\chi_{\widehat{\theta}}|t_n\right)\int_{\theta\in\chi_{\widehat{\theta}}}\pi\left(t,\theta|t_n,\theta\in\pi_{\widehat{\theta}}\right)x_y(t_n,t_{-n},\theta)\,d\theta d\chi_{\widehat{\theta}}\\
=\ &-\sum_{t\in T}\int_{\chi_{\widehat{\theta}}\in\Upsilon}\int_{\theta\in\chi_{\widehat{\theta}}}\pi(t,\theta|t_n)\,x_y(t_n,t_{-n},\theta)\,d\theta d\chi_{\widehat{\theta}}\\
=\ &-\sum_{t\in T}\int_{\theta}\pi(t,\theta|t_n)\,x_y(t_n,t_{-n},\theta)\,d\theta
\end{aligned}$$

[20] I could have used any other ordering. In fact, I only order the evidence for expositional convenience.

By following the same steps with the right hand side, it follows that x_y is incentive compatible, according to definition 5.

9.2.13 Appendix B13

Recall that message c represents the choice of confessing, while message $\overline{c}$ represents the choice of not confessing. I divide the proof into two lemmas.

Lemma B13.1 *For all n, there is $\left(\beta_n^g, \beta_n^i\right) \in [0,1]^N$ such that either A) for all (t_n, β_n),*

$$s_n\left(t_n, \beta_n\right) = \begin{cases} c \text{ if } \beta_n \geq \beta_n^{t_n} \\ \overline{c} \text{ otherwise} \end{cases}$$

or B) for all (t_n, β_n),

$$s_n\left(t_n, \beta_n\right) = \begin{cases} c \text{ if } \beta_n \leq \beta_n^{t_n} \\ \overline{c} \text{ otherwise} \end{cases}$$

Proof of Lemma B13.1 Let pair (t_n, β_n) denote the agent n's extended type. Notice that a CIS is determined by the pair (s, x) where $s = \left\{\{s_n\left(t_n, \beta_n\right)\}_{\beta_n \in [0,1]}\right\}_{t_n \in T_n}$ and $x : \{T_n \times [0,1]\}_{n=1}^N \times \Theta \rightarrow [0,1]$. For all n, let $B_n^{t_n}\left(\beta_n\right)$ denote the expected punishment that agent n receives if his extended type is (t_n, β_n). Divide the set of agent n's extended types into 6 smaller sets. In particular, for $t_n \in \{i, g\}$, let $\Gamma_{\overline{c}}^{t_n}$ denote the set of $\beta_n \in [0,1]$ such that the agent strictly prefers $\overline{c}$, $\Gamma_c^{t_n}$ denote the set of $\beta_n \in [0,1]$ such that the agent strictly prefers c and $\Gamma_{=}^{t_n}$ denote the set of $\beta_n \in [0,1]$ such that the agent is indifferent. Also, let $\beta = (\beta_1, .., \beta_N)$ and m_{-n} to be the set of actions (c or $\overline{c}$) that all other agents choose.

The principal chooses punishments in order to maximize the following objective function

$$\pi\left(t_n=g\right)\pi\left(\beta_n\in\Gamma^g_c\cup\Gamma^g_=|t_n=g\right)x_n\left(c\right)-\alpha\pi\left(t_n=i\right)\pi\left(\beta_n\in\Gamma^i_c\cup\Gamma^i_=|t_n=i\right)x_n\left(c\right)$$

$$+\int\limits_{\beta_n\in\Gamma^g_{\bar{c}}}\int\limits_{\beta_{-n}}\int\limits_{\theta}\sum_{t_{-n}}\sum_{m_{-n}}\pi\left(g,t_{-n}\right)\pi\left(\beta|t_n=g,t_{-n}\right)\pi\left(m_{-n},\theta|t_{-n},\beta\right)x_n\left(\bar{c},m_{-n},\theta\right)d\theta d\beta$$

$$-\alpha\int\limits_{\beta_n\in\Gamma^i_{\bar{c}}}\int\limits_{\beta_{-n}}\int\limits_{\theta}\sum_{t_{-n}}\sum_{m_{-n}}\pi\left(i,t_{-n}\right)\pi\left(\beta|t_n=i,t_{-n}\right)\pi\left(m_{-n},\theta|t_{-n},\beta\right)x_n\left(\bar{c},m_{-n},\theta\right)d\theta d\beta$$

subject to the respective incentive constraints - agents that choose message c prefer it to message $\bar{c}$ and vice-versa. Agents that are not indifferent have loose constraints - a slight change in the punishments still leaves them strictly preferring the same message. Hence, the only constraints that might bind are the ones of agents that are indifferent. In particular, it must be that, for all $\beta_n\in\Gamma^g_=$,

$$\begin{aligned}&x_n\left(c\right)\pi\left(t_n=g\right)\\=&\int\limits_{\beta_{-n}}\int\limits_{\theta_{-n}}\sum_{t_{-n}}\sum_{m_{-n}}\pi\left(g,t_{-n}\right)\pi\left(\beta_{-n}|t_{-n}\right)\pi\left(m_{-n},\theta|t_{-n},\beta\right)x_n\left(\bar{c},m_{-n},\theta\right)d\theta_{-n}d\beta_{-n}\end{aligned}$$

and for all $\beta_n\in\Gamma^i_=$,

$$\begin{aligned}&x_n\left(c\right)\pi\left(t_n=i\right)\\=&\int\limits_{\beta_{-n}}\int\limits_{\theta_{-n}}\sum_{t_{-n}}\sum_{m_{-n}}\pi\left(i,t_{-n}\right)\pi\left(\beta_{-n}|t_{-n}\right)\pi\left(m_{-n},\theta|t_{-n},\beta\right)x_n\left(\bar{c},m_{-n},\theta\right)d\theta_{-n}d\beta_{-n}\end{aligned}$$

For all $\beta_n\in\Gamma^g_=$ and $\beta_n\in\Gamma^i_=$ let $\lambda^g\left(\beta_n\right)$ and $\lambda^i\left(\beta_n\right)$ denote the lagrange multipliers of the conditions above respectively. Also, for all $\beta_n\in\Gamma^g_{\bar{c}}$ and $\beta_n\in\Gamma^i_{\bar{c}}$, write $\lambda^g\left(\beta_n\right)=\lambda^i\left(\beta_n\right)=1$.

For all m_{-n} and θ, the first order condition with respect to $x_n\left(\bar{c},m_{-n},\theta\right)$ is given by

$$\begin{aligned}
& \int\limits_{\beta_n \in \Gamma^g_{=} \cup \Gamma^g_{\bar{c}}} \pi\left(\beta_n | t_n = g\right) \lambda^g\left(\beta_n\right) \int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi\left(g, t_{-n}\right) \pi\left(\beta_{-n} | t_{-n}\right) \pi\left(m_{-n}, \theta | t_{-n}, \beta\right) d\beta \\
& -\alpha \int\limits_{\beta_n \in \Gamma^i_{=} \cup \Gamma^i_{\bar{c}}} \pi\left(\beta_n | t_n = i\right) \lambda^i\left(\beta_n\right) \int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi\left(i, t_{-n}\right) \pi\left(\beta_{-n} | t_{-n}\right) \pi\left(m_{-n}, \theta | t_{-n}, \beta\right) d\beta \\
= \; & \zeta^{\bar{c}}_n\left(m_{-n}, \theta\right) - \eta^{\bar{c}}_n\left(m_{-n}, \theta\right)
\end{aligned}$$

where $\zeta^{\bar{c}}_n\left(m_{-n}, \theta\right) \geq 0$ and $\eta^{\bar{c}}_n\left(m_{-n}, \theta\right) \geq 0$ denote the lagrange multipliers associated with constraints $\left\{x_n\left(\bar{c}, m_{-n}, \theta\right) \geq 0\right\}$ and $\left\{x_n\left(\bar{c}, m_{-n}, \theta\right) \leq 1\right\}$ respectively.

The left hand side (LHS) has the following property:

$$LHS \begin{cases} > 0 \text{ if } k\left(m_{-n}, \theta_{-n}\right) h\left(\theta_n\right) > 1 \\ = 0 \text{ if } k\left(m_{-n}, \theta_{-n}\right) h\left(\theta_n\right) = 1 \\ < 0 \text{ if } k\left(m_{-n}, \theta_{-n}\right) h\left(\theta_n\right) < 1 \end{cases}$$

where

$$h\left(\theta_n\right) = \frac{\displaystyle\int\limits_{\beta_n \in \Gamma^g_{=} \cup \Gamma^g_{\bar{c}}} \pi\left(\beta_n | t_n = g\right) \lambda^g\left(\beta_n\right) \pi\left(\theta_n | \beta_n\right) d\beta_n}{\displaystyle\int\limits_{\beta_n \in \Gamma^i_{=} \cup \Gamma^i_{\bar{c}}} \pi\left(\beta_n | t_n = i\right) \lambda^i\left(\beta_n\right) \pi\left(\theta_n | \beta_n\right) d\beta_n}$$

and

$$k\left(m_{-n}, \theta_{-n}\right) = \frac{\displaystyle\int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi\left(g, t_{-n}\right) \pi\left(\beta_{-n} | t_{-n}\right) \pi\left(\theta_{-n} | \beta_{-n}\right) \pi\left(m_{-n} | t_{-n}, \beta_{-n}\right) d\beta_{-n}}{\displaystyle\alpha \int\limits_{\beta_{-n}} \sum_{t_{-n}} \pi\left(i, t_{-n}\right) \pi\left(\beta_{-n} | t_{-n}\right) \pi\left(\theta_{-n} | \beta_{-n}\right) \pi\left(m_{-n} | t_{-n}, \beta_{-n}\right) d\beta_{-n}}$$

Notice that

$$h'\left(\theta_n\right) \begin{cases} > 0 \text{ if } A > B \\ = 0 \text{ if } A = B \\ < 0 \text{ if } A < B \end{cases}$$

where

$$A = \int\limits_{\beta_n \in \Gamma^g_{=} \cup \Gamma^g_{\overline{c}}} \pi\left(\beta_n | t_n = g\right) \lambda^g\left(\beta_n\right) \beta_n d\beta_n \int\limits_{\beta_n \in \Gamma^i_{=} \cup \Gamma^i_{\overline{c}}} \pi\left(\beta_n | t_n = i\right) \lambda^i\left(\beta_n\right)\left(1 - \beta_n\right) d\beta_n$$

and

$$B = \int\limits_{\beta_n \in \Gamma^i_{=} \cup \Gamma^i_{\overline{c}}} \pi\left(\beta_n | t_n = i\right) \lambda^i\left(\beta_n\right) \beta_n d\beta_n \int\limits_{\beta_n \in \Gamma^g_{=} \cup \Gamma^g_{\overline{c}}} \pi\left(\beta_n | t_n = g\right) \lambda^g\left(\beta_n\right)\left(1 - \beta_n\right) d\beta_n$$

Given that A and B are independent of θ_n, it follows that h is either a constant or strictly monotone. If it is a constant, then the punishment an agent receives is independent of the evidence he produces. In that case, an agent's β_n is irrelevant. Therefore, if this is the case, the statement follows with $\beta_n^{t_n}$ being either equal to 0 or 1. If it is strictly monotone it means that there is a strict ordering over β_n and so there is at most one indifferent β_n per type and the statement follows.

In the next lemma, I show that A) follows.

Lemma B13.2 For all n, there is $\left(\beta_n^g, \beta_n^i\right) \in [0,1]^N$ such that for all $\left(t_n, \beta_n\right)$,

$$s_n\left(t_n, \beta_n\right) = \begin{cases} c \text{ if } \beta_n \geq \beta_n^{t_n} \\ \overline{c} \text{ otherwise} \end{cases}$$

Proof of Lemma B13.2 Suppose not. Following the previous lemma, it must be that $h\left(\cdot\right)$ is strictly decreasing and

$$s_n\left(t_n, \beta_n\right) = \begin{cases} c \text{ if } \beta_n \leq \beta_n^{t_n} \\ \overline{c} \text{ otherwise} \end{cases}$$

This implies that

$$\frac{\int\limits_{\beta_n^i}^{1} \pi\left(\beta_n|t_n=i\right)\beta_n d\beta_n}{\int\limits_{\beta_n^i}^{1} \pi\left(\beta_n|t_n=i\right) d\beta_n} > \frac{\int\limits_{\beta_n^g}^{1} \pi\left(\beta_n|t_n=g\right)\beta_n d\beta_n}{\int\limits_{\beta_n^g}^{1} \pi\left(\beta_n|t_n=g\right) d\beta_n}$$

where, without loss of generality, $\lambda^{t_n}\left(\beta_n\right)=1$ for all t_n and β_n because there are only two pairs $\left(i,\beta_n^i\right)$ and $\left(g,\beta_n^g\right)$ that are indifferent and they have a 0 measure. Notice that if $\beta_n^i=\beta_n^g$ the condition does not hold because the right hand side is strictly larger. So it follows that $\beta_n^i>\beta_n^g$.

To complete the proof I show that an innocent agent with $\beta_n=\beta_n^g$ prefers to send message $\overline{c}$ (or is indifferent). I do this by showing that, for any fixed β_n, the expected punishment of choosing $\overline{c}$ is larger if the agent is guilty. Notice that

$$x_n\left(\overline{c},m_{-n},\theta\right)=\begin{cases} 1 \text{ if } \alpha\frac{\pi\left(t_n=i,\beta_n\geq\beta_n^i\right)}{\pi\left(t_n=g,\beta_n\geq\beta_n^g\right)}\frac{\pi\left(\theta_n|\beta_n\geq\beta_n^i\right)}{\pi\left(\theta_n|\beta_n\geq\beta_n^g\right)}\frac{\pi\left(m_{-n},\theta_{-n}|t_n=i\right)}{\pi\left(m_{-n},\theta_{-n}|t_n=g\right)}<1 \\ 0 \text{ otherwise}\end{cases}$$

and let $E_n^{\theta_n}=\left\{\left(m_{-n},\theta_{-n}\right):x_n\left(\overline{c},m_{-n},\theta_n,\theta_{-n}\right)=1\right\}$. Notice that the expected punishment of an agent of type $\left(t_n,\beta_n\right)$ of sending $\overline{c}$ is given by

$$\int\limits_{\theta_n\in[0,1]} \pi\left(\theta_n|\beta_n\right) \int\limits_{e_n\in E_n^{\theta_n}} \pi\left(e_n|t_n\right) de_n d\theta_n$$

Take any β_n and any θ_n. I want to show that

$$\int\limits_{e_n\in E_n^{\theta_n}} \pi\left(e_n|t_n=g\right) de_n \geq \int\limits_{e_n\in E_n^{\theta_n}} \pi\left(e_n|t_n=i\right) de_n$$

If $E_n^{\theta_n}=\varnothing$ or $E_n^{\theta_n}=\varnothing$ then the statement is trivially true. If $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)}<1$ for all $e_n\in E_n^{\theta_n}$, then the statement follows by definition. Finally, suppose there is $e_n'\in E_n^{\theta_n}$ such that $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)}\geq 1$. Then, it must be that for all $e_n\notin E_n^{\theta_n}$, $\frac{\pi(e_n|t_n=i)}{\pi(e_n|t_n=g)}>1$, which implies that $\int\limits_{e_n \notin E_n^{\theta_n}} \pi\left(e_n | t_n = i\right) de_n > \int\limits_{e_n \notin E_n^{\theta_n}} \pi\left(e_n | t_n = g\right) de_n$ which implies the statement.

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] Cremer, Jacques, and Richard P. McLean. "Full extraction of the surplus in Bayesian and dominant strategy auctions." Econometrica (1988): 1247-1257.

[9] 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.

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

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

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

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

[17] Laudan, Larry. "Truth, error, and criminal law: an essay in legal epistemology." Cambridge University Press (2006).

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

[19] Maskin, Eric, and Jean Tirole. "The principal-agent relationship with an informed principal: The case of private values." Econometrica (1990): 379-409.

[20] Midjord, Rune. "Competitive Pressure and Job Interview Lying: A Game Theoretical Analysis", *mimeo* (2013).

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

[22] Myerson, Roger B. "Mechanism design by an informed principal." Econometrica (1983): 1767-1797.

[23] Mylovanov, Tymofiy, and Andriy Zapechelnyuk. "Mechanism Design with expost Verification and Limited Punishments", *mimeo* (2014).

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

[25] Siegel, Ron and Bruno Strulovici. "Improving Criminal Trials by Reflecting Residual Doubt: Multiple verdicts and Plea Bargains", *mimeo* (2016).

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

[27] 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.